NBER WORKING PAPER SERIES

MEASURING THE IMPACTS OF TEACHERS II: TEACHER VALUE-ADDED AND STUDENT OUTCOMES IN ADULTHOOD

Raj Chetty
John N. Friedman
Jonah E. Rockoff

Working Paper 19424
http://www.nber.org/papers/w19424

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2013

John Friedman is currently on leave from Harvard, working at the National Economic Council. This work was completed before he began that position and does not represent the views of the NEC. We thank Joseph Altonji, Josh Angrist, David Card, Gary Chamberlain, David Deming, Caroline Hoxby, Guido Imbens, Brian Jacob, Thomas Kane, Lawrence Katz, Michal Kolesar, Adam Looney, Phil Oreopoulos, Jesse Rothstein, Douglas Staiger, Danny Yagan, anonymous referees, and numerous seminar participants for helpful discussions and comments. This paper is the second of two companion papers on teacher quality. The results in the two papers were previously combined in NBER Working Paper No. 17699, entitled "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," issued in December 2011. On May 4, 2012, Raj Chetty was retained as an expert witness in Vergara vs. California by Gibson, Dunn, and Crutcher LLP to testify about the importance of teacher effectiveness for student learning based on the findings in NBER Working Paper No. 17699. All results based on tax data contained in this paper were originally reported in an IRS Statistics of Income white paper (Chetty, Friedman, and Rockoff 2011a). Sarah Abraham, Alex Bell, Peter Ganong, Sarah Griffis, Jessica Laird, Shelby Lin, Alex Olssen, Heather Sarsons, and Michael Stepner provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard and the National Science Foundation is gratefully acknowledged. Publicly available portions of the analysis code are posted at: http://obs.rc.fas.harvard.edu/chetty/cfr_analysis_code.zip The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood
Raj Chetty, John N. Friedman, and Jonah E. Rockoff
NBER Working Paper No. 19424
September 2013, Revised January 2014
JEL No. H0

ABSTRACT

Are teachers' impacts on students' test scores ("value-added") a good measure of their quality? This question has sparked debate partly because of a lack of evidence on whether high value-added (VA) teachers who raise students' test scores improve students' long-term outcomes. Using school district and tax records for more than one million children, we find that students assigned to high-VA teachers in primary school are more likely to attend college, earn higher salaries, live in higher SES neighborhoods, and have higher savings rates. They are also less likely to have children as teenagers. Teachers have substantial 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 1.3% 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 approximately $250,000 per classroom.

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
john_friedman@harvard.edu

1 Introduction

How can we measure and improve the quality of teaching in primary schools? One prominent but controversial method is to evaluate teachers based on their impacts on their students' test scores, commonly termed the "value-added" (VA) approach.[1] School districts from Washington D.C. to Los Angeles have begun to calculate 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 (e.g., Baker et al. 2010, Corcoran 2010). The debate about teacher VA stems primarily from two questions. First, do the differences in test-score gains across teachers measured by VA capture causal impacts of teachers or are they biased by student sorting? Second, do teachers who raise test scores improve their students' outcomes in adulthood or are they simply better at teaching to the test?

We addressed the first question in the previous paper in this volume (Chetty, Friedman, and Rockoff 2013) and concluded that VA measures that control for lagged test scores exhibit little or no bias. This paper addresses the second question. 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 there is no evidence to date on the long-term impacts of teacher quality as measured by value-added. Understanding this issue is important for policy because a policy maker's ultimate objective is presumably to identify teachers who help students succeed in the long run rather than simply score higher on standardized tests.

We study the long-term impacts of teachers by linking information from an administrative dataset on students and teachers in grades 3-8 from a large urban school district spanning 1989-2009 with selected data from United States tax records spanning 1996-2011. We match approximately 90% of the observations in the school district data to the tax data, allowing us to track approximately one million individuals from elementary school to early adulthood, where we measure outcomes such as earnings, college attendance, and teenage births.

We use two research designs to estimate the long-term impacts of being assigned to a high VA teacher. The first design compares the outcomes of students who were assigned to teachers with different VA, controlling for a rich set of student characteristics such as prior test scores and

[1] Value-added models of teacher quality were pioneered by Hanushek (1971) and Murnane (1975). More recent examples include Rockoff (2004), Rivkin, Hanushek, and Kain (2005), Aaronson, Barrow, and Sander (2007), and Kane and Staiger (2008).

demographics. We implement this approach by regressing long-term outcomes on the VA estimates constructed in our companion paper. The identification assumption underlying this research design is selection on observables: unobserved determinants of outcomes in adulthood such as student ability must be balanced across teachers conditional on the observable characteristics.

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 0.82 percentage points, relative to a sample mean of 37%. 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. Students who are assigned higher VA teachers have steeper earnings trajectories in their 20s. At age 28, the oldest age at which we currently 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 1.3%. If the impact on earnings remains constant at 1.3% over the lifecycle, students would gain approximately $39,000 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 $7,000 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 participation rates in 401(k) retirement savings plans.

Our second design relaxes the assumption of selection on observables using a quasi-experimental approach based on teacher turnover. To understand this research design, suppose a high-VA 4th grade teacher moves from school A to another school in 1995. Because of this staff change, students entering grade 4 in school A in 1995 will have lower quality teachers on average than those in the prior cohort. If high VA teachers improve long-term outcomes, we would expect college attendance rates and earnings for the 1995 cohort to be lower on average than the previous cohort. Building on this idea, we estimate teachers' impacts by regressing changes in mean adult outcomes across consecutive cohorts of children within a school on changes in the mean VA of the teaching staff.

Using this design, we find that a 1 SD improvement in teacher VA raises the probability of college attendance at age 20 by 0.86 percentage points, very similar to the estimate from the first design. Improvements in average teacher VA also increase the quality of colleges that students attend. The impacts on college outcomes are statistically significant with $p < 0.01$. Unfortunately, we have insufficient precision to obtain informative estimates for later outcomes such as earnings –

for which we have data for fewer cohorts – using the quasi-experimental design.

Our quasi-experimental results rest on the identification assumption that high-frequency teacher turnover within school-grade cells is uncorrelated with student and school characteristics. Several pieces of evidence support this identification assumption: (1) predetermined student and parent characteristics are uncorrelated with changes in the quality of teaching staff; (2) changes in teacher VA across cohorts have sharp effects on college attendance exactly in the year of the change but not in prior years or subsequent years; (3) within-school variation in teacher quality across grades yields similar results; and (4) students' prior test scores and contemporaneous scores in the other subject are uncorrelated with changes in the quality of teaching staff in a given subject. Hence, the results from the two research designs together strongly support the view that teacher quality has long lasting impacts on students.

We analyze the heterogeneity of teachers' impacts along several dimensions. The impacts of teacher VA are slightly larger for females than males. Improvements in English teacher quality have larger long-term impacts than improvements in math teacher quality. The impacts of teacher VA are roughly constant in percentage terms by parents' income. Hence, higher income households, whose children have higher earnings on average, should be willing to pay larger amounts for higher teacher VA. We also find teachers' impacts are significant and large throughout grades 4-8, showing that improvements in the quality of education can have large returns well beyond early childhood.[2]

Our conclusion that teachers 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 teachers' impacts on earnings are similar to what one would predict based on the cross-sectional correlation between earnings and contemporaneous test score gains. 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 evaluate Hanushek's (2009) proposal to replace teachers in the bottom 5% of the VA distribution with teachers of average quality. We estimate that replacing a teacher whose current VA is in the bottom 5 percent with an average teacher would increase the mean present value of students' lifetime income by $250,000 per classroom over

[2] 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 2002).

a teacher's career, accounting for drift in teacher quality over time.[3] However, because VA is estimated with noise, the gains from deselecting teachers based on data from a limited number of classrooms are smaller. The present value gain from deselecting the bottom 5% of teachers using three years of test score data is $185,000 per classroom on average. This gain is still about 10 times larger than recent estimates of the additional salary one would have to pay teachers to compensate them for the risk of evaluation based on VA measures (Rothstein 2013). This result suggests that VA could potentially be a useful tool for evaluating teacher performance if the signal quality of VA for long-term impacts does not fall substantially when it is used to evaluate teachers.

We also evaluate the expected gains from policies that pay bonuses to high-VA teachers to increase retention rates. The gains from such policies are only slightly larger than their costs because most bonus payments end up going to high-VA teachers who would have stayed even without the additional payment. 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.

The paper is organized as follows. In Section 2, we present a stylized 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 test scores and adult outcomes. Sections 4 and 5 present results on teachers' long-term impacts using the two research designs described above. We analyze the heterogeneity of teachers' impacts in Section 6. Section 7 presents policy simulations and Section 8 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.

[3] This calculation discounts the earnings gains at a rate of 5% to age 12. The estimated total undiscounted earnings gains from this policy are approximately $50,000 per child and $1.4 million for the average classroom.

2.1 Dynamic Model of Student Outcomes

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

Classroom Assignment Rule. School principals assign student i in school year t to a classroom $c = c(i,t)$ based on observed and unobserved determinants of student achievement. Principals then assign a teacher j to each classroom c based on classroom characteristics. For simplicity, assume that each teacher teaches one class per year, as in elementary schools.

Test Scores. Let $j = j(c(i,t))$ denote student i's teacher in school year t. Let μ_{jt} represent teacher j's "test-score value-added" in year t, i.e., the impact of teacher j on test scores.[4] We scale μ_{jt} in student test-score SDs so that the average teacher has $\mu_{jt} = 0$ and the effect of a 1 unit increase in teacher value-added on end-of-year test scores is 1. We allow teacher quality μ_{jt} to vary with time t to account for the stochastic drift in teacher quality documented in our companion paper (Chetty, Friedman, and Rockoff 2013).

Let $t_i(g)$ denote the calendar year in which student i reaches grade g; $t_i(0)$ denotes the year in which the student starts school (Kindergarten). Student i's test score in year t, A^*_{it}, is given by

$$\begin{aligned} A^*_{it} &= \beta X_{it} + \nu_{it} \\ \text{where } \nu_{it} &= \mu_{jt} + \sum_{s=1}^{t-t_i(0)} \xi_s \mu_{j,t-s} + \theta_c + \widetilde{\varepsilon}'_{it} \end{aligned} \tag{1}$$

Here, X_{it} denotes observable determinants of student achievement, such as lagged test scores and family characteristics. We decompose the error term ν_{it} into four components: current teacher value-added μ_{jt}, the impacts of prior teachers' value-added $\sum_{s=1}^{t-t_{i0}} \xi_s \mu_{j,t-s}$, exogenous class shocks θ_c, and idiosyncratic student-level variation $\widetilde{\varepsilon}'_{it}$. The parameters ξ_s capture the "fade-out" of test score impacts: they measure the impact of teacher quality s years ago on current test scores. The model in (1) coincides with the static model in equation (1) in our companion paper except that here, we decompose the original error term $\widetilde{\varepsilon}_{it} = \sum_{s=1}^{t-t_i(0)} \xi_s \mu_{j,t-s} + \widetilde{\varepsilon}'_{it}$ into the contribution of prior teachers vs. other idiosyncratic fluctuations in scores in order to identify teachers' long-term impacts.

Earnings. Earnings are a function of teacher quality over all years in school, up to grade $G = 12$. Let τ_{jt} represent teacher j's "earnings value-added," i.e. the direct impact of teacher j on earnings holding other factors fixed. We scale τ_{jt} so that the average teacher has $\tau_{jt} = 0$ and the standard

[4] To simplify notation, we write $\mu_{j(i,t),t}$ as μ_{jt} and always denote by j the teacher who taught student i in the relevant year t. For instance, $\mu_{j,t-s}$ denotes the value-added in year $t-s$ of the teacher j who taught student i in year $t-s$. We adopt the same convention with τ_{jt} below as well.

deviation of τ_{jt} is 1. We assume that a teacher's earnings value-added τ_{jt} is linearly related to her test-score value-added μ_{jt}:

$$\tau_{jt} = \phi\mu_{jt} + \tau_{jt}^{\perp}$$

where ϕ measures the relationship between earnings- and test-score VA and $\tau_{jt}^{\perp}$ represents the portion of a teacher's earnings impact that is orthogonal to her test-score impacts.

Earnings Y_i^* are given by

$$Y_i^* = \beta^Y X_{it} + \nu_{it}^Y \tag{2}$$

$$\text{where } \nu_{it}^Y = \sum_{g=0}^{G} \gamma_g \tau_{j,t_i(g)} + \varepsilon_{it}^Y \tag{3}$$

where γ_g measures the effect of teacher quality in grade g on earnings and ε_{it}^Y reflects individual heterogeneity in earnings ability, which may be correlated with academic ability $\widetilde{\varepsilon}_{it}'$.[5] The error ε_{it}^Y may also be correlated with μ_{jt} and τ_{jt} because the principal may systematically sort certain types of students to certain teachers. Accounting for such selection is the key challenge in obtaining unbiased estimates of teachers' causal impacts.

Throughout our analysis, we focus on earnings residuals after removing the effect of observable characteristics:

$$Y_{it} = Y_i^* - \beta^Y X_{it} = \sum_{g=0}^{G} \gamma_g \tau_{j,t_i(g)}^Y + \varepsilon_{it}^Y \tag{4}$$

Note that earnings residuals Y_{it} vary across school years because the control vector X_{it} varies across school years. We estimate the coefficient vector β^Y using variation across students taught by the same teacher using an OLS regression

$$Y_i^* = \alpha_j + \beta^Y X_{it} \tag{5}$$

where α_j is a teacher fixed effect. We estimate β^Y using within-teacher variation to account for the potential sorting of students to teachers based on VA. If teacher VA is correlated with X_{it}, estimates of β^Y in a specification without teacher fixed effects overstate the impact of the X's because part of the teacher effect is attributed to the covariates. See Section 2.2 of our companion paper for further discussion of this issue.

[5] We do not explicitly model class-level shocks to earnings in (2), but none of the results that follow are affected by allowing ε_i^Y to be correlated across students assigned to the same class.

2.2 Reduced-Form Treatment Effects

In this section, we define two notions of a teacher's treatment effect on earnings. To define these treatment effects, suppose that the principal randomly assigns students to teachers in a given year t instead of following his usual assignment rule.

Total Earnings Value-Added. One natural definition of a teacher's impact on earnings is the effect of changing the teacher of class c in grade g from j to j' in year t on expected earnings:

$$\mu_{jt}^{Y} - \mu_{j't}^{Y} = \mathbb{E}Y_{it}(j(i,t)) - \mathbb{E}Y_{it}(j'(i,t)) \tag{6}$$

$$= \gamma_g \left(\tau_{jt}^{Y} - \tau_{j't}^{Y}\right) + \sum_{s=g+1}^{G} \gamma_s \left(\mathbb{E}\left[\tau_{j,t_i(s)}^{Y} \mid j\left(i,t\right)\right] - \mathbb{E}\left[\tau_{j,t_i(s)}^{Y} \mid j'\left(i,t\right)\right]\right). \tag{7}$$

Being assigned teacher j instead of j' affects earnings through two channels. The first term in (7) 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. We refer to μ_{jt}^{Y} as a teacher's "earnings VA" in what follows. The reduced-form parameter μ_{jt}^{Y} is a function of several structural parameters, but is of direct relevance to certain questions, such as the net impact of retaining teachers on the basis of their VA (Todd and Wolpin 2003).

In principle, one can estimate teachers' earnings VA using an approach identical to the one we used to estimate teachers' test-score VA in our first paper. In particular, we could predict a teacher's earnings VA $\widehat{\mu}_{jt}^{Y}$ in year t based on mean residual earnings for students in other years with observational data. In observational data, such a prediction would yield unbiased forecasts of teachers' impacts on earnings if

$$\frac{Cov\left(\mu_{jt}^{Y}, \widehat{\mu}_{jt}^{Y}\right)}{Var\left(\widehat{\mu}_{jt}^{Y}\right)} = 1 \Rightarrow \frac{Cov\left(\varepsilon_{it}^{Y}, \widehat{\mu}_{jt}^{Y}\right)}{Var\left(\widehat{\mu}_{jt}^{Y}\right)} = 0. \tag{8}$$

This condition requires that unobserved determinants of students' earnings are orthogonal to earnings VA estimates. Although conceptually analogous to the requirement for forecast unbiasedness of test-score VA, (8) turns out not to hold in practice. Tests for sorting on pre-determined characteristics analogous to those in Section 5 of our first paper reveal that (8) is violated for earnings VA estimates based on the same control vector (prior test scores and student and classroom de-

mographics) that we used to estimate test score VA. In particular, we find substantial "effects" of earnings VA estimates on parent income and family characteristics, indicating that our baseline control vector is unable to fully account for sorting when estimating earnings VA.

Why are we able to construct unbiased estimates of test score VA but not earnings VA? We believe the central reason is that controlling for lagged test scores effectively absorbs most unobserved determinants of student achievement on which students are sorted to classrooms, but does not account for unobserved determinants of earnings. To see how this can occur, let ζ_i denote a student's academic ability, which affects both test scores and earnings, and ζ_i^Y denote determinants of earnings that are orthogonal to academic achievement, such as family connections. Suppose students are sorted to teachers on the basis of both of these characteristics. The key difference between the two characteristics is that latent student ability ζ_i appears directly in $A_{i,t-1}$, whereas latent student earnings ability ζ_i^Y does not directly appear in $A_{i,t-1}$. As a result, variation in academic ability ζ_i can be largely purged from the error term $\widetilde{\varepsilon}'_{it}$ in the specification for test scores in (1) by controlling for $A_{i,t-1}$.[6] In contrast, family connections are not reflected in $A_{i,t-1}$ and therefore appear in the error term ε_{it}^Y in the specification for earnings in (2). Under such a data generating process, we would be able to identify teachers' causal impacts on test scores by controlling for $A_{i,t-1}$, but would not be able to identify teachers' causal impacts on earnings because there is systematic variation across teachers in students' earnings purely due to variation in family connections ζ_i^Y even conditional on $A_{i,t-1}$.

Consistent with this reasoning, we showed in our first paper that the key to obtaining forecast unbiased estimates of test-score VA was to control for prior test scores, $A_{i,t-1}$. If we observed an analog of lagged scores such as lagged expected earnings, we could effectively control for ζ_i^Y and potentially satisfy (8). Lacking such a control in our data, we cannot identify teachers' total earnings VA and defer this task to future work.

Impacts of Test-Score VA on Earnings. An alternative objective is to identify the impacts of teachers' test-score based VA μ_{jt} on earnings. Let σ_μ denote the standard deviation of teachers' test-score VA. The reduced-form earnings impact of having a 1 SD better teacher, as measured by

[6] In general, controlling for lagged test scores need not completely account for the variation in ζ_i because lagged test scores are noisy measures of latent ability. The fact that controlling for $A_{i,t-1}$ does eliminate bias in practice (as shown in our first paper) suggests that students are allocated to classrooms based on factors highly correlated with $A_{i,t-1}$ and other factors that directly affect earnings (ζ_i^Y).

test-score VA, in grade g is

$$\kappa_g = \mathbb{E}[Y_{it} \mid \mu_{j't} = \mu_{jt} + \sigma_\mu] - \mathbb{E}[Y_{it} \mid \mu_{jt}] \tag{9}$$

$$= \sigma_\mu \phi \gamma_g \left(\tau^Y_{j'(i,g)} - \tau^Y_{j(i,g)} \right) + \sum_{s=g+1}^{G} \gamma_s \left(\mathbb{E}\left[\tau^Y_{j,t_i(s)} \mid \mu_{j't} \right] - \mathbb{E}\left[\tau^Y_{j,t_i(s)} \mid \mu_{jt} \right] \right). \tag{10}$$

As above, the reduced-form impact κ_g consists of two terms. The first is the direct effect of having a better teacher in grade g in school year T, which is attenuated by $\phi = \frac{Cov(\tau_{jt}, \mu_{jt})}{Var(\mu_{jt})}$ because we only pick up the portion of earnings impacts that projects onto test-score VA. The second is the impact of having different teachers in subsequent grades.

Let $m_{jt} = \mu_{jt}/\sigma_\mu$ denote teacher j's "normalized value-added," i.e. teacher quality scaled in standard deviation units of the teacher VA distribution. Under the assumption that the conditional expectation $\mathbb{E}[Y_{it} \mid \mu_{jt}]$ is a linear function of μ_{jt}, we can write

$$Y_{it} = a + \kappa_g m_{jt} + \eta_{it} \tag{11}$$

where η_{it} is orthogonal to m_{jt} under the assumption of random assignment in period t. Intuitively, if student unobservables are orthogonal to teachers' test-score VA, regressing Y_{it} on m_{jt} identifies the impact of a 1 SD increase in teachers' test-score VA on earnings.

The parameter κ_g is of interest for two reasons. First, from a policy perspective, it is important to determine the extent to which existing test-score based VA measures predict teachers' long-term impacts. Second, κ_g is a lower bound for the teachers' total impacts on earnings ($SD\left(\mu^Y_{jt}\right) \geq \kappa_g$) if the effects of current teacher quality on future teacher quality are small, as is the case empirically.[7] Intuitively, κ^2_g measures the portion of $Var\left(\mu^Y_{jt}\right)$ due to variation in teachers' test-score VA.

We cannot directly estimate κ_g using (11) because true test-score VA m_{jt} is unobserved. We can substitute an estimate of teacher VA $\widehat{m}_{jt} = \hat{\mu}_{jt}/\sigma_\mu$ for true teacher VA in (11) under the following assumption.

Assumption 1 [Forecast Unbiasedness of Test-Score VA] Test-score value-added estimates are forecast unbiased predictors of test scores:

$$\frac{Cov\left(A_{it}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)} = 1.$$

[7] When current teacher assignments have no impact on future teacher assignments, earnings $Y_{it} = \gamma_g \tau_{jt} + \varepsilon^Y_{it}$. In this case, $\kappa_g \equiv \frac{Cov(Y_{it}, m_{jt})}{Var(m_{jt})} = \frac{Cov(\gamma_g \tau_{jt}, \mu_{jt})}{Var(\mu_{jt})}\sigma_\mu$ and hence earnings VA $\mu^Y_{jt} = \gamma_g \tau_{jt} = \kappa_g m_{jt} + \gamma_g \tau^{\perp}_{jt}$. It follows that $Var(\mu^Y_{jt}) \geq Var(\kappa_g m_{jt}) = \kappa^2_g$. With tracking, this identity need not hold because the tracking process for test-score VA and total earnings VA could differ. Empirically, we find that tracking based on past teacher quality is small, suggesting that our estimates do give a lower bound on teachers' total earnings VA.

In our companion paper, we demonstrate that Assumption 1 holds for the VA estimates that we use in this paper. Under Assumption 1, $\frac{Cov\left(A_{jt},\widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)} = \frac{Cov\left(\mu_{jt},\widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)} = \frac{Cov(m_{jt},\widehat{m}_{jt})}{Var(\widehat{m}_{jt})} = 1$. Since $Cov\left(\widehat{m}_{jt},\eta_{it}\right) = 0$ under random assignment in year t, it follows from (11) that $Cov\left(Y_{it},\widehat{m}_{jt}\right) = \kappa_g Cov\left(m_{jt},\widehat{m}_{jt}\right)$ and hence

$$\kappa_g = \frac{Cov\left(Y_{it},\widehat{m}_{jt}\right)}{Cov\left(m_{jt},\widehat{m}_{jt}\right)} = \frac{Cov\left(Y_{it},\widehat{m}_{jt}\right)}{Var\left(\widehat{m}_{jt}\right)}.$$

That is, we can identify κ_g by regressing earnings residuals Y_{it} on the teacher VA estimates $\hat{\mu}_{jt}$ constructed using observational data as in our first paper by estimating the following OLS regression specification:

$$Y_{it} = \alpha + \kappa_g \widehat{m}_{jt} + \eta'_{it} \tag{12}$$

Intuitively, we can identify κ_g using 2SLS by instrumenting for m_{jt} with our teacher VA estimates under random assignment in year t. Forecast unbiasedness of test-score VA implies that the first stage of this 2SLS regression has a coefficient of 1. Thus, the reduced form coefficient obtained from an OLS regression of earnings on the VA estimate coincides with κ_g, provided that the error term η'_{it} is orthogonal to the variation in $\widehat{m}_{jt}$.[8] We develop research designs that isolate such variation in $\widehat{m}_{jt}$ using observational data below.

Impacts of Multiple Teachers. The treatment effects defined above measure the total impact of having a better teacher in a single grade g, including both direct effects and the impacts of being tracked to better teachers in future grades. One may also wish to identify the direct impacts of teachers in each grade. Let $\tilde{\kappa}_g$ denote the impact of teacher VA in grade g on earnings holding fixed teacher VA in other grades. An intuitive specification to identify $\tilde{\kappa}_g$ is to regress earnings on teacher VA in all grades simultaneously:

$$Y_i^* = \sum_{g=0}^{G} \left[\tilde{\kappa}_g \widehat{m}_{j,t_i(g)} + \tilde{\beta}_g X_{i,t_i(g)}\right] + \varepsilon_i^m. \tag{13}$$

Identifying $\{\tilde{\kappa}_g\}$ in (13) requires the orthogonality condition $Cov\left(\widehat{m}_{j,t_i(g)},\varepsilon_i^m\right) = 0$. This orthogonality condition is violated if we do not include grade $g-1$ test scores $A_{i,g-1}$ in the control vector X because teacher assignment is correlated with lagged test scores and other factors that directly affect earnings, as shown in Table 7 of our companion paper. But $A_{i,g-1}$ is endogenous to grade

[8] Another way to identify κ_g is using a correlated random effects or factor model, in which a teacher's random effect on test scores (μ_{jt}) is correlated with her random effect on earnings (μ_{jt}^Y). One can then directly estimate the covariance of μ_{jt}^Y and μ_{jt}. Chamberlain (2013) develops such an approach and obtains estimates similar to those reported here.

$g-1$ teacher VA $\widehat{m}_{j,t_i(g-1)}$, implying that we cannot estimate (13) to identify $\tilde{\kappa}_g$. Conceptually, estimating the effects of multiple teachers requires simultaneous quasi-random assignment of teachers in multiple grades. Our primary research design, which requires conditioning on lagged test scores, only yields quasi-random variation in teacher assignment one grade at a time. As a result, we cannot directly estimate (13) and we also cannot identify the substitutability or complementarity of teachers' impacts across grades.

Given this problem, we develop an iterative method of recovering the net impacts $\tilde{\kappa}_g$ from our reduced form estimates κ_g and estimates of the degree of teacher tracking in Section 6.3. The degree of tracking turns out to be small in our data, and thus the reduced-form impacts reported below are very similar to the direct impacts of each teacher.

3 Data

We draw information from two administrative databases: school district records and federal income tax returns. We obtain information on students – including math and English test scores and teacher assignments – from the 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. We obtain information on student outcomes in adulthood and their parents' characteristics from U.S. federal income tax returns spanning 1996-2011.

The structure of both datasets, how they are linked, and the sample restrictions we impose to arrive at our core sample are described in Section 3 of our companion paper (Chetty, Friedman, and Rockoff 2013). Because the adult outcomes we analyze are measured at age 20 or afterward, in this paper we restrict the core sample of 10.7 million observations to students who would have graduated high school by the 2008-09 school year (and thus turned 20 by 2011) if they progressed through school at a normal pace.[9] This leaves a sample of 6.8 million student-subject-year observations (covering 1.1 million students) that we use to study teachers' long-term impacts.[10] We refer to this sample as the "linked analysis sample." Within the linked analysis sample, we match 87.4% of students and 89.2% of student-subject-year observations in the school district data to the tax data. We find little or no correlation between match rates and teacher VA in the various subsamples we use

[9] 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 2008-09 if she progressed at the normal pace. That is, we include all classrooms in which $\min_i(12 + \text{school year} - \text{grade}_i) \leq 2009$.

[10] For much of the analysis in our first paper, we restricted attention to the subset of observations in the core sample that have lagged scores and other controls needed to estimate the baseline VA model. Because we do not control for individual-level variables in most of the specifications in this paper, we do not impose that restriction here.

in our analysis and obtain very similar estimates of teachers' impacts on long-term outcomes when restricting the sample to school-grade-subject cells with above-median match rates (see Appendix Table 7).

The linked analysis sample 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). 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 the estimation of (12). We account for the fact that each student appears multiple times in the dataset by clustering standard errors as described in Section 4.1.

3.1 Definitions of Outcomes in Adulthood

In this subsection, we describe the outcomes in adulthood for children that we measure using information from tax returns. All variables from the school district data and parent characteristics from the tax data are defined in Section 3 of our companion paper.

Earnings. Individual wage earnings data come from W-2 forms, which are available from 1999-2011.[11] Importantly, W-2 data are available for *both* tax filers and non-filers, eliminating concerns about missing data on formal sector earnings. We cap earnings in each year at $100,000 to reduce the influence of outliers; 1.3% of individuals in the sample report earnings above $100,000 at age 28. We measure income in 2010 dollars, adjusting for inflation using the Consumer Price Index. Individuals with no W-2 are coded as having 0 earnings. 33.1% of individuals have 0 wage earnings at age 28 in our sample.

Total Income. To obtain a more comprehensive definition of income, we define "total income" as the sum of W-2 wage earnings and household self-employment earnings (as reported on the 1040). For non-filers we define total income as just W-2 wage earnings; those with no W-2 income are coded as having zero total income. 29.6% of individuals have 0 total income in our sample.[12] We show that similar results are obtained using this alternative definition of income, but use W-2

[11] Here and in what follows, the year refers to the tax year, i.e. the calendar year in which income is earned. In most cases, tax returns for tax year t are filed during the calendar year $t+1$.

[12] According to the Current Population Survey, 27.2% of the non-institutionalized population between the ages of 25 and 29 was not employed in 2011. The non-employment rate in our sample may differ from this figure for several reasons, including the following: (1) it is based on annual rather than weekly data, (2) it includes the institutionalized population in the denominator, and (3) it applies to a relatively low-income public school district in an urban city.

wage earnings as our baseline measure because it (1) is unaffected by the endogeneity of tax filing and (2) provides a consistent definition of individual (rather than household) income for both filers and non-filers.

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 eligible for federal student aid – 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 independent of whether an individual files a tax return, we have complete records on college attendance for all indivduals. However, 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 1999-2011.

Comparisons to other data sources indicate that 1098-T forms capture college enrollment quite accurately.[13] The correlation between enrollment counts for students age 18-21 based on 1098-T's and enrollment counts for colleges listed in the IPEDS dataset from the Department of Education exceeds 0.95. The aggregate counts are also aligned as one would expect.[14] Finally, two independent evaluations of the Project STAR class size experiment using data from 1098-T's (Chetty et al. 2011) and the National Student Clearinghouse (Dynarski et al. 2011) obtained nearly identical point estimates of the impacts of class size on college attendance.

College Quality. We construct an earnings-based index of college quality by building upon the work of Chetty et al. (2011). Using the population of all current U.S. citizens born in 1979 or 1980, we group individuals by the higher education institution they attended at age 20. We pool individuals who were not enrolled in any college at age 20 together in a separate "no college" category. For each college or university (including the "no college" group), we then compute the mean W-2 earnings of the students when they are age 31 (in 2010 and 2011). Among colleges attended by students in the school district studied in this paper, the average value of our earnings index is \$44,048 for four-year colleges and \$30,946 for two-year colleges. For students who did not attend college, the mean earnings level is \$17,920.[15]

[13] Legally, 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 appear to be available even for such cases, perhaps because of automated reporting to the IRS by universities.

[14] In 2009, 27.4 million 1098-T forms were issued (Internal Revenue Service, 2010). According to the Current Population Survey (US Census Bureau, 2010, Tables V and VI), in October 2008, there were 22.6 million students in the U.S. (13.2 million full time, 5.4 million part-time, and 4 million vocational). As an individual can be a student at some point during the year but not in October and can receive a 1098-T form from more than one institution, the number of 1098-T forms for the calendar year should indeed be higher than the number of students as of October.

[15] If students attended two or more colleges in a given year, we assign them the maximum college quality across

In Appendix A, we analyze the robustness of the college quality index to alternative specifications, such as measuring earnings and college attendance at different ages and defining the index based on total income instead of W-2 earnings. We find that rankings of college quality are very stable across cohorts and are robust to alternative specifications provided that earnings are measured after age 28 (Appendix Figure 1, Appendix Table 2).

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 1999-2011. We define saving for retirement as an indicator for contributing to a 401(k) at age 28.

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 2011. We observe dates of birth and death for all dependents and tax filers until the end of 2011 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 2011. Second, the mother must herself claim the child as a dependent at some point during the sample years. If a child is claimed as a dependent by another individual (e.g., a 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 who are not biological children, but were born between 13 and 19 years after the female who claims them as a dependent. Because most dependents who are not biological children 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 Appendix Figure 2d and Appendix Table 3). For instance, women

all colleges attended. We code college quality as missing for all institutions with fewer than 100 students and for institutions founded in or after 2001. As a result, 0.21% of students who attend college at age 20 are missing information on college quality.

who score higher on tests, attend college, or have higher income parents are significantly less likely to have teenage births according to our measure.

3.2 Summary Statistics

Our core analysis sample for long-term outcomes contains 6.8 million student-year-subject observations. Table 1 reports summary statistics for this dataset. 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 1.1 million students in our analysis dataset; on average, each student has 6.25 subject-school year observations.

The mean test score in the analysis sample is positive and has a standard deviation below 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. 77.1% of students are eligible for free or reduced price lunches.

The availability of data on adult outcomes naturally varies across cohorts. There are more than 5.9 million observations for which we observe college attendance at age 20. We observe earnings at age 25 for 2.3 million observations and at age 28 for 1.3 million observations. Because many of these observations at later ages are from older cohorts of students who were in middle school in the early 1990s, we were not able to obtain information on teachers. As a result, there are only 1.6 million student-subject-school year observations for which we see *both* teacher VA and earnings at age 25, 750,000 at age 28, and only 220,000 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 individual earnings at age 28 is $20,885 (in 2010 dollars), while mean total income is $21,272.

For students whom we are able to link to parents, the mean parent household income is $40,808, while the median is $31,834. 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 $34,515, with 10% of parents earning more than $100,000.

3.3 Cross-Sectional Correlations

As a benchmark for evaluating the magnitude of the causal effects estimated below, Appendix Tables 3-6 report coefficients from OLS regressions of various adult outcomes 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 \$7,700 (36%) increase in earnings at age 28. Conditional on the student- and class-level controls X_{it} that we define in Section 4.1 below, a 1 SD increase in the current test score is associated with \$2,600 (12%) increase in earnings on average.

Appendix Figure 2 presents binned scatter plots of selected outcomes vs. test scores both with and without controls. The unconditional relationship between scores and outcomes is S-shaped, while the relationship conditional on prior scores and other covariates is almost perfectly linear. We return to these results below and show that the causal impacts of teacher VA on earnings and other outcomes are commensurate to what one would predict based on these correlations.

4 Research Design 1: Cross-Class Comparisons

Our first method of estimating teachers' long-term impacts builds on our finding that conditioning on prior test scores and other observables is adequate to obtain unbiased estimates of teachers' causal impacts on test scores (Chetty, Friedman, and Rockoff 2013). Given this result, one may expect that comparing the long-term outcomes of students assigned to different teachers conditional on the same control vector will yield unbiased estimates of teachers' long-term impacts. The next subsection formalizes the identification assumptions and estimating equations underlying this approach. We then present results for three sets of impacts: college attendance, earnings, and other outcomes such as teenage birth. We report mean teacher impacts across all grades and subjects in this section and analyze heterogeneity in impacts across these groups in Section 6.

4.1 Methodology

As in Section 2.1, let X_{it} denote a vector of observable characteristics on which students are sorted to teachers. Let $Y_{it} = Y_i^* - \beta^Y X_{it}$ denote earnings (or other long-term outcome) residuals adjusted for X_{it}, where β^Y is estimated using within-teacher variation as in (5). Following (12), we regress students' earnings residuals on their teachers' normalized VA $\widehat{m}_{jt}$, pooling all grades and subjects:

$$Y_{it} = \alpha + \kappa \widehat{m}_{jt} + \eta'_{it} \tag{14}$$

Recall that each student appears in our dataset once for every subject-year with the same level of Y_{it} but different values of $\widehat{m}_{jt}$. Hence, κ represents the mean reduced-form impact of having a higher VA teacher for a *single* grade between grades 4-8. Estimating (14) using OLS yields an

unbiased estimate of κ under the following assumption.

Assumption 2 [Selection on Observables] Test-score value-added estimates are orthogonal to unobserved determinants of earnings conditional on X_{it}:

$$Cov\left(\widehat{m}_{jt}, \eta'_{it}\right) = 0. \tag{15}$$

Assumption 1 is weaker than the assumption needed to identify total earnings VA in (8) because it only requires that there be no correlation between teachers' test-score VA and unobservables. In our example in Section 2.2, Assumption 2 allows students with better family connections ζ_i^Y to be systematically tracked to certain teachers as long as those teachers do not systematically have higher levels of test-score VA (conditional on X_{it}).

Four methodological issues arise in estimating (14): (1) estimating test-score VA $\widehat{m}_{jt}$, (2) specifying a control vector X_{it}, (3) calculating the standard error on κ, and (4) accounting for outliers in $\widehat{m}_{jt}$. The remainder of this subsection addresses these four issues.

Estimating Test-Score VA. We define normalized VA $\widehat{m}_{jt} = \widehat{\mu}_{jt}/\sigma_\mu$, where $\widehat{\mu}_{jt}$ is the baseline estimate of test-score VA for teacher j in year t constructed in our companion paper.[16] We define σ_μ as the standard deviation of teacher effects for the corresponding subject and school-level using the estimates in Table 2 of our companion paper: in elementary school, 0.163 for math and 0.124 for English and in middle school, 0.134 for math and 0.098 for English. With this scaling, a 1 unit increase in $\widehat{m}_{jt}$ corresponds to a teacher who is rated 1 SD higher in the distribution of true teacher quality for her subject and school-level. Note that because $\widehat{\mu}_{jt}$ is shrunk toward the sample mean to account for noise in VA estimates, $SD(\widehat{\mu}_{jt}) < \sigma_\mu$ and hence the standard deviation of the normalized VA measure $\widehat{m}_{jt}$ is less than 1. We demean $\widehat{m}_{jt}$ within each of the four subject (math vs. English) by school level (elementary vs. middle) cells in the estimation sample in (14) so that κ is identified purely from variation within the subject-by-school-level cells.[17]

Importantly, the VA estimates $\widehat{m}_{jt}$ are predictions of teacher quality in year t based on test score data from all years excluding year t. For example, when predicting teachers' effects on the outcomes of students they taught in 1995, we estimate $\widehat{m}_{j,1995}$ based on residual test scores from students in all years of the sample *except* 1995. To maximize precision, the VA estimates are

[16] Unless otherwise specified, the independent variable in all the regressions and figures in this paper is normalized test-score VA $\widehat{m}_{jt}$. For simplicity, we refer to this measure as "value-added" or VA below.

[17] By construction, the normalized VA estimates have mean 0 within each subject by school level cell in the full sample. However, there are slight deviations in some of the estimation samples due to missing data. In practice, demeaning m_{jt} has little or no impact on our estimates. Note that the estimates we report equal what one would obtain by running separate regressions within the four subject-by-school-level cells and taking weighted means of the coefficients.

based on data from all years for which school district data with teacher assignments are available (1991-2009), not just the subset of older cohorts for which we observe long-term outcomes.

Using a leave-year-out estimate of VA is necessary to obtain unbiased estimates of teachers' long-term impacts because of correlated errors in students' test scores and later outcomes. Intuitively, if a teacher is randomly assigned unobservably high ability students, her estimated VA will be higher. The same unobservably high ability students are likely to have high levels of earnings $\eta_{it}^{'}$, generating a mechanical correlation between VA and earnings even if teachers have no causal effect ($\kappa = 0$). The leave-year-out approach eliminates this correlated estimation error bias because $\widehat{m}_{jt}$ is estimated using a sample that excludes the observations on the left hand side of (14).[18]

Control Vectors. We construct residuals Y_{it} using separate models for each of the four subject-by-school-level cells. Within each of these groups, we regress raw outcomes Y_i^* on a vector of covariates X_{it} with teacher fixed effects, as in (5), and compute residuals Y_{it}. We partition the control vector X_{it} that we used to construct our baseline VA estimates into two components: student-level controls X_{it}^I that vary across students within a class and classroom-level controls X_{ct} that vary only at the classroom level. The student-level control vector X_{it}^I includes cubic polynomials in prior-year math and English scores. We interact these cubics with the student's grade level to permit flexibility in the persistence of test scores as students age. We also control for the following student level characteristics: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, free or reduced-price lunch, special education, and limited English. The class-level controls X_{ct} consist of the following elements: (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_{it}^I, and (4) grade and year dummies.[19]

In our baseline analysis, we control only for the class-level controls X_{ct} when estimating the residuals Y_{it}. Omitting individual-level controls allow us to estimate Y_{it} and (14) using a dataset collapsed to classroom means, which greatly reduces computational costs given the size of the

[18] 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.

[19] The control vector X_{it} that we use here exactly matches the control vector used to construct the VA estimates $\widehat{m}_{jt}$. There is no reason that the two control vectors must match exactly; we adopted this approach to avoid specification searching. However, unconditional regressions of Y_{it}^* on $\widehat{m}_{jt}$ (with no controls) yield substantially upward-biased estimates of κ relative to the quasi-experimental estimates in the next section. This is not surprising because teacher VA is correlated with characteristics such as prior scores and parent income unconditionally, as shown in Table 7 of our first paper.

student-level dataset.[20] In particular, let $Y_{ct} = Y_c^* - \beta_C X_{ct}$ denote the residual of mean outcomes Y_c^* in class c in year t, where β_C is estimated at the class level using within-teacher variation across classrooms as in (5), weighting by class size. We estimate the impact of teacher VA on mean outcomes using a class-level OLS regression analogous to (14), again weighting by class size:

$$Y_{ct} = \alpha + \kappa_C \widehat{m}_{jt} + \eta'_{ct} \tag{16}$$

We show in Appendix A that the point estimate $\hat{\kappa}_C$ in (16) coincides exactly with $\hat{\kappa}$ in (14). Intuitively, because teacher VA varies only at the classroom level, deviations of individual-level controls $(X_{it}^I - X_{ct})$ and outcomes $(Y_i^* - Y_{ct}^*)$ from class means are uncorrelated with $\widehat{m}_{jt}$ and thus have no impact on the estimate of κ.

In practice, the identity $\hat{\kappa} = \hat{\kappa}_C$ does not hold exactly because the class means X_{ct} are defined using all observations with non-missing data for the relevant variable. Some students are not matched to the tax data and hence are missing Y_i^*, while other students are missing some of the individual-level covariates X_{it}^I (e.g., prior-year test scores). As a result, X_{ct} does not exactly equal the mean of X_{it}^I within classrooms in the final estimation sample. To verify that the small discrepancies between X_{ct} and X_{it}^I do not affect our estimates of κ, we show in Appendix Table 7 that the inclusion of individual controls X_{it}^I has little impact on the point estimates of κ by estimating (14) for a selected set of specifications on the individual data.

Standard Errors. The dependent variable in (14) has 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 would be to cluster standard errors by both student and classroom (Cameron, Gelbach, and Miller 2011). Unfortunately, two-way clustering of this form requires running regressions on student-level data and thus was computationally infeasible at the Internal Revenue Service. We instead cluster standard errors at the school by cohort level when estimating (16) at the class level, which adjusts for correlated errors across classrooms and repeat student observations within a school. The more conservative approach of clustering by school does not affect our hypothesis tests at conventional levels of statistical significance but yields slightly wider confidence intervals as expected (Appendix Table 7).[21]

[20] We estimated test-score VA $\widehat{m}_{jt}$ using individual-level controls in Chetty, Friedman, and Rockoff (2013) because that estimation did not use any information from the tax data and could be done entirely outside the Internal Revenue Service (IRS). Outcomes in adulthood can only be analyzed internally at the IRS, where computational capacity is restricted.

[21] In Appendix Table 7 of our working paper (Chetty, Friedman, and Rockoff 2011b), we evaluated the robustness of

Outliers. In our baseline specifications, we exclude classrooms taught by teachers whose estimated VA $\widehat{m}_{jt}$ falls in the top one percent for their subject and school level (above 2.03 in math and 1.94 in English in elementary school and 1.93 in math and 1.19 in English in middle school). We do so because these teachers' impacts on test scores appear suspiciously consistent with testing irregularities indicative of test manipulation. 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 show that this proxy for cheating is highly correlated with unusual answer sequences that directly reveal test manipulation. Teachers in the top 1% of our estimated VA distribution are significantly more likely to show suspicious patterns of test score gains followed by steep losses, as defined by Jacob and Levitt's proxy (see Appendix Figure 3).[22] We therefore trim the top 1% of outliers in all the specifications reported in the main text. We investigate how trimming at other cutoffs affects our 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 1% reduces our estimates of teachers' impacts on long-term outcomes by 10-30%. In contrast, excluding the bottom 1% 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 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 99th percentile of VA.

4.2 College Attendance

We begin by analyzing the impact of teachers' test-score VA on college attendance at age 20, the age at which college attendance rates are maximized in our sample. Figure 1a plots residual college attendance rates for students in school year t vs. $\widehat{m}_{jt}$, the leave-year-out estimate of their teacher's VA in year t. To construct this binned scatter plot, we first residualize college attendance rates

our results to additional forms of clustering for selected specifications. We found that school-cohort clustering yields more conservative confidence intervals than more computationally intensive techniques such as two-way clustering by student and classroom.

[22] Appendix Figure 3 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 0.47 SD in year t followed by mean test score *losses* of 0.42 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.

with respect to the class-level control vector X_{ct} separately within each subject by school-level cell, using within-teacher variation to estimate the coefficients on the controls as described above. We then divide the VA estimates $\widehat{m}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the mean of the college attendance residuals in each bin against the mean of $\widehat{m}_{jt}$ in each bin. Finally, we add back the mean college attendance rate in the estimation sample to facilitate interpretation of the scale.[23] Note that this binned scatter plot provides a non-parametric representation of the conditional expectation function but does not show the underlying variance in the individual-level data. The regression coefficient and standard error reported in this and all subsequent figures are estimated on the class-level data using (16), with standard errors clustered by school-cohort.

Figure 1a shows that 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 11 ($p < 0.001$). On average across subjects and grades, a 1 SD increase in a teacher's test score VA in a single grade increases the probability of college attendance by $\kappa = 0.82$ percentage points at age 20, relative to a mean of 37.22%. This impact of a 2.2% increase in college attendance rates for a 1 SD improvement in teacher VA is roughly similar to the impacts on other outcomes we document below.

The relationship in Figure 1a can be interpreted as a causal effect of teacher quality only if the selection on observables assumption in (15) holds. One way to evaluate the validity of this assumption is to assess the degree of selection bias due to observable characteristics that were excluded from our baseline control vector. As in our companion paper, we use two sets of variables for this purpose: parent characteristics and lagged test score gains.

The parent characteristics P_{it}^{*} consist of the following variables: 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.[24] Let P_{ct} denote the classroom means of these parent characteristics residualized on X_{ct} using within-teacher variation. These parent characteristics are ideal variables to test for selection because they are strong predictors of college attendance even conditional on the baseline controls X_{ct}. The F-statistic on the parent characteristics in the regression of the baseline college residuals Y_{ct} on P_{ct} exceeds 300.

[23] In this and all subsequent scatter plots, we also demean $\widehat{m}_{jt}$ within subject-by-school-level groups to isolate variation within these cells as in the regressions, and then add back the unconditional mean of $\widehat{m}_{jt}$ in the estimation sample.

[24] We code the parent characteristics as 0 for the 5.2% of students whom we matched to the tax data but were unable to link to 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 the small number of observations where we match parents but do not have data on parents' ages, and include an indicator for such cases.

Column 1 of Table 2 replicates the specification in Figure 1a with the baseline control vector X_{ct} as a reference. Column 2 replicates Column 1, adding P^*_{ct} to the control vector used to residualize college attendance. The estimate in Columns 2 is quite similar to that in Column 1, indicating that selection on parent characteristics has a modest impact on the estimated effect of teacher quality on college attendance rates.

Next, we assess selection on lagged test score gains. Let $A^*_{i,t-2}$ denote twice-lagged test scores, and let $A_{c,t-2}$ denote the classroom means of this variable residualized on X_{ct}. Twice-lagged test scores have considerable predictive power for college attendance even conditional on X_{ct} because test scores are a noisy measure of latent ability. The F-statistic on $A_{c,t-2}$ in a regression of the baseline college residuals Y_{ct} on $A_{c,t-2}$ exceeds 400. Column 3 of Table 2 replicates Column 1, adding class means of twice-lagged test scores $A^*_{c,t-2}$ to the control vector instead of parent characteristics. Again, the coefficient does not change appreciably, indicating that our estimate of κ is not significantly biased by selection on unobserved determinants of prior achievement.[25] Although these tests for selection on excluded observables are not conclusive, they support the identification assumption in (15).

College Quality. We use the same set of specifications to analyze whether high-VA teachers also improve the quality of colleges that their students attend, as measured by the earnings of students who previously attended the same college (see Section 3.2). Students who do not attend college are assigned the mean earnings of individuals who do not attend college. Figure 1b plots the earnings-based index of quality for college attended at age 20 vs. teacher VA, using the same baseline controls X_{ct} and technique as in Figure 1a. Again, there is a highly significant relationship between the quality of colleges students attend and the quality of the teachers they had in grades 4-8 ($t = 14.4$, $p < 0.001$). A 1 SD improvement in teacher VA raises college quality by \$299 (or 1.11%) on average, as shown in Column 4 of Table 2. Columns 5 and 6 replicate Column 4 adding parent characteristics and lagged test score gains to the baseline control vector. As with college attendance, the inclusion of these controls has only a modest effect on the point estimates.

The \$299 estimate in Column 4 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 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

[25] The sample in Column 3 has fewer observations than in Column 1 because twice lagged test scores are not observed in 4th grade. Replicating the specification in Column 1 on exactly the estimation sample used in Column 3 yields an estimate of 0.81% (0.09).

attendance. We take two approaches to overcome this selection problem and identify intensive-margin effects. First, we define colleges with earnings-based quality above the student-weighted median in our sample ($43,914) as "high quality." We regress this high quality college indicator on teacher VA in the full sample, including students who do not attend college, and find that a 1 SD increase in teacher VA raises the probability of attending a high quality college by 0.72%, relative to a mean of 13.41% (Column 7 of Table 2). 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 $41,756, while the quality for all those who do not attend college is $17,920. This suggests that at most $(41,756 - 17,920) \times 0.82\% = \$195$ of the $299 impact is due to the extensive margin response, confirming that teachers improve the quality of colleges that students attend.

Finally, we analyze the impact of teacher quality on the number of years in college. Column 8 replicates the baseline specification in Column 1, replacing the dependent variable with an indicator variable for attending college in at least 4 years between 18 and 22. A 1 SD increase in teacher quality increases the fraction of students who spend 4 or more years in college by 0.79 percentage points (3.2% of the mean).[26] While we cannot directly measure college completion in our data, this finding suggests that higher quality teachers increase not just attendance but also college completion rates.

Figure 1c plots the impact of a 1 SD improvement in teacher quality on college attendance rates at all ages from 18-28. We run separate regressions of college attendance at each age on teacher VA, using the same specification as in Column 1 of Table 2. As one would expect, teacher VA has the largest impacts on college attendance rates before age 22. However, the impacts remain significant even in the mid 20's, perhaps because of increased attendance of graduate or professional schools. These continued impacts on higher education affect our analysis of earnings impacts, to which we now turn.

[26] The magnitude of the four-year attendance impact (0.79 pp) is very similar to the magnitude of the single-year attendance impact (0.82 pp). Since the students who are on the margin of attending for one year presumably do not all attend for four years, this suggests that better teachers increase the number of years that students spend in college on the intensive margin.

4.3 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. 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 within the tax data in Appendix Figure 4.[27]

Figure 2a plots individual (W-2) wage earnings at age 28 against VA $\widehat{m}_{jt}$, conditioning on the same set of classroom-level controls as above. Being assigned to a higher value-added teacher has a significant impact on earnings, with the null hypothesis of $\kappa = 0$ rejected with $p < 0.01$. A 1 SD increase in teacher VA in a single grade increases earnings at age 28 by \$350, 1.65% of mean earnings in the regression sample.

Columns 1-3 of Table 3 evaluate the robustness of this estimate to the inclusion of parent characteristics and lagged test score gains. These specifications mirror Columns 1-3 of Table 2, but use earnings at age 28 as the dependent variable. As with college attendance, controlling for these additional observable characteristics has relatively small effects on the point estimates, supporting the identification assumption in (15). The smallest of the three estimates implies that a 1 SD increase in teacher VA raises earnings by 1.34%.

To interpret the magnitude of this 1.34% impact, consider the lifetime earnings gain from having a 1 SD higher VA teacher in a single grade. Assume that the percentage gain in earnings remains constant at 1.34% 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.[28] 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 $1.34\% \times \$522,000 \simeq \$7,000$ per grade. The undiscounted lifetime earnings gain (assuming a 2% growth rate but 0% discount rate) is approximately \$39,000 per student.[29]

[27] Appendix Figure 4 correlates earnings at age t with earnings at age $t+12$ for the all individuals in the tax data. The correlation of wage earnings at age 28 with wage earnings at age 40 is fairly close to the maximum 12-year-ahead correlation over the lifecycle, suggesting that our earnings measures provide reasonably reliable proxies for lifetime income.

[28] 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.

[29] These gains reflect the value of a 1 SD improvement in *actual* teacher VA m_{jt}. Being assigned to a teacher

A second benchmark is the increase in earnings from an additional year of schooling, which is around 9% (Gunderson and Oreopoulos 2010, Oreopoulos and Petronijevic 2013). Having a teacher in the first percentile of the value-added distribution (2.33 SD below the mean) is equivalent to missing $\frac{2.33 \times 1.34\%}{9\%} = 1/3$ of the school year when taught by a teacher of average quality.

A third benchmark is the cross-sectional relationship between test scores and earnings. A 1 SD increase in teacher quality raises end-of-year scores by 0.13 SD of the student test score distribution on average across grades and subjects. A 1 SD increase in student test scores, controlling for the student- and class-level characteristics X_{it}, is associated with a 12% increase in earnings at age 28 (Appendix Table 3, Column 3, Row 2). The predicted impact of a 1 SD increase in teacher VA on earnings is therefore $0.13 \times 12\% = 1.55\%$, similar to the observed impact of 1.34%.

Extensive Margin Responses and Earnings Trajectories. The increase in wage earnings comes from a combination of extensive and intensive margin responses. In Column 4 of Table 3, we regress an indicator for having positive W-2 wage earnings on teacher VA using the same specification as in Column 1. A 1 SD increase in teacher VA raises the probability of working by 0.38%. If the marginal entrant into the labor market were to take a job that paid the mean earnings level in the sample ($21,256), this extensive margin response would raise mean earnings by $81. Since the marginal entrant most likely has lower earnings than the mean, this implies that the extensive margin accounts for at most $81/350 = 23\%$ of the total earnings increase due to better teachers.

As noted above, W-2 wage earnings do not include self-employment and other potential sources of income and therefore may provide an incomplete picture of teachers' impacts. To evaluate this concern, Column 5 replicates the baseline specification in Column 1 using total earnings (as defined in Section 3) instead of wage earnings. Reassuringly, the point estimate of teachers' impacts changes relatively little with this broader income definition. We therefore use wage earnings – which provides an individual rather than household measure of earnings and is unaffected by the endogeneity of filing – for the remainder of our analysis.

Next, we analyze how teacher VA affects the trajectory of earnings by examining wage earnings impacts at each age from 20 to 28. We run separate regressions of wage earnings at each age on teacher VA using the same specification as in Column 1 of Table 3. Figure 2b plots the coefficients from these regressions (which are reported in Appendix Table 9), divided by average earnings at each age to obtain percentage impacts. The impact of teacher quality on earnings

with higher *estimated* VA yields smaller gains because of noise in $\widehat{m}_{jt}$ and drift in teacher quality, which we revisit in Section 7.

rises almost monotonically with age. At early ages, the impact of higher VA is *negative* and statistically significant, 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 than students who had lower VA teachers in grades 4-8. Earnings impacts become positive at age 23, become statistically significant at age 24, and grow through age 28, where the earnings impact reaches 1.65%, as in Figure 2a.

An alternative way to state the result in Figure 2b is that better teachers increase the growth rate of students' earnings in their 20s. In Column 6 of Table 3, we verify this result directly 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 $286 (2.5%) over this period. This finding suggests that teachers' impacts on lifetime earnings could be larger than the 1.34% impact observed at age 28.

4.4 Other Outcomes

In this subsection, we 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 3.1). 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 6.6 percentage points less likely to have a teenage birth relative to a mean of 13.4% (Appendix Table 3). The correlation is 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 qualitative proxy for teenage births even though we can only identify children who are claimed as dependents in the tax data. However, one must be cautious in interpreting the quantitative magnitudes of results using this measure, as our proxy might understate the total number of children born to teenagers.

Column 1 of Table 4 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 0.61 percentage points, a reduction of roughly 4.6%, as shown in Figure 3a. This impact is similar to the raw cross-sectional correlation between scores and teenage births, echoing our results on earnings and college attendance.

Column 2 of Table 4 analyzes the impact of teacher VA on the socio-economic status of the neighborhood in which students live at age 28, measured by the percent of college graduates living

in that neighborhood. A 1 SD increase in teacher VA raises neighborhood SES by 0.25 percentage points (1.8% of the mean) by this metric, as shown in Figure 3b.

Column 3 of Table 4 studies the effect of teacher quality on the probability of having a 401(k) at age 28. Increasing teacher VA by 1 SD increases the likelihood of saving by 0.55 percentage points (or 2.8% of the mean), as shown in Figure 3c.[30]

Fade-Out of Test Score Impacts. The final set of outcomes we consider are teachers' impacts on test scores in subsequent grades. Figure 4 plots the impacts of teacher VA on test scores in subsequent years; see Appendix Table 10 for the underlying coefficients. To construct this figure, we residualize raw test scores $A^*_{i,t+s}$ with respect to the class-level controls X_{ct} using within-teacher variation and then regress the residuals $A_{i,t+s}$ on $\widehat{\mu}_{jt}$ using all observations in the core sample. We scale teacher VA in units of student test-score SDs in these regressions – by using $\widehat{\mu}_{jt}$ as the independent variable instead of $\widehat{m}_{jt} = \widehat{\mu}_{jt}/\sigma_0$ – to facilitate interpretation of the regression coefficients, which are plotted in Figure 4. The coefficient at $s = 0$ is not statistically distinguishable from 1, as shown in our companion paper. Teachers' impacts on test scores fade out rapidly in subsequent years and appear to stabilize at approximately 25% of the initial impact after 3-4 years.[31] This result aligns with existing evidence that improvements in education raise contemporaneous scores, then fade out in later scores, only to reemerge in adulthood (Deming 2009, Heckman et al. 2010c, Chetty et al. 2011).

5 Research Design 2: Teacher Switching Quasi-Experiments

Our preceding estimates rely on the strong assumption that the unobserved determinants of students' long-term outcomes are uncorrelated with teacher quality. In this section, we estimate teachers' long-term impacts using a quasi-experimental design that relaxes this identification assumption. This research design parallels the quasi-experimental approach used to estimate the degree of bias in VA estimates in our companion paper (Chetty, Friedman, and Rockoff 2013), and the methodology described in the next subsection draws heavily from that paper.

[30] We also investigated the impacts of teacher quality on marital status, homeownership, and the probability of living out of state at age 25. Because all of these are infrequent outcomes in the sample and age group we study, we find no significant impacts of teacher quality on these measures (not reported).

[31] Prior studies (e.g., Kane and Staiger 2008, Jacob, Lefgren, and Sims 2010, Rothstein 2010, Cascio and Staiger 2012) 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. 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 5.6 million student-year-subject observations and four-year persistence using more than 1.3 million student-year-subject observations.

5.1 Methodology

Adjacent cohorts of students within a school are frequently exposed to different teachers. We exploit this teacher turnover to obtain a quasi-experimental estimate of teachers' long-term impacts. To understand our research design, consider a school with three 4th grade classrooms. Suppose one of the teachers leaves the school in 1995 and is replaced by a teacher whose VA estimate is 0.3 higher, so that the mean test-score VA of the teaching staff rises by $0.3/3 = 0.1$. If the distribution of unobserved determinants of students' long-term outcomes does not change between 1994 and 1995, the change in mean college attendance rates between the 1994 and 1995 cohorts of students will reveal the impact of a 0.1 improvement in 4th grade teachers' test-score VA. More generally, we can estimate teachers' long-term impacts by comparing the change in mean student outcomes across cohorts to the change in mean VA driven by teacher turnover provided that student quality is stable over time.[32]

To formalize this approach, let $\widehat{m}_{jt}^{-\{t,t-1\}}$ denote the test-score VA estimate for teacher j in school year t constructed as in our companion paper using data from all years except $t-1$ and t. Similarly, let $\widehat{m}_{j,t-1}^{-\{t,t-1\}}$ denote the VA estimate for teacher j in school year $t-1$ based on data from all years except $t-1$ and t. Let Q_{sgt} denote the student-weighted mean of $\widehat{m}_{jt}^{-\{t,t-1\}}$ across teachers in school s in grade g, which is the average estimated quality of teachers in a given school-grade-year cell; define $Q_{sg,t-1}$ analogously.[33] Let $\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$ denote the change in mean teacher value-added from year $t-1$ to year t in grade g in school s. Define mean changes in student outcome residuals ΔY_{sgt} analogously. Note that because we exclude both years t and $t-1$ when estimating VA, the variation in ΔQ_{sgt} is driven purely by changes in the teaching staff and not by changes in teachers' VA estimates. As above, this leave-out technique ensures that changes in ΔY_{sgt} are not spuriously correlated with ΔQ_{sgt} due to estimation error in VA.[34]

We estimate teachers' long-term impacts by regressing changes in mean outcomes across cohorts

[32] By analyzing student outcomes at the school-grade-subject level, we do *not* exploit information on classroom assignment, thus overcoming the non-random assignment of students across classrooms.

[33] In our baseline specifications, we impute teacher VA as the sample mean (0) for students for whom we have no leave-out-year VA estimate $\widehat{m}_{jt}^{-\{t,t-1\}}$, either because we have no teacher information or because the teacher did not teach in the district outside of years $\{t-1, t\}$. We show below that we obtain similar results when restricting to the subset of school-grade-subject-year cells with no missing data on teacher VA. See Section 6 of our companion paper for additional discussion on the effects of this imputation.

[34] Formally, not using a two-year leave out would immediately violate Assumption 3 below, because unobserved determinants of scores ε_{sgt} or $\varepsilon_{sg,t-1}$ would appear in ΔQ_{sgt} and ε_{sgt} is correlated with unobserved determinants of earnings ε_{sgt}^{Y}.

on changes in mean test-score VA:

$$\Delta Y_{sgt} = \alpha + \kappa \Delta Q_{sgt} + \Delta \eta'_{sgt} \tag{17}$$

The coefficient in (17) identifies the effect of a 1 SD improvement in teacher quality as defined in (14) under the following assumption.

Assumption 3 [Teacher Switching as a Quasi-Experiment] Changes in teacher quality across cohorts within a school-grade are orthogonal to changes in other determinants of student outcomes $\Delta \eta'_{sgt}$ across cohorts:

$$Cov\left(\Delta Q_{sgt}, \Delta \eta'_{sgt}\right) = 0. \tag{18}$$

This assumption could potentially be violated by endogenous student or teacher sorting. In practice, 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 spouses. Moreover, in our first paper, we present direct evidence supporting (18) by showing that both prior scores and *contemporaneous* scores in the other subject (e.g., English) are uncorrelated with changes in mean teacher quality in a given subject (e.g., math). We present additional evidence supporting (18) below.

Note that if observable characteristics X_{it} are also orthogonal to changes in teacher quality across cohorts (i.e., satisfy Assumption 3), we can implement (17) simply by regressing the change in raw outcomes ΔY^*_{sgt} on ΔQ_{sgt}. If the quasi-experiment is a good approximation to a true experiment, one would expect the controls to be balanced across cohorts as well. We therefore begin with regressions of ΔY^*_{sgt} on ΔQ_{sgt} and then confirm that changes in control variables across cohorts are uncorrelated with ΔQ_{sgt}.

5.2 Results

Figure 5a presents a binned scatter plot of changes in mean college attendance rates ΔY^*_{sgt} against changes in mean teacher value-added ΔQ_{sgt} across cohorts. We include year fixed effects (demeaning both the x and y variables by school year), so that the estimate is identified purely from differential changes in teacher value-added across school-grade-subject cells over time. The cor-

responding regression coefficient, which is based on estimating (17) with year fixed effects but no other controls, is reported in Column 1 of Table 5a.

Changes in the quality of the teaching staff have significant impacts on changes in college attendance rates across consecutive cohorts of students in a school-grade-subject cell. The null hypothesis that $\kappa = 0$ is rejected with $p < 0.01$. The point estimate implies that a 1 SD improvement in teacher quality raises college attendance rates by 0.86 percentage points, with a standard error of 0.23. This estimate is not statistically distinguishable from the estimate of 0.82% obtained in Column 1 of Table 2 using the first research design. However, as expected, the quasi-experimental estimate is much less precise because it exploits much less variation.

This analysis identifies teachers' causal impacts provided that (18) holds. One natural concern is that improvements in teacher quality may be correlated with other improvements in a school – such as better resources in other dimensions – that also contribute to students' long-term success and thus lead us to overstate teachers' true impacts. To address this concern, Column 2 of Table 5a replicates the baseline specification in Column 1 including school by year fixed effects instead of just year effects. In this specification, the only source of identifying variation comes from differential changes in teacher quality across subjects and grades *within* a school in a given year. The coefficient on ΔQ_{sgt} changes very little relative to the baseline estimate that pools all sources of variation. Column 3 further accounts for secular trends in subject- or grade-specific quality by controlling for the change in mean teacher VA in the prior and subsequent year as well as cubics in the change in prior-year mean own-subject and other-subject scores across cohorts. Controlling for these variables has little impact on the estimate. This result shows that fluctuations in teacher quality relative to trend in specific grades generate significant changes in the affected students' college attendance rates.

In the preceding specifications, we imputed the sample mean of VA (0) for classrooms for which we could not calculate actual VA. This generates downward bias in our estimates because we mismeasure the change in teacher quality ΔQ_{sgt} across cohorts. Column 4 of Table 5a replicates Column 2, limiting the sample to school-grade-subject-year cells in which we can calculate the leave-two-year-out mean for all teachers in the current and preceding year. As expected, the point estimate of a one SD increase in teacher quality increases, but the confidence interval is significantly wider because the sample size is considerably smaller.

Finally, we further evaluate (18) using a series of placebo tests. In Column 5 on Table 5a, we replicate Column 2, replacing the change in actual college attendance with the change in predicted

college attendance based on parent characteristics. We predict college attendance using an OLS regression of Y_{it}^* on the same five parent characteristics P_{it}^* used in Section 4.2, with no other control variables. Changes in mean teacher VA have no effect on predicted college attendance rates, supporting the assumption that changes in the quality of the teaching staff are unrelated to changes in student quality at an annual level.

In Figure 6a, we present an alternative set of placebo tests based on the sharp timing of the change in teacher quality. To construct this figure, we replicate the specification in Column 1 but include changes in mean teacher VA for the four preceding and subsequent cohorts as placebo effects (as well as year fixed effects):

$$\Delta Y_{sgt}^* = \alpha_t + \sum_{n=-4}^{4} \kappa_n \Delta Q_{sg,t+n} + \varphi_t \tag{19}$$

Figure 6a plots the vector of coefficients $\overrightarrow{\kappa} = (\kappa_{-4}, ..., \kappa_0, ..., \kappa_4)$, which represent the impacts of changes in the quality of teaching staff at different horizons on changes in college attendance rates at time 0. As one would expect, κ_0 is positive and highly significant while all the other coefficients are near 0 and statistically insignificant. That is, contemporaneous changes in teacher quality have significant effects on college attendance rates, but past or future changes have no impact, as they do not directly affect the current cohort of students. This figure – which is analogous to an event study based on teacher entry and exit – strongly supports the view that the changes in college attendance rates documented above reflect teachers' casual effects.[35]

Figures 5b and 6b and Table 5b replicate the preceding analysis using the earnings-based index of college quality as the outcome. Consistent with the preceding results, we find that improvements in teachers' test score VA across cohorts lead to sharp changes in the quality of colleges that students attend. This result is robust across the specifications in Columns 1-4 of Table 5 described above. We find no evidence that predicted college quality based on parent characteristics is correlated with changes in teacher quality. In addition, changes in teacher VA again affect college quality in the year of the change rather than in preceding or subsequent years, as shown in Figure 6b. We conclude based on this evidence that students who happen to be in a cohort in their school that is

[35] In Figure 3 of our companion paper, we directly use event studies around the entry and exit of teachers in the top and bottom 5% to demonstrate the impacts of VA on test scores. We do not have adequate power to identify the impacts of these exeptional teachers on college attendance using such event studies. In the cross-cohort regression that pools all teaching staff changes, the t-statistic for college attendance is 3.78 (Column 1 of Table 5a in this paper). The corresponding t-statistic for test scores is 34.0 (Column 2 of Table 5 of the first paper). We have much less power here both because the college attendance is only observed for the older half of our sample and because college is a much noisier outcome than end-of-grade test scores.

taught by higher VA teachers are significantly more likely to go to college and attend higher ranked colleges.

We used specifications analogous to those in Table 5 to investigate the impacts of teaching quality on other outcomes, including earnings at age 28. Unfortunately, our sample size for earnings at age 28 is roughly 1/7th the size of the sample available to study college attendance at age 20. This is both because we have fewer cohorts of students who are currently old enough to be observed at age 28 in the tax data and because we have data on teacher assignments for much fewer schools in the very early years of our school district data. Because of the considerably smaller sample, we obtain very imprecise and fragile estimates of the impacts of teacher quality on earnings using the quasi-experimental design.[36] While we cannot obtain quasi-experimental estimates of the impacts of teacher quality on earnings, the close alignment between the quasi-experimental and cross-class OLS regression estimates for college outcomes validates the selection on observables assumption underlying our first research design. Given that cross-class comparisons conditional on observables provide accurate forecasts of teachers' impacts on test scores, college attendance, and college quality, we would expect the same to be true of earnings impacts as well.

6 Heterogeneity of Teachers' Impacts

In this section, we analyze whether teachers' impacts vary across demographic groups, subjects, and grades. Because analyzing subgroup heterogeneity requires considerable statistical precision, we use the first research design – comparisons across classrooms conditional on observables. We analyze teachers' impacts on college quality at age 20 (rather than earnings at age 28) to maximize precision and obtain a quantitative metric based on projected earnings gains.

6.1 Demographic Groups

In Panel A of Table 6, we study the heterogeneity of teachers' impacts across demographic subgroups. Each value reported in the first row of the table is a coefficient estimate from a separate regression of college quality on teacher VA conditional on controls. To be conservative, we include both student characteristics X_{ct} and parent characteristics P^*_{ct} in the control vector throughout this section and estimate specifications analogous to Column 5 of Table 2 on various subsamples. Columns 1 and 2 consider heterogeneity by gender. Columns 3 and 4 consider heterogeneity by

[36] For instance, estimating the specification in Column 1 of Table 5 with earnings at age 28 as the dependent variable yields a confidence interval of (-$581,$665), which contains both 0 and values nearly twice as large as the estimated earnings impacts based on our first research design.

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 Table 6a. First, the point estimates of the impacts of teacher VA are larger for females than males, although we cannot reject equality of the impacts ($p = 0.102$). 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 \$190 for children whose parents have below-median income, compared with \$380 for those whose parents have above-median income. However, the impacts are more similar as a percentage of mean college quality: 0.80% for low-income students vs. 1.25% for high-income students.

The larger absolute 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 6a shows that a 1 SD increase in teacher VA raises test scores by approximately 0.13 SD on average in all the subgroups, consistent with the findings of Lockwood and McCaffrey (2009). In contrast, the cross-sectional 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.[37] Overall, the heterogeneity in treatment effects on college quality indicates that teacher quality is complementary to family inputs and resources, i.e. the marginal effect of better teaching is *larger* for students from high SES families. This result implies that higher income families should be willing to pay more for teacher quality.

6.2 Subjects: Math vs. English

Panel B of Table 6 analyzes differences in teachers' impacts across subjects. For these regressions, we split the sample into elementary (Columns 1-3) and middle (Columns 4-5) schools. This distinction is important because students have the same teacher for both subjects in elementary school but not middle school.

In Column 1, we replicate the baseline specification in Column 5 of Table 2, restricting the sample to math classrooms in elementary school. Column 2 repeats this specification for English. In Column 3, we include each teacher's math and English VA together in the same specification,

[37] Importantly, the relationship between college quality and test scores conditional on prior characteristics X_{it} is linear throughout the test score distribution (Appendix Figure 2b). Hence, the heterogeneity is not due to non-linearities in the relationship between scores and college outcomes but rather the fact that the same increase in scores translates to a bigger change in college outcomes for high SES families.

reshaping the dataset to have one row for each student-year (rather than one row per student-subject-year, as in previous regressions). Because a given teacher's math and English VA are highly correlated ($r = 0.6$), the magnitude of the two subject-specific coefficients drops by an average of 40% when included together in a single regression for elementary school. 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.

We find that a 1 SD increase in teacher VA in English has larger impacts on college quality than a 1 SD improvement in teacher VA in math. This is despite the fact that the variance of teacher effects in terms of test scores is *larger* in math than English. In Table 2 of our companion paper, we estimated that the standard deviation of teacher effects on student test scores in elementary school is 0.124 in English and 0.163 in math. Using the estimates from Column 3 of Table 6b, this implies that an English teacher who raises her students' test scores by 1 SD raises college quality by $\frac{189/0.124}{106/0.163} = 2.3$ times as much as a math teacher who generates a commensurate test score gain. Hence, the returns to better performance in English are especially large, although it is much harder for teachers to improve students' achievement in English (e.g., Hanushek and Rivkin 2010, Kane et al. 2013).

We find a similar pattern in middle school. In Column 4 of Table 6b, we replicate the baseline specification for the subset of observations in math in middle school. We control for teacher VA in English when estimating this specification by residualizing college quality Y_{it}^* with respect to the student and parent class-level control vectors X_{ct} and P_{it}^* as well as $\widehat{m}_{jt}$ in English using a regression with math teacher fixed effects as in (5). Column 5 of Table 6b replicates the same regression for observations in English in middle school, controlling for math teacher VA. A 1 SD improvement in English teacher quality raises college quality by roughly twice as much as a 1 SD improvement in math teacher quality. We conclude that even though teachers have much smaller impacts on English test scores than math test scores, the small improvements that good teachers generate in English are associated with substantial long-term impacts.

6.3 Impacts of Teachers by Grade

We estimate the impact of a 1 SD improvement in teacher quality in each grade $g \in [4, 8]$ on college quality (κ_g) by estimating the specification in Column 5 of Table 2 but interacting $\widehat{m}_{jt}$ with grade dummies. 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 in this subsection.[38]

The series in circles in Figure 7 plots the estimates of κ_g, which are also reported in Appendix Table 11. We find that 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 0.6 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. Hence, the fact that high-VA math and English teachers continue to have substantial impacts even in middle school strongly suggests that higher quality education has substantial returns well beyond early childhood.

Tracking and Net Impacts. The reduced-form estimates of κ_g reported above include the impacts of being tracked to a better teacher in subsequent grades, as shown in (10). While a parent may be interested in the reduced-form impact of teacher VA in grade g, a policy reform that raises teacher quality in grade g will not allow every child to get a better teacher in grade $g+1$. We now turn to identifying teachers' net impacts $\tilde{\kappa}_g$ in each grade, holding fixed future teachers' test-score VA.

Because we have no data after grade 8, we can only estimate teachers' net effects holding fixed teacher quality up to grade 8.[39] We therefore set $\tilde{\kappa}_8 = \kappa_8$. We recover $\tilde{\kappa}_g$ from estimates of κ_g by subtracting out the impacts of future teachers on earnings iteratively. The net impact of a 7th grade teacher is her reduced-form impact κ_7 minus her indirect impact via tracking to a better 8th grade teacher:

$$\tilde{\kappa}_7 = \kappa_7 - \rho_{78}\tilde{\kappa}_8, \tag{20}$$

[38] Restricting the sample in the same way does not affect the conclusions above about heterogeneity across subjects or demographic groups, because these groups are balanced across cohorts.

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

where ρ_{78} is the extent to which teacher VA in grade 7 increases teacher VA in grade 8 conditional on controls. We can identify ρ_{78} using an OLS regression that parallels (16) with future teacher VA as the dependent variable:

$$\widehat{m}_{j,t_i(8)} = \alpha + \rho_{78}\widehat{m}_{j,t_i(7)} + \gamma_1 X_{ct} + \gamma_2 P^*_{ct} + \eta^{\mu}_{ct78}. \tag{21}$$

As above, we estimate ρ_{78} in two steps. First, we residualize $\widehat{m}_{j,t_i(8)}$ with respect to the controls by regressing $\widehat{m}_{j,t_i(8)}$ on X_{ct} and P^*_{ct} with grade 7 teacher fixed effects, as in (5). We then run a univariate OLS regression of the residuals on $\widehat{m}_{j,t_i(7)}$ to estimate $\hat{\rho}_{78}$. We apply (20) to identify $\tilde{\kappa}_7$ from the reduced-form estimates of κ_g in Figure 7. Iterating backwards, we can calculate κ_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 $\tilde{\kappa}_g$ in Appendix C.

The series in triangles in Figure 7 plots the estimates of the net impacts $\tilde{\kappa}_g$. The net impacts are very similar to the reduced-form impacts because the tracking coefficients $\rho_{g,g'}$ are generally quite small, as shown in Appendix Table 12. Tracking is larger in middle school, as one would expect, but still has a relatively modest impact on $\tilde{\kappa}_g$.

These results suggest that the reduced-form estimates reported above largely reflect a teacher's own direct impact rather than the impacts of being tracked to better teachers in later grades. However, we caution that this approach to calculating teachers' net impacts has three important limitations. First, it assumes that all tracking to future teachers occurs exclusively via teachers' test-score VA. We allow students who have high-VA teachers in grade g to be tracked to higher test-score VA (m_{jt}) teachers in grade $g+1$, but *not* to teachers with higher total earnings VA μ^Y_{jt}. 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, $\tilde{\kappa}_g$ does not net out potential changes in other factors besides teachers, such as peer quality or parental inputs. Hence, $\tilde{\kappa}_g$ 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 a child has affects her outcomes in adulthood.

7 Policy Analysis

In this section, we use our estimates to predict the potential earnings gains from selecting and retaining teachers on the basis of their VA. We focus on earnings gains because they are easily quantifiable; however, improvements in teacher quality may have non-monetary returns as well (Oreopoulos and Salvanes 2010), as suggested by our findings on teenage birth rates and neighborhood quality.

We make four assumptions in our calculations. First, we assume that the percentage impact of a 1 SD improvement in teacher VA on earnings observed at age 28 is constant at $b = 1.34\%$ (Table 3, Column 2) over the lifecycle.[40] Second, we ignore general equilibrium effects that may reduce wage rates if all children are better educated. 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 $\$522,000$ in 2010 dollars, as noted above. Finally, we assume that teacher VA m_{jt} is normally distributed.

To quantify the value of improving teacher quality, we evaluate Hanushek's (2009, 2011) proposal to replace teachers whose VA ratings are in the bottom 5 percent of the distribution with teachers of average quality. To simplify exposition, we calculate these impacts for elementary school teachers, who teach one class per day. We first calculate the earnings gains from selecting teachers based on their true (unobserved) test-score VA m_{jt} and then calculate the feasible gains from selecting teachers based on VA estimates $\widehat{m}_{jt}$.

Selection on True VA. Elementary school teachers teach both math and English and therefore have two separate VA measures on which they could be evaluated. First consider the simple case in which teachers are evaluated based purely on their VA in one subject (say math) and VA in the other subject is discarded. Consider a student whose teacher's true math VA is Δm_σ standard deviations below the mean. Replacing this teacher with a teacher of mean quality (for a single school year) would raise the student's expected earnings by

$$G = \Delta m_\sigma \times \$522,000 \times b. \tag{22}$$

Under the assumption that m_{jt} is normally distributed, a teacher in the bottom 5% of the true VA

[40] We have inadequate precision to estimate wage earnings impacts separately by subject and grade level. To obtain a rough estimate, we therefore assume that a 1 SD improvement in teacher VA raises earnings by 1.34% in all subjects and grade levels in the calculations that follow.

distribution is on average 2.063 standard deviations below the mean teacher quality. Therefore, replacing a teacher in the bottom 5% of math (or English) VA with an average teacher generates a present value lifetime earnings gain per student of

$$G = \$522,000 \times 2.063 \times 1.34\% = \$14,500.$$

For a class of average size (28.2), the total NPV earnings impact from this replacement is $G_C = \$407,000$. The undiscounted cumulative lifetime earnings gains from deselection are 5.5 times larger than these present value gains ($80,000 per student and $2.25 million per classroom), as shown in Appendix Table 13.[41] These simple calculations show that the potential gains from improving the quality of teaching – whether using selection based on VA, teacher training, or other policy tools – are quite large.

The preceding approach discards information because it rates teachers only on the basis of one subject, whereas we typically have two VA ratings per teacher. School districts that use VA for evaluation purposes typically average math and English VA ratings to calculate a single measure of teacher performance in elementary schools (e.g., District of Columbia Public Schools 2012). To simulate such a policy, suppose we deselect the 5% of elementary school teachers with the lowest mean standardized VA across subjects. Simulating a bivariate normal distribution with a within-year correlation between $\widehat{m}_{jt}$ across math and English of $r = 0.6$, we calculate that teachers whose mean VA across subjects is in the bottom 5% have a standardized VA that is $\Delta m_\sigma = 1.84$ SD below the mean in both math and English.

To calculate the long-term earnings impact of replacing such teachers, we must identify the impacts of changes in VA in one subject holding fixed VA in the other subject. Given the between-subject VA correlation of $r = 0.6$, our earnings impact estimate of $b = 1.34\%$ reflects the effect of a 1 SD improvement in a given subject (e.g. math) combined with a 0.6 SD improvement in the other subject (English). Under the simplifying assumption that earnings impacts do not vary across subjects, the impact of a 1 SD improvement in VA in a given subject is $b_s = \frac{b}{1+0.6} = 0.84\%$. Therefore, replacing a teacher with mean VA in the bottom 5% with an average teacher for one

[41] These calculations do not account for the fact that deselected teachers may be replaced by rookie teachers, who have lower VA. Mean test score residuals for students taught by first-year teachers are on average 0.05 lower (in units of standardized student test scores) than those taught by more experienced teachers. Given that the median teacher remains in the district for approximately 10 years, accounting for the effect of inexperience in the first year would reduce the expected benefits of deselection over a typical horizon by $\frac{0.05/10}{2.063 \times \sigma(m_{jt})} = 2\%$, where $\sigma(m_{jt}) = 0.14$ is the mean SD of teacher effects across elementary school subjects in our data (see Table 2 of the first paper).

school year in elementary school increases the present value of a student's earnings by

$$G' = \$522,000 \times 2 \times 1.84 \times 0.84\% = \$16,100$$

and yields total gains of $\$454,000$ for an average-sized classroom. The gains from evaluating teachers based on mean math and English VA are only 12% larger than the gains from using information based on only one subject because math and English VA estimates are quite highly correlated. Therefore, we focus on the case in which the teacher is rated based on VA in only one subject in what follows.

Selection on Estimated VA. In practice, we cannot observe actual VA m_{jt} and therefore can only select teachers on the basis of estimated VA $\widehat{m}_{jt}$. This reduces the gains from selection for two reasons: (1) estimation error in VA and (2) drift in teacher quality over time. To quantify the impact of these realities, suppose we use test score data from years $t = 1, ..., n$ to estimate teacher VA in school year $n + 1$. The gain in year $n + 1$ from replacing the bottom 5% of teachers based on VA estimated using the preceding n years of data is

$$G(n) = -\mathbb{E}\left[m_{j,n+1} \mid \widehat{m}_{j,n+1} < F^{-1}_{\widehat{m}_{j,n+1}}(0.05)\right] \times \$522,000 \times b, \tag{23}$$

where $\mathbb{E}\left[m_{j,n+1} \mid \widehat{m}_{j,n+1} < F^{-1}_{\widehat{m}_{j,n+1}}(0.05)\right]$ denotes the expected value of $m_{j,n+1}$ conditional on the teacher's estimated VA falling below the 5th percentile. We calculate this expected value separately for math and English using Monte Carlo simulations of a Multivariate Normal distribution as follows.[42] First, we construct Σ_A, the VCV matrix of $\overrightarrow{A}_j^{-t}$, the vector of past class average scores, using the parameters of the autocovariance vector of test scores reported in Columns 1 and 2 of Table 2 of our companion paper.[43] We then simulate draws of average class scores from a $N(0, \Sigma_A)$ distribution for one million teachers and calculate $\widehat{m}_{j,n+1}$ based on scores from the first n periods using the same method used to construct the VA estimates in our companion paper. Finally, we calculate the conditional expectation in (23) as the mean test score in year $n + 1$ for teachers with $\widehat{m}_{j,n+1}$ in the bottom 5% of the distribution.

Figure 8a plots the mean gain per classroom $G_C(n) = 28.2 \times G(n)$, averaging over math and

[42] Without drift, the formula in (23) reduces to $r(n)^{1/2} \times 2.063 \times \$522,000 \times b$, where $r(n)$ denotes the reliability of the VA estimate using n years of data, which is straightforward to calculate analytically. The working paper version of our study (Chetty, Friedman, and Rockoff 2011b) used this version of the formula and an estimate of b based on a model that did not account for drift.

[43] We define the off-diagonal elements of Σ_A directly based on the autocovariances σ_{As} reported in Table 2 of our first paper, setting the autocovariance $\sigma_{As} = \sigma_{A7}$ for $s > 7$. We define the diagonal elements of Σ_A as the variance of mean class test scores, which we compute based on the estimates in Table 2 as (Class+Teacher Level SD)^2 + (Individual-Level SD)^2/28.2, where 28.2 is the average number of students per class.

English, for $n = 1, ..., 10$. The values underlying this figure are reported in Appendix Table 13. The gain from deselecting teachers based on true VA, $G_C = \$407,000$, is shown by the horizontal line in the figure. The gains from deselecting teachers based on estimated VA are significantly smaller because of noise in VA estimates and drift in teacher quality.[44] With one year of data, the expected gain per class is \$226,000, 56% of the gain from selecting on true VA. The gains grow fairly rapidly with more data in the first 3 years, but the marginal value of additional information is small. With three years of test score data, the gain is \$266,000, but the gain increases to only \$279,000 after 10 years. After three years, waiting for one more year would increase the gain by \$4,000 but has an expected cost of \$266,000. 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). Adding data from prior classes yields relatively little information about current teacher quality both because of decreasing returns to additional observations and drift.

The calculations above assume that VA estimates have zero forecast bias. While the estimates in our first paper do not reject this hypothesis, the upper bound on the 95% confidence interval for our quasi-experimental estimate of forecast bias is 9%, which would imply $\mathbb{E}\left[m_{j,n+1} \mid \widehat{m}_{j,n+1}\right] = 0.91\widehat{m}_{j,n+1}$. This degree of forecast bias has modest impacts on the gains from deselection: for instance, the earnings gains per class in year 4 based on 3 years of test score data are $G_C(3) = \$242,000$.[45]

Drift in Quality over Subsequent School Years. The values in Figure 8a reflect the gains in the first year after the deselection of teachers, based on $\widehat{m}_{j,n+1}$ in school year $n+1$. Now consider the impacts of such a policy on the earnings of students in a subsequent school year $n+m$:

$$G(m,n) = -\mathbb{E}\left[m_{j,n+m} \mid \widehat{m}_{j,n+1} < F^{-1}_{\widehat{m}_{j,n+1}}(0.05)\right] \times \$522,000 \times b,$$

where $\mathbb{E}\left[m_{j,n+m} \mid \widehat{m}_{j,n+1} < F^{-1}_{\widehat{m}_{j,n+1}}(0.05)\right]$ denotes the mean VA of teachers in year $n+m$ conditional on having estimated VA in year $n+1$ below the 5th percentile. We calculate this expected value using the same Monte Carlo simulation as above.

[44] In Panel B of Appendix Table 13, we distinguish these two factors by eliminating estimation error and predicting current VA based on past VA instead of past scores. Without estimation error, $G_C(1) = \$340,000$. Hence, drift and estimation error each account for roughly half of the difference between $G_C(1)$ and G_C.

[45] We also replicated the simulations using VA estimates that do not account for drift. When the estimation window n is short, drift has little impact on the weights placed on test scores across years. As a result, drift-unadjusted measures yield rankings of teacher quality that are very highly correlated with our measures and thus produce similar gains. For instance, selection based on 3 years of data using VA estimates that do not adjust for drift yields gains that are 98% as large as those reported above. Hence, while accounting for drift is important for evaluating out-of-sample forecasts accurately, it may not be critical for practical policy applications for VA.

The lower series in Figure 8b plots $G_C(m,3) = 28.2 \times G(m,3)$, the per-class gains in school year m from deselecting teachers based on their estimated VA for year 4 ($\widehat{m}_{j,4}$), constructed using the first 3 years of data. The first point in this series coincides with the value of \$266,000 in Figure 8a reported for $n = 3$. Because teacher quality drifts over time, the gains fall in subsequent school years, as some of the teachers who were deselected based on their predicted VA in school year n would have reverted toward the mean in subsequent years. Deselection based on VA estimates at the end of year 3 generates an average gain of \$184,000 per classroom per year over the subsequent 10 years, the median survival time in the district for teachers who have taught for 3 years.[46]

The upper series in Figure 8b plots the analogous gains when teachers are deselected based on their true VA $m_{j,4}$ in year 4 instead of their estimated VA $\widehat{m}_{j,4}$.[47] The first point in this series coincides with the maximum attainable gain of \$407,000 shown in Figure 8a. The gains again diminish over time because of drift in teacher quality. The average present value gain from deselection based on true VA over the subsequent ten years is approximately \$250,000 per classroom. This corresponds to an undiscounted lifetime earnings gain per classroom of students of approximately \$1.4 million.

We conclude that the potential gains from selecting teachers based on VA remain substantial even when estimation error and drift are taken into account. However, because VA estimates are imperfect predictors of m_{jt}, there is substantial room to use other measures of quality – such as principal evaluations or student surveys – to complement VA estimates.[48] What is clear from these calculations is that improving teacher quality is likely to yield substantial returns for students; the best way to accomplish that goal is less clear.

Costs of Teacher Selection. The calculations above do not account for the costs associated with a policy that deselects teachers with the lowest estimated performance ratings. First, they ignore downstream costs that may be required to generate earnings gains, most notably the cost

[46] If one's goal is to maximize expected gains over a teacher's tenure, one should ideally deselect teachers after n years based on mean predicted VA over all future years, discounted by the survival probabilities. We find that this more complex policy increases gains by less than 1% over 10 years. Intuitively, because the VA drift process is close to an AR(1) process, the relative weights on average scores from a teacher's first three years do not change much when projecting beyond year 4.

[47] We calculate these gains using a Monte Carlo simulation analogous to that above, except that we draw scores from the VCV matrix of true VA Σ_μ instead of test scores Σ_A. The off-diagonal elements of the two matrices are identical, but the diagonal elements of Σ_μ reflect only the variance of teacher quality σ^2_μ. We use the quadratic estimates of σ_μ reported in the last row of Table 2 in our companion paper for this simulation.

[48] These other measures will also be affected by drift and estimation error. For instance, classroom observations have significant noise and may capture transitory fluctuations in teacher quality (Kane et al. 2013). More generally, issues of drift and estimation error are not unique to the teaching profession. Applying the techniques here to quantify the impacts of estimation error and drift on personnel evaluation in various professions would be a useful direction for future research.

associated with higher college attendance rates. Second, and more importantly, they ignore the fact that teachers need to be compensated for the added employment risk they face from such an evaluation system. Rothstein (2013) estimates the latter cost using a structural model of the labor market for teachers. Rothstein estimates that a policy that fires teachers if their estimated VA after 3 years falls below the 5th percentile would require a mean salary increase of 1.4% to equilibrate the teacher labor market.[49] In our sample, mean teacher salaries were approximately $50,000, implying that annual salaries would have to be raised by approximately $700 for all teachers to compensate them for the additional risk. Based on our calculations above, the deselection policy would generate NPV gains of $184,000 per teacher deselected, or $9,250 for all teachers on average (because only one out of twenty teachers would actually be deselected). Hence, the estimated gains from this policy are more than 10 times larger than the costs.

Together with the results in this paper, Rothstein's (2013) findings imply that deselecting low VA teachers could be a very cost effective policy. However, as Rothstein emphasizes, this conclusion assumes that the signal quality of VA measures for long-term impacts remains unchanged when it is used to evaluate teachers. If erosion in the signal quality of VA measures is substantial, the gains from selection could be eliminated and one would need to turn to other measures to identify high quality teachers.[50]

Retention of High VA Teachers. An alternative approach to improving teacher quality that may impose lower costs on teachers is to increase the retention of high-VA teachers by paying them bonuses. Using Monte Carlo simulations analogous to those above, we estimate that retaining a teacher at the 95th percentile of the estimated VA distribution (using 3 years of data) for an extra year would yield present value earnings gains in the subsequent school year of $\$522,000 \times 28.2 \times 1.34\% \times \mathbb{E}\left[m_{j,n+1} | \widehat{m}_{j,n+1} = F^{-1}_{\widehat{m}_{j,n+1}}(0.95)\right] = \$212,000$. In our data, roughly 9% of teachers in their third year do not return to the school district for a fourth year.[51] 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

[49] In the working paper version of his study, Rothstein calculates the wage gains needed to compensate teachers for a policy that deselects teachers below the 20th percentile after 2 years. Jesse Rothstein kindly provided the corresponding estimates for the policy analyzed here in personal correspondence.

[50] In practice, one need not switch to evaluation based purely on VA measures. School districts typically use VA metrics in conjunction with other measures of performance. If the signal quality of VA is a continuous function of its weight in evaluation decisions, one would optimally place some non-zero weight on VA, because the net gains would fall from the initial level of $184,000 in proportion to the weight on VA.

[51] The rate of attrition bears little or no relation to VA, consistent with the findings of Boyd et al. (2008).

benefit of $3,180. The expected benefit of offering a bonus to even an excellent (95th percentile) teacher is only modestly larger than the cost because one must pay bonuses to $(100-9)/1.5 \approx 60$ additional teachers for every extra teacher retained.

Replacing ineffective teachers is more cost-effective than attempting to retain high VA teachers because most teachers stay for the following school year and are relatively inelastic to salary increases. Of course, increasing the salaries of high VA teachers could attract more talented individuals into teaching to begin with. The preceding calculations, which focus on the stock of current teachers, do not account for this potentially important benefit.[52]

8 Conclusion

Our first paper (Chetty, Friedman, and Rockoff 2013) showed that existing test-score value-added measures are a good proxy for a teacher's ability to raise students' test scores. This paper has shown that the same VA measures are also an informative proxy for teachers' long-term impacts. Although these findings are encouraging for the use of value-added metrics, two important issues must be resolved before one can determine how VA should be used for policy.

First, using VA measures to evaluate teachers could induce responses such as teaching to the test or cheating, eroding the signal in VA measures (e.g., Jacob 2005, Neal and Schanzenbach 2010).[53] One can estimate the magnitude of such effects by replicating the analysis in this paper in a district that evaluates teachers based on their VA. If behavioral responses substantially reduce the signal quality of VA, policy makers may need to develop metrics that are more robust to such responses, as in Barlevy and Neal (2012). For instance, districts may also be able to use data on the persistence of test score gains to identify test manipulation and develop a more robust estimate of teacher quality, as in Jacob and Levitt (2003).

Second, one should compare the long-term impacts of evaluating teachers on the basis of VA to other metrics, such as principal evaluations or classroom observation. One can adapt the methods developed in this paper to evaluate these other measures of teacher quality. When a teacher who is rated highly by principals enters a school, do subsequent cohorts of students have higher college attendance rates and earnings? What fraction of a teacher's long-term impact is captured by test-score VA vs. other measures of teacher quality? By answering these questions, one could

[52] Increasing salaries or paying bonuses based on VA could also increase teacher effort. The evidence on the importance of this margin is mixed (Springer et al. 2010, Imberman and Lovenheim 2012).

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

ultimately estimate the optimal weighting of available metrics to identify teachers who are most successful in improving students' long-term outcomes.

More generally, there are many aspects of teachers' long-term impacts that remain to be explored and would be helpful in designing education policy. For example, in this paper we only identified the impact of a single teacher on long-term outcomes. Are teachers' impacts additive over time? Do good teachers complement or substitute for each other across years? Similarly, it would be useful to go beyond the mean treatment effects that we have estimated here and determine whether some teachers are especially effective in improving lower-tail outcomes or producing stars.

Whether or not VA is ultimately used as a policy tool, our results show that parents should 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 mean who is contemplating leaving a school. Each child would gain approximately $39,000 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 $200,000 ($7,000 per parent) to stay and teach their children during the next school year. Our analysis of teacher entry and exit demonstrates that retaining such a high-VA teacher would improve students' outcomes. Hence, the most important lesson of this study is that improving the quality of teaching – whether via the use of value-added metrics or other policy levers – is likely to have substantial economic and social benefits.

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. 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.

3. Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* 102 (5): 1805-1831.

4. Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. "Who Leaves? Teacher Attrition and Student Achievement." NBER Working Paper 14022.

5. 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.

6. 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.

7. Cascio, Elizabeth and Douglas Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper 18038.

8. Chamberlain 2013. 2013. "Predictive Effects of Teachers and Schools on Test Scores, College Attendance, and Earnings." Unpublished Harvard Working Paper.

9. 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.

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

11. Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011b. "The Long Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper 17699.

12. Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2013. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." Harvard University Working Paper.

13. 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.

14. 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.

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

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

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

18. District of Columbia Public Schools. 2012. "IMPACT: The District of Columbia Public Schools Effectiveness Assessment System for School-Based Personnel."

19. 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.

20. 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.

21. Gunderson, Morley K. and Philip Oreopoulos. 2010. "Returns to Education in Developed Countries ", in *International Encyclopedia of Education, 3rd edition* (edited by E. Barker, M. McGaw and P. Peterson), Elsevier Publishers, USA.

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

23. 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.

24. Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations About Using Value-Added Measures of Teaching Quality." *American Economic Review Papers and Proceedings* 100(2): 267-71.

25. 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.

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

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

28. 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.

29. 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.

30. 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.

31. Imberman, Scott A. and Michael F. Lovenheim. 2012. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System "NBER Working Paper No. 18439.

32. 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.

33. 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.

34. 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.

35. 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.

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., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Seattle, WA: Bill & Melinda Gates Foundation.

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

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

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

41. 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.

42. Oreopoulos, Philip and Uros Petronijevic. 2013. "Making College Worth It: A Review of Research on the Returns to Higher Education," NBER Working Paper 19053.

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

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

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

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

47. Rothstein, Jesse. 2013. "Teacher Quality Policy When Supply Matters," UC-Berkeley mimeo.

48. 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.

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

50. 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.

51. 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.

Online Appendix A: Earnings-Based College Quality Index

Our index of college quality is based on the average earnings of the individuals who attend each college. The construction of such an index requires several choices, including (1) the age at which college attendance is measured, (2) the age at which earnings are measured, (3) the cohort of students used, and (4) the definition of earnings. In this appendix, we assess the stability of rankings of colleges with respect to these four choices.

We begin by constructing measures of college quality that vary the four parameters above. In each case, we first identify all individuals who are U.S. citizens as of February 19, 2013 to remove those who were temporarily in the United States for college and for whom we do not have post-college earnings data.[54] We group individuals by the higher education institution they attended and by age of attendance, as measured on December 31 of each year.[55] We group individuals not enrolled at a higher education institution at a given age (i.e., those who have no 1098-T form filed on their behalf during the tax year) in a separate "no college" category. For each college (including the "no college" group), we then compute earnings of the students at various ages (in real 2010 dollars). We begin by defining earnings based on individual W-2 wage earnings and then consider broader income measures. We top code individual earnings at $10 million to reduce the influence of outliers and we include only those who are alive at the age at which we measure earnings.

We first evaluate the stability of rankings of college quality with respect to the age at which we measure earnings. Appendix Figure 1a plots the percentile ranking of colleges based on earnings measured at age 23 (one-year after most students graduate from 4 year colleges) and age 27 (five-years post-college) vs. the oldest age at which we can measure earnings of college graduates in our sample, which is 32 (ten-years post-college). We hold the age of college attendance constant at 20 and focus on the cohort of students born in 1979. To construct this figure, we bin colleges into 100 percentiles based on their ranking using age 32 earnings (without any weighting) and compute the mean percentile ranking based on earnings at age 23 and 27 within each bin. Rankings at age 27 are very well aligned with rankings at age 32, but rankings at age 23 are very poorly aligned with measures based on older data. Colleges that have the highest-earning graduates at age 32 are commonly ranked average or even below-average based on data at age 23.

In Appendix Figure 1b, we extend this analysis to cover all ages from 23-32. This figure plots the rank correlation between college quality measured at age 32 with college quality measured using earnings at earlier ages. Each point shows the correlation of an earnings-based percentile ranking at a given age with the ranking based on earnings at age 32. The correlation is very low at age 23 and rises steeply at first before asymptoting to 1. At age 28 and after, the correlations are all above 0.95, implying that we obtain similar rankings irrespective of what age one uses to measure earnings of college graduates beyond this point. The stability of the index starting in the late 20's is consistent with evidence from other studies that annual earnings starting in the late 20's are quite highly correlated with lifetime earnings (Haider and Solon 2006).

Panel A of Appendix Table 2 presents the rank correlations to corresponding to Appendix Figure 1b. The rest of Appendix Table 2 studies the rank correlations between college quality measures as we vary the other parameters. In Panel B, we vary the age at which we measure college attendance from 18 to 25, holding fixed the age of earnings measurement at 30 for the cohort born in 1981 (which is the oldest cohort for which we can measure college attendance at age 18). When we measure attendance between 18 and 22, college quality rankings are very highly

[54] Only current citizenship status is recorded in the database. As a result, the date at which we determine citizenship is simply the date we accessed the data.

[55] We include the small fraction of students who attend more than one college in a single year in the determination of college quality for each unique institution to which they are matched.

correlated with each other. The correlations begin to fall when we measure attendance at later ages. This is intuitive, as ages 18-22 correspond to those at which most students would attend a 4-year college if they proceed through school at a normal pace.

Panel C of Appendix Table 2 varies the cohort of students included in the measure, including students born between 1979 and 1981. We hold fixed the ages of college attendance and earnings measurement at 20 and 30, respectively. The measures are very highly correlated across cohorts, showing that the reliability of our index of college quality is quite high.

Finally, Panel D of Appendix Table 2 shows the relationship between college quality measures based on alternative definitions of earnings. In addition to our baseline measure of mean W-2 earnings, we consider median W-2 earnings and mean total income (W-2 wages plus self-employment income from Form 1040). In each case we hold fixed age in college at 20, age of earnings measurement at 30, and focus on the 1979 cohort. The correlation between these measures exceeds 0.94, showing that the rankings are not sensitive to the concept of income used to measure earnings. We view W-2 earnings as the preferred measure because it is unaffected by marriage and the endogeneity of filing.

Based on these results, we construct our preferred measure of college quality measuring college attendance at age 20 and mean W-2 earnings at age 31. These choices allow us to combine data from two cohorts — students born in 1979 and 1980 — for whom we measure earnings in 2010 and 2011, respectively. We code college quality as missing for a small number of institutions with fewer than 100 students across the two cohorts. College quality is also coded as missing for institutions founded in 2001 or later. If students attended two or more colleges in a given year, we assign them the maximum college quality across all colleges attended.

Online Appendix B: Equivalence of Individual and Class-Level Regressions

This appendix shows that the inclusion of individual controls does not affect our point estimate of κ in the absence of missing data. To simplify notation, assume that X_{it}^I is a scalar and that each class has the same number of students. Let Y_c^* denote the mean of Y_i^* in classroom c. For simplicity, we use notation that corresponds to the asymptotic values of various moments of the data, but the result holds in finite samples as well.

Recall that to estimate κ in individual-level data, we first residualize Y_i^* with respect to the control vector using within-teacher variation by estimating the following regression:

$$Y_i^* = a_j + \beta_C X_{ct} + \beta_I (X_{it}^I - X_{ct}) + \varepsilon_{it}^Y \tag{24}$$

We then estimate the univariate regression in (14) using OLS, which yields a coefficient

$$\hat{\kappa} = \frac{Cov\left(Y_i^* - \beta_C X_{ct} + \beta_I (X_{it}^I - X_{ct}), \widehat{m}_{jt}\right)}{Var(\widehat{m}_{jt})}$$

Observe that $Cov\left(X_{it}^I - X_{ct}, \widehat{m}_{jt}\right) = 0$ and $Cov\left(Y_i^* - Y_c^*, \widehat{m}_{jt}\right) = 0$, as the individual deviations have mean 0 in all classrooms and $\widehat{m}_{jt}$ does not vary within classrooms. It follows that the coefficient on the class mean in (24) is

$$\beta_C = \frac{Cov(Y_i^* - \beta_I (X_{it}^I - X_{ct}), X_{ct})}{Var(X_{ct})} = \frac{Cov(Y_i^*, X_{ct})}{Var(X_{ct})} = \frac{Cov(Y_c^*, X_{ct})}{Var(X_{ct})}$$

Similarly, the coefficient on $\widehat{m}_{jt}$ can be written in terms of class means as

$$\hat{\kappa} = \frac{Cov\left(Y_i^* - \beta_C X_{ct}, \widehat{m}_{jt}\right)}{Var\left(\widehat{m}_{jt}\right)} = \frac{Cov\left(Y_c^* - \beta_C X_{ct}, \widehat{m}_{jt}\right)}{Var\left(\widehat{m}_{jt}\right)} = \hat{\kappa}_C,$$

where $\hat{\kappa}_C$ is the coefficient obtained from regressing the residuals of the class means on $\widehat{m}_{jt}$.

It follows that β_C can be estimated from a regression of mean outcomes on mean covariates:

$$Y_c^* = a + \beta_C X_{ct} + \varepsilon_c^Y$$

and $\hat{\kappa}$ can be estimated by regressing the resulting residuals $Y_{ct} = Y_c^* - \beta_C X_{ct}$ on $\widehat{m}_{jt}$:

$$Y_{ct} = \alpha + \kappa \widehat{m}_{jt} + \eta'_{ct}. \tag{25}$$

Online Appendix C: Identifying Teachers' Net Impacts

This appendix shows that the iterative method described in Section 6.3 recovers the net impacts of teacher VA, $\tilde{\kappa}_g$, defined as the impact of raising teacher VA in grade g on earnings, holding fixed VA in subsequent grades. To simplify notation, we omit controls in this derivation; in practice, we residualize all the dependent variables in the regressions below with respect to the standard control vector. Furthermore, we simplify the notation by replacing the year subscript t with a grade subscript g, so that $\widehat{m}_{jg} = \widehat{m}_{j,t_i(g)}$.

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

$$Y_{ig} = \kappa_g \widehat{m}_{jg} + \varepsilon_{ig}^m \tag{26}$$

$$\widehat{m}_{jg'} = \rho_{gg'} \widehat{m}_{jg} + \eta_{igg'}^{\rho} \quad \forall g' > g \tag{27}$$

The first set of equations identifies the reduced-form impact of teacher VA in grade g on earnings. The second set of equations identifies the impact of teacher VA in grade g on teacher VA in future grade g'. Note that identification of the tracking coefficients $\rho_{gg'}$ using (21) 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 conditional on controls:

$$Cov\left(\widehat{m}_{jg}, \eta_{igg'}^{\rho}\right) = 0.$$

After estimating $\{\kappa_g\}$ and $\left\{\rho_{gg'}\right\}$, we recover the net impacts $\tilde{\kappa}_g$ as follows. Under our definition of $\tilde{\kappa}_g$, earnings Y_{ig} can be written as $\sum_{g'=4}^{8} \tilde{\kappa}_{g'} \widehat{m}_{jg'} + \varepsilon_{ig}^m$. Substituting this definition of Y_{ig} into (26) and noting that $\rho_{gg'} = Cov\left(\widehat{m}_{jg'}, \widehat{m}_{jg}\right) / Var\left(\widehat{m}_{jg}\right)$ yields

$$\kappa_g = \frac{Cov\left(\sum_{g'=4}^{8} \tilde{\kappa}_{g'} \widehat{m}_{jg'} + \varepsilon_{ig'}^m, \widehat{m}_{jg}\right)}{Var\left(\widehat{m}_{jg}\right)} = \sum_{g'=4}^{8} \rho_{gg'} \tilde{\kappa}_{g'}.$$

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

$$Cov\left(\widehat{m}_{jg'}, \widehat{m}_{jg}\right) = 0 \quad \text{for } g' < g$$

since past teacher quality $\hat{\mu}_{j(i,g')}$ is one component of the error term ε^{μ}_{igt} in (26). Combined with the fact that $\rho_{gg} = 1$ by definition, these equations imply that

$$\begin{aligned} \kappa_g &= \tilde{\kappa}_g + \sum_{g'=g+1}^{8} \rho_{gg'} \tilde{\kappa}_{g'} \quad \forall g < 8 \\ \kappa_8 &= \tilde{\kappa}_8. \end{aligned}$$

Rearranging this triangular set of equations yields the following system of equations, which can be solved by iterating backwards as in Section 6.3:

(28)
$$\begin{aligned} \tilde{\kappa}_8 &= \kappa_8 \\ \tilde{\kappa}_g &= \kappa_g - \sum_{g'=g+1}^{8} \rho_{gg'} \tilde{\kappa}_{g'} \quad \forall g < 8. \end{aligned}$$

TABLE 1
Summary Statistics for Linked Analysis Dataset

Variable	Mean (1)	Std. Dev. (2)	Observations (3)
Student Data:			
Class size (not student-weighted)	28.2	5.8	240,459
Number of subject-school years per student	6.25	3.18	1,083,556
Test score (SD)	0.12	0.91	6,035,726
Female	50.4%		6,762,896
Age (years)	11.7	1.6	6,762,679
Free lunch eligible (1999-2009)	77.1%		3,309,198
Minority (Black or Hispanic)	72.1%		6,756,138
English language learner	4.9%		6,734,837
Special education	3.1%		6,586,925
Repeating grade	2.7%		6,432,281
Matched to tax data	89.2%		6,770,045
Matched to parents (cond. on match to tax data)	94.8%		6,036,422
Adult Outcomes:			
Annual wage earnings at age 20	5,670	7,733	5,939,022
Annual wage earnings at age 25	17,194	19,889	2,321,337
Annual wage earnings at age 28	20,885	24,297	1,312,800
Total income at age 28	21,780	24,281	1,312,800
In college at age 20	35.6%		5,939,022
In college at age 25	16.5%		2,321,337
More than 4 years of college, ages 18-22	22.7%		4,514,758
College quality at age 20	26,408	13,461	5,934,570
Contributed to a 401(k) at age 28	19.1%		1,312,800
Pct. college graduates in ZIP at age 28	13.7%		929,079
Had a child while a teenager (for women)	14.3%		3,032,170
Owned a house at age 25	4.3%		2,321,337
Married at age 25	11.3%		2,321,337
Parent Characteristics:			
Annual Household income	40,808	34,515	5,720,657
Owned a house	34.8%		5,720,657
Contributed to a 401k	31.3%		5,720,657
Married	42.2%		5,720,657
Age at child birth	28.3	7.8	5,615,400

Notes: All statistics reported are for the linked analysis dataset described in Section 3, which includes students from classrooms in which at least one student would graduate high school in or before 2009 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 1996-2011 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. Test score refers to standardized scale score in math or English. Free lunch is an indicator for receiving free or reduced-price lunches. We link students to their parents by finding the earliest 1040 form from 1996-2011 on which the student is claimed as a dependent. We are unable to link 10.8% of observations to the tax data; the summary statistics for adult outcomes and parent characteristics exclude these observations. Wage earnings are measured from W-2 forms; we assign 0's to students with no W-2's. Total income includes both W-2 wage earnings and self-employment income reported on the 1040. College attendance is measured from 1098-T forms. College quality is the average W-2 earnings at age 31 for students who attended a given college at age 20 (see Section 3.1 for more details). 401(k) contributions are reported on W-2 forms. ZIP code of residence is determined from the address on a 1040 (or W-2 for non-filers); percent college graduates in ZIP is based on the 2000 Census. We measure teen births for female students as an indicator for claiming a dependent who was born fewer than 20 years after the student herself was born. We measure home ownership from the payment of mortgage interest, reported on either the 1040 or a 1099 form. We measure marriage by the filing of a joint return. Conditional on linking to the tax data, we are unable to link 5.2% of observations to a parent; the summary statistics for parents exclude these observations. Parent income is average adjusted gross income during the three tax-years between 2005-2007. For parents who do not file, household income is defined as zero. Parent age at child birth is the difference between the age of the mother (or father if single father) and the student. All parent indicator variables are defined in the same way as the equivalent for the students and are equal to 1 if the event occurs in any year between 2005-2007.

TABLE 2
Impacts of Teacher Value-Added on College Attendance

Dep. Var.:	College at Age 20	College at Age 20	College at Age 20	College Quality at Age 20	College Quality at Age 20	College Quality at Age 20	High Quality College	4 or More Years of College, Ages 18-22
	(%)	(%)	(%)	($)	($)	($)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher VA	0.82	0.71	0.74	298.63	265.82	266.17	0.72	0.79
	(0.07)	(0.06)	(0.09)	(20.74)	(18.31)	(26.03)	(0.05)	(0.08)
Mean of Dep. Var.	37.22	37.22	37.09	26,837	26,837	26,798	13.41	24.59
Baseline Controls	X	X	X	X	X	X	X	X
Parent Chars. Controls		X			X			
Lagged Score Controls			X			X		
Observations	4,170,905	4,170,905	3,130,855	4,167,571	4,167,571	3,128,478	4,167,571	3,030,878

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 (as described in the notes to Table 1). Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure described in Section 4.1. The dependent variable in Columns 1-3 is an indicator for college attendance at age 20. The dependent variable in Columns 4-6 is the earnings-based index of college quality. See notes to Table 1 and Section 3 for more details on the construction of these variables. The dependent variable in Column 7 is an indicator for attending a high-quality college, defined as quality greater than the median college quality among those attending college, which is $43,914. The dependent variable in Column 8 is an indicator for attending four or more years of college between the ages of 18 and 22. All columns control for the baseline class-level control vector, which includes: class size and class-type indicators; cubics in class and school-grade means of lagged own- and cross-subject scores, interacted with grade level; class and school-year means of student-level characteristics including ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, free or reduced-price lunch, and limited English; and grade and year dummies. Columns 2 and 5 additionally control for class means of parent characteristics, including 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. Columns 3 and 6 include the baseline controls and class means of twice-lagged test scores. We use within-teacher variation to identify the coefficients on all controls as described in Section 2.1; the estimates reported are from regressions of outcome residuals on teacher VA with school by subject level fixed effects.

TABLE 3
Impacts of Teacher Value-Added on Earnings

Dep. Var.:	Earnings at Age 28	Earnings at Age 28	Earnings at Age 28	Working at Age 28	Total Income at Age 28	Wage Growth Ages 22-28
	($)	($)	($)	(%)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)
Teacher VA	349.84	285.55	308.98	0.38	353.83	286.20
	(91.92)	(87.64)	(110.17)	(0.16)	(88.62)	(81.86)
Mean of Dep. Var.	21,256	21,256	21,468	68.09	22,108	11,454
Baseline Controls	X	X	X	X	X	X
Parent Chars. Controls		X				
Lagged Score Controls			X			
Observations	650,965	650,965	510,309	650,965	650,965	650,943

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 (as described in the notes to Table 1). 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 described in Section 4.1. The dependent variable in Columns 1-3 is the individual's wage earnings reported on W-2 forms at age 28. The dependent variable in Column 4 is an indicator for having positive wage earnings at age 28. The dependent variable in Column 5 is total income (wage earnings plus self-employment income). The dependent variable in Column 6 is wage growth between ages 22 and 28. All columns control for the baseline class-level control vector; Column 2 additionally controls for parent characteristics, while Column 3 additionally controls for twice-lagged test scores (see notes to Table 2 for details). We use within-teacher variation to identify the coefficients on all controls as described in Section 2.1; the estimates reported are from regressions of outcome residuals on teacher VA with school by subject level fixed effects.

TABLE 4
Impacts of Teacher Value-Added on Other Outcomes

Dep. Var.:	Teenage Birth	Percent College Grad in ZIP at Age 28	Have 401(k) at Age 28
	(%)	(%)	(%)
	(1)	(2)	(3)
Teacher VA	-0.61	0.25	0.55
	(0.06)	(0.04)	(0.16)
Mean of Dep. Var.	13.24	13.81	19.81
Baseline Controls	X	X	X
Observations	2,110,402	468,021	650,965

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 (as described in the notes to Table 1). 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 described in Section 4.1. The dependent variables in Column 1-3 are an indicator for having a teenage birth, the fraction of residents in an individual's zip code of residence at age 28 with a college degree or higher, and an indicator for whether an individual made a contribution to a 401(k) plan at age 28 (see notes to Table 1 and Section 3 for more details). Column 1 includes only female students. All regressions include the baseline class-level control vector (see notes to Table 2 for details). We use within-teacher variation to identify the coefficients on all controls as described in Section 2.1; the estimates reported are from regressions of outcome residuals on teacher VA with school by subject level fixed effects.

TABLE 5
Impacts of Teacher Value-Added on College Outcomes: Quasi-Experimental Estimates

Panel A: College Attendance at Age 20

Dep. Var.:	College Attendance (%)				Pred. Coll. Attendance (%)
	(1)	(2)	(3)	(4)	(5)
Teacher VA	0.86	0.73	0.67	1.20	0.02
	(0.23)	(0.25)	(0.26)	(0.58)	(0.06)
Year FE	X				
School x Year FE		X	X	X	X
Lagged Score Controls			X		
Lead and Lag Changes in Teacher VA			X		
Number of School x Grade x Subject x Year Cells	33,167	33,167	26,857	8,711	33,167
Sample:	Full Sample	Full Sample	Full Sample	No Imputed Scores	Full Sample

Panel B: College Quality at Age 20

Dep. Var.:	College Quality ($)				Pred. Coll. Quality ($)
	(1)	(2)	(3)	(4)	(5)
Teacher VA	197.64	156.64	176.51	334.52	2.53
	(60.27)	(63.93)	(64.94)	(166.85)	(18.30)
Year FE	X				
School x Year FE		X	X	X	X
Lagged Score Controls			X		
Lead and Lag Changes in Teacher VA			X		
Number of School x Grade x Subject x Year Cells	33,167	33,167	26,857	8,711	33,167
Sample:	Full Sample	Full Sample	Full Sample	No Imputed Scores	Full Sample

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 (as described in the notes to Table 1), collapsed to school-grade-year-subject means. The independent variable for each regression is the difference in mean teacher value-added between adjacent school-grade-year-subject cells, where we estimate teacher value-added using data that omits both years (see Section 5.1 for more details). Similarly, dependent variables are defined as changes in means across consecutive cohorts at the school-grade-year-subject level. In Panel A, the dependent variable is college attendance at age 20; in Panel B, the dependent variable is the earnings-based index of college quality (see Table 1 for details). In Column 1 we regress the mean change in the dependent variable on the mean change in teacher value-added, controlling only for year fixed-effects. Column 2 replicates Column 1 including school-year fixed effects. In Column 3, we add a cubic in the change in mean lagged scores to the specification in Column 2, as well as controls for the lead and lag change in mean teacher value-added. In Column 4, we restrict the sample to cells with no imputed VA; other columns impute the sample mean of 0 for classes with missing VA. Column 5 replicates Column 2, except that the dependent variable is the predicted value from an individual-level regression of the original dependent variable on the vector of parent characteristics defined in the notes to Table 2.

TABLE 6
Heterogeneity in Impacts of Teacher Value-Added

Panel A: Impacts by Demographic Group

	Female	Male	Low Income	High Income	Minority	Non-Minority
	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	College Quality at Age 20 ($)					
Teacher VA	290.65	237.93	190.24	379.89	215.51	441.08
	(23.61)	(21.94)	(19.63)	(27.03)	(17.09)	(42.26)
Mean of Dep. Var.	27,584	26,073	23,790	30,330	23,831	33,968
Impact as % of Mean	1.05%	0.91%	0.80%	1.25%	0.90%	1.30%
Dep. Var.:	Test Score (SD)					
Teacher VA	0.135	0.136	0.128	0.129	0.136	0.138
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Mean of Dep. Var.	0.196	0.158	-0.003	0.331	-0.039	0.651

Panel B: Impacts by Subject

	Elementary School			Middle School	
	(1)	(2)	(3)	(4)	(5)
Dep. Var.:	College Quality at Age 20 ($)				
Math Teacher VA	207.81		106.34	265.59	
	(21.77)		(28.50)	(43.03)	
English Teacher VA		258.16	189.24		521.61
		(25.42)	(33.07)		(63.67)
Control for Average VA in Other Subject				X	X

Notes: In the first row of estimates in Panel A, we replicate the specification in Column 5 of Table 2 within various population subgroups. In Columns 1 and 2, we split the sample between males and females; in Columns 3 and 4, we split the sample based on the median parent household income (which is $31,905); in Columns 5 and 6, we split the sample based on whether a student belongs to an ethnic minority (black or hispanic). In the second row of estimates in Panel A, we replicate all of the regressions from the first row replacing college quality with score as the dependent variable. In Panel B, we split the sample into elementary 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). Columns 1 and 2 replicate the specification in Column 5 of Table 2, splitting the sample by subject. In Column 3, we regress college quality on measures of math teacher value-added and English teacher value-added together in a dataset reshaped to have one row per student by school year. We restrict the sample so that the number of teacher-year observations is identical in Columns 1-3. Columns 4 and 5 replicate Column 5 of Table 2 for middle schools with an additional control for the average teacher value-added in the other subject for students in a given class.

APPENDIX TABLE 1
Structure of Linked Analysis Dataset

Student	Subject	Year	Grade	Class	Teacher	Test Score	Matched to Tax Data?	Earnings at Age 28
Bob	Math	1992	4	1	Jones	0.5	1	$35K
Bob	English	1992	4	1	Jones	-0.3	1	$35K
Bob	Math	1993	5	2	Smith	0.9	1	$35K
Bob	English	1993	5	2	Smith	0.1	1	$35K
Bob	Math	1994	6	3	Harris	1.5	1	$35K
Bob	English	1994	6	4	Adams	0.5	1	$35K
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 linked analysis sample which combines information from the school district database and the tax data. There is one row for each student-subject-school year. 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 are for illustrative purposes only.

APPENDIX TABLE 2

Correlation of College Rankings Based on Alternative Measures

Panel A: Correlation of College Rankings Across Ages at Which Earnings are Measured

	Age 23	Age 24	Age 25	Age 26	Age 27	Age 28	Age 29	Age 30	Age 31	Age 32
Age 23	1.000									
Age 24	0.858	1.000								
Age 25	0.747	0.949	1.000							
Age 26	0.676	0.901	0.967	1.000						
Age 27	0.614	0.852	0.928	0.972	1.000					
Age 28	0.577	0.822	0.900	0.950	0.979	1.000				
Age 29	0.553	0.802	0.882	0.934	0.962	0.983	1.000			
Age 30	0.519	0.774	0.860	0.916	0.948	0.968	0.980	1.000		
Age 31	0.505	0.761	0.848	0.905	0.937	0.958	0.971	0.986	1.000	
Age 32	0.495	0.750	0.838	0.897	0.930	0.952	0.964	0.977	0.987	1.000

Panel B: Correlation of College Rankings Across Ages at Which College Attendance is Measured

	Age 18	Age 19	Age 20	Age 21	Age 22	Age 23	Age 24	Age 25
Age 18	1.000							
Age 19	0.948	1.000						
Age 20	0.930	0.975	1.000					
Age 21	0.909	0.947	0.972	1.000				
Age 22	0.880	0.914	0.940	0.968	1.000			
Age 23	0.850	0.886	0.909	0.933	0.960	1.000		
Age 24	0.803	0.830	0.851	0.873	0.893	0.932	1.000	
Age 25	0.766	0.790	0.806	0.830	0.851	0.883	0.935	1.000

Panel C: Correlation of College Rankings Across Birth Cohorts

	Cohort 1979	Cohort 1980	Cohort 1981
Cohort 1979	1.000		
Cohort 1980	0.931	1.000	
Cohort 1981	0.933	0.942	1.000

Panel D: Correlation of College Rankings Across Earnings Definitions

	Mean W-2	Median W-2	Mean W-2 + S-E
Mean W-2 Earnings	1.000		
Median W-2 Earnings	0.960	1.000	
Mean W-2 + Self-Employment Income	0.989	0.943	1.000

Notes: This table displays Spearman rank correlations between alternative earnings-based indices of college quality, each of which is defined by four characteristics: age of earnings measurement, age of college attendance, cohort of students, and definition of earnings. Throughout this table, we construct college quality measures from only a single birth cohort of students; however, the preferred measure used in the text combines two cohorts. Panel A varies the age of earnings measurement from 23 to 32, holding fixed the age of college attendance at 20, using only the 1979 cohort of students, and using mean W-2 wage earnings. Panel B varies the age of college attendance from 18 to 25, holding fixed the age of earnings measurement at 30, using only the 1981 cohort, and using mean W-2 wage earnings. Panel C varies the birth cohort of students, holding fixed the age of college attendance at 20, the age of earnings measurement at 30, and using mean W-2 wage earnings. Panel D varies the measure of earnings between the baseline (mean W-2 wage earnings by college) and two alternatives (median W-2 wage earnings by college and mean total income by college), holding fixed the age of college attendance at 20, the age of earnings measurement at 30, and using only the 1979 cohort of students.

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

Dep. Var.:	College at Age 20	College Quality at Age 20	Earnings at Age 28	Teenage Birth	Percent College Grads in ZIP at Age 28
	(%)	($)	($)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)
No Controls	18.37	6,366	7,709	-6.57	1.87
	(0.02)	(6)	(23)	(0.02)	(0.01)
With Controls	5.54	2,114	2,585	-1.58	0.34
	(0.04)	(11)	(59)	(0.05)	(0.01)
Math	6.04	2,295	2,998	-1.21	0.31
Full Controls	(0.06)	(16)	(83)	(0.07)	(0.02)
English	5.01	1,907	2,192	-2.01	0.37
Full Controls	(0.06)	(16)	(88)	(0.06)	(0.02)
Mean of Dep. Var.	37.71	26,963	21,622	13.25	13.43

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 observations from the linked analysis sample (as described in the notes to Table 1). 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 an indicator for attending college at age 20 in column 1, our earnings-based index of college quality in column 2, wage earnings at age 28 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 28 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 Section 4.1, as well as teacher fixed effects. 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. The final row displays of the mean of the dependent variable in the sample for which we have the full control vector (i.e., the sample used in the 2nd row).

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

	Dependent Variable: Earnings ($)								
Age:	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
No Controls	889	1,098	1,864	3,592	4,705	5,624	6,522	7,162	7,768
	(20)	(25)	(28)	(34)	(39)	(44)	(48)	(51)	(54)
With Controls	392	503	726	1,372	1,759	1,971	2,183	2,497	2,784
	(64)	(79)	(91)	(110)	(125)	(139)	(152)	(161)	(171)
Mean Earnings	6,484	8,046	9,559	11,777	14,004	16,141	18,229	19,834	21,320
Pct. Effect (with controls)	6.1%	6.2%	7.6%	11.6%	12.6%	12.2%	12.0%	12.6%	13.1%

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. See notes to Table 1 for our definition of earnings. We restrict this table to students born in cohorts 1979 and 1980, so that regressions are estimated on a constant subsample of the linked analysis sample. 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 Section 4.1 as well as teacher fixed effects. Means of earnings for the estimation sample with controls 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	2,408	5.36	1,976	n/a
	(88)	(0.06)	(16)	
	[22,179]	[34.24]	[26,205]	
Female	2,735	5.74	2,262	-1.58
	(80)	(0.06)	(17)	(0.05)
	[21,078]	[41.07]	[27,695]	[13.25]
Non-minority	2,492	5.11	2,929	-0.72
	(139)	(0.08)	(27)	(0.04)
	[31,587]	[59.67]	[34,615]	[2.82]
Minority	2,622	5.65	1,734	-1.96
	(62)	(0.05)	(12)	(0.06)
	[17,644]	[28.98]	[23,917]	[17.20]
Low Parent Inc.	2,674	5.14	1,653	-1.72
	(85)	(0.06)	(15)	(0.07)
	[18,521]	[26.91]	[23,824]	[16.67]
High Parent Inc.	2,573	5.73	2,539	-1.29
	(92)	(0.06)	(18)	(0.06)
	[26,402]	[49.92]	[30,420]	[9.21]

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses and the mean of the dependent variable in the estimation sample in brackets. These regressions replicate the second row (full sample, with controls and teacher fixed effects) of estimates in Columns 1-4 of Appendix Table 3, splitting the sample based on student demographic characteristics. The demographic groups are defined in exactly the same way as in Panel A of Table 6. We split rows 1 and 2 by the student's gender. 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 the sample, which is $31,905.

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	7,561	18.29	6,378	2,970	6.78	2,542
	(57)	(0.05)	(13)	(122)	(0.09)	(23)
Grade 5	7,747	18.27	6,408	2,711	5.28	2,049
	(50)	(0.05)	(13)	(108)	(0.08)	(23)
Grade 6	7,524	17.95	6,225	2,395	4.92	1,899
	(51)	(0.05)	(14)	(140)	(0.10)	(27)
Grade 7	7,891	18.23	6,197	2,429	4.48	1,689
	(54)	(0.05)	(14)	(198)	(0.11)	(29)
Grade 8	7,795	19.10	6,596	2,113	5.43	2,106
	(48)	(0.05)	(13)	(141)	(0.11)	(28)

Notes: Each column reports coefficients from an OLS regression, with standard errors in parentheses. The regressions in the first three columns replicate the first row (full sample, no controls) of estimates in Columns 1-3 of Appendix Table 3, splitting the sample by grade. The regressions in the second set of three columns replicate the second row (full sample, with controls and teacher fixed effects) of estimates in Columns 1-3 of Appendix Table 3, again splitting the sample by grade.

APPENDIX TABLE 7

Robustness of Baseline Results to Student-Level Controls, Clustering, and Missing Data

Dep. Var.:	College at Age 20 (%) (1)	College Quality at Age 20 ($) (2)	Earnings at Age 28 ($) (3)
Panel A: Individual Controls			
Teacher VA, with baseline controls	0.825	299	350
	(0.072)	(21)	(92)
Observations	4,170,905	4,167,571	650,965
Teacher VA, with additional individual controls	0.873	312	357
	(0.072)	(21)	(90)
Observations	4,170,905	4,167,571	650,965
Panel B: Clustering			
Teacher VA, school clustered	0.825	299	350
	(0.115)	(36)	(118)
Observations	4,170,905	4,167,571	650,965
Panel C: Missing Data			
Teacher VA, cells > 95% VA coverage	0.819	277	455
	(0.090)	(26)	(202)
Observations	2,238,143	2,236,354	363,392
Teacher VA, cells > median match rate	0.912	345	563
	(0.094)	(28)	(203)
Observations	2,764,738	2,762,388	278,119

Notes: The table presents robustness checks of the main results in Tables 2 and 3. In Panel A, the first row replicates Columns 1 and 4 of Table 2 and Column 1 of Table 3 as a reference. The second row adds individual controls, so that the control vector exactly matches that used to estimate the value-added model (see Section 4.1 for details). Panel B clusters standard errors by school. In Panel C, the first row limits the sample to those school-grade-year-subject cells in which we are able to calculate teacher value-added for at least 95% of students. The second row limits the sample to those school-grade-year-subject cells in which the rate at which we are able to match observations to the tax data is more than the school-level-subject-specific median across cells.

APPENDIX TABLE 8

Impacts of Teacher Value-Added: Sensitivity to Trimming

	Percent Trimmed in Upper Tail						Bottom and Top 1%	Jacob and Levitt Proxy
	5%	4%	3%	2%	**1%**	0%		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test Score	0.972	0.975	0.981	0.987	0.993	1.005	0.996	0.918
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	0.006	(0.006)
College at Age 20	0.93	0.90	0.88	0.86	0.82	0.72	0.79	1.10
	(0.08)	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.09)
College Quality at Age 20	329	320	315	307	299	276	292	371
	(23)	(22)	(22)	(21)	(21)	(21)	(21)	(25)
Earnings at Age 28	404	405	390	356	350	248	337	391
	(102)	(100)	(99)	(96)	(92)	(91)	(94)	(118)

Notes: This table presents results that use alternative approaches to trimming the tails of the distribution of teacher VA. Each coefficient reports the coefficient on teacher VA from a separate OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions in the first row replicate the baseline specification used in Column 1 of Table 3 in our companion paper (using VA scaled in units of student test-score SDs), except that we include only the class-level controls that correspond to the baseline set of controls in this paper (as in Section 4.1). The regressions in rows 2-4 replicate the baseline specification used in Columns 1 and 4 of Table 2 and Column 1 of Table 3. Columns 1-6 report results for trimming the upper tail at various cutoffs. Column 7 shows estimates when both the bottom and top 1% 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 3. The results in Column 5 (1% trimming) correspond to those reported in the main text.

APPENDIX TABLE 9
Impacts of Teacher Value-Added on Outcomes by Age

Panel A: College Attendance

Dependent Variable:	College Attendance (%)										
Age:	18	19	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Teacher Value-Added	0.61	0.81	0.82	0.98	0.71	0.44	0.58	0.46	0.50	0.46	-0.01
	(0.06)	(0.07)	(0.07)	(0.08)	(0.07)	(0.07)	(0.07)	(0.08)	(0.07)	(0.09)	(0.11)
Mean Attendance Rate	29.4	36.8	37.2	35.7	32.2	24.4	20.31	17.3	15.7	13.9	12.3

Panel B: Wage Earnings

Dependent Variable:	Earnings ($)								
Age:	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Teacher Value-Added	-32	-35	-18	44	74	141	230	254	350
	(11)	(14)	(18)	(25)	(32)	(44)	(47)	(63)	(92)
Mean Earnings	5,696	7,293	9,473	12,582	15,080	17,547	18,833	20,229	21,256

Notes: These results present the regression estimates underlying the results in Panel C of Figure 1 (in Panel A) and Panel B of Figure 2 (in Panel B). The regressions in Panel A match the specification from Column 1 of Table 2, with college attendance measured at different ages; those in Panel B match the specification from Column 1 of Table 3.

APPENDIX TABLE 10

Impacts of Teacher Value-Added on Current and Future Test Scores

Dep. Var.:	Test Score (SD)				
	t	t+1	t+2	t+3	t+4
	(1)	(2)	(3)	(4)	(5)
Teacher VA	0.993	0.533	0.362	0.255	0.221
	(0.006)	(0.007)	(0.007)	(0.008)	(0.012)
Observations	7,401,362	5,603,761	4,097,344	2,753,449	1,341,266

Notes: This table presents the regression estimates plotted in Figure 4; see notes to that figure for details.

APPENDIX TABLE 11
Impacts of Value-Added on College Quality by Grade

College Quality at Age 20					
Panel A: Reduced-Form Coefficients					
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Teacher Value-Added	226	289	292	482	198
	(31)	(33)	(48)	(61)	(48)
Panel B: Coefficients Net of Teacher Tracking					
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Teacher Value-Added	204	275	216	433	198

Notes: This table presents the regression estimates plotted in Figure 7; see notes to that figure for details.

APPENDIX TABLE 12
Tracking: Impact of Current Teacher VA on Future Teachers' VA

	Future Teacher Value-Added			
	Grade 5	Grade 6	Grade 7	Grade 8
Grade 4 Teacher VA	0.017	0.035	0.015	0.017
	(0.004)	(0.003)	(0.003)	(0.003)
Grade 5 Teacher VA		0.036	0.009	0.012
		(0.004)	(0.003)	(0.008)
Grade 6 Teacher VA			0.121	0.120
			(0.007)	(0.009)
Grade 7 Teacher VA				0.248
				(0.010)

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, with standard errors at the school-cohort level. As in Figure 7, we first residualize each dependent variable (i.e. lead VA, two-year lead VA, etc.) with respect to the classroom-level baseline control vector (see notes to Table 2 for more details). We then regress residualized future VA on current VA interacted with grade. All regressions are estimated using observations in the linked analysis sample for which the student is progressing through grades at normal pace (e.g., the student is in sixth grade two years after fourth grade).

APPENDIX TABLE 13

Earnings Impacts of Replacing Teachers Below 5th Percentile with Average Teachers

Panel A: Impacts in First Year After Deselection

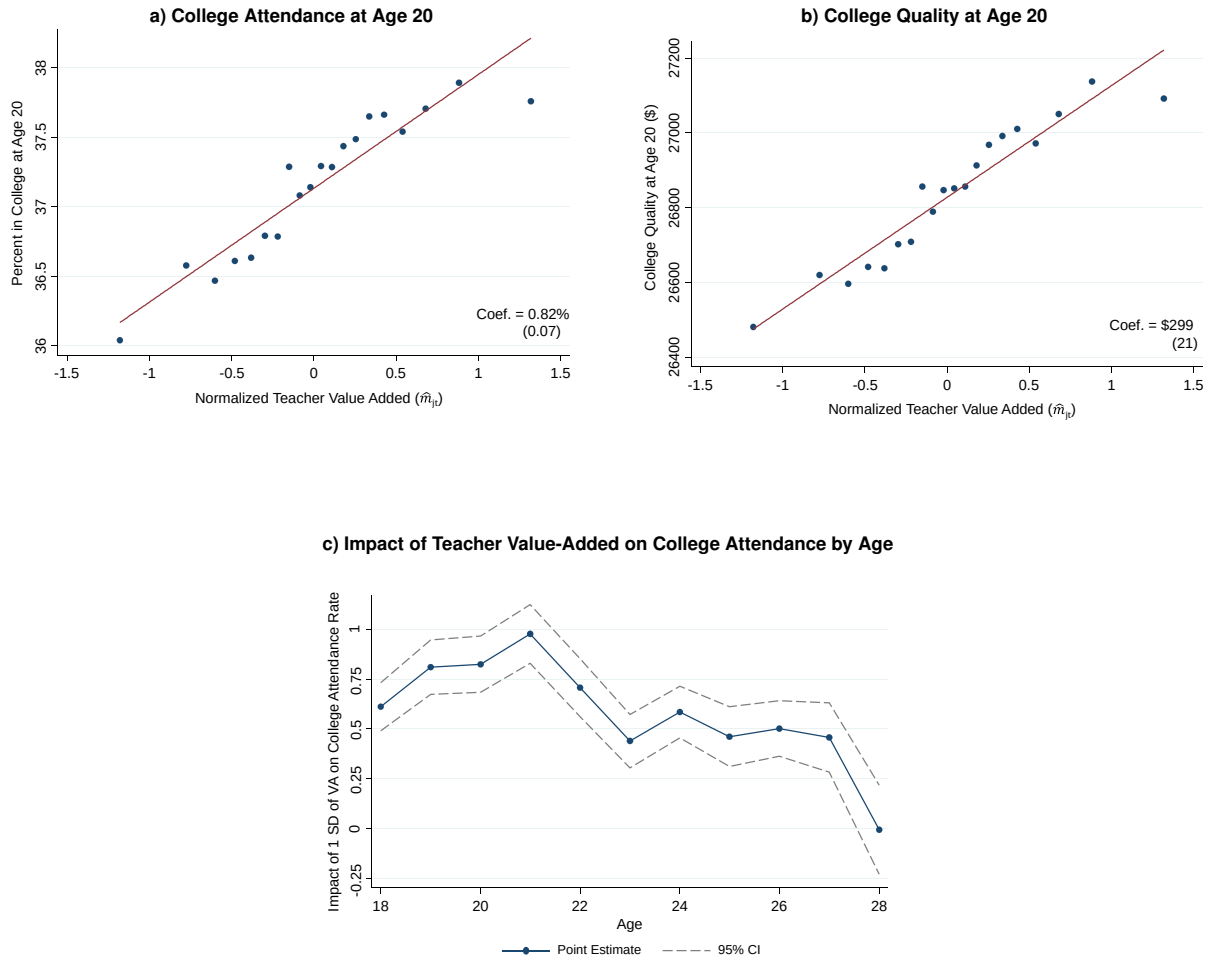
	Selection on Estimated VA		Selection on True VA	
Years Used to Estimate VA	Present Value of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class	Present Value of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class
1	$225,843	$1,249,636	$406,988	$2,251,954
2	$256,651	$1,420,105		
3	$265,514	$1,469,147		
4	$269,297	$1,490,081		
5	$272,567	$1,508,174		
6	$274,143	$1,516,891		
7	$275,232	$1,522,918		
8	$276,665	$1,530,845		
9	$278,112	$1,538,851		
10	$279,406	$1,546,013		

Panel B: Impacts in Subsequent School Years

	Selection on Estimated VA in Yr. 4		Selection on True VA in Yr. 4	
School Years Since Teacher was Hired	Present Value of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class	Present Value of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class
4	$265,514	$1,469,147	$406,988	$2,251,954
5	$229,923	$1,272,213	$339,870	$1,880,574
6	$202,631	$1,121,202	$297,569	$1,646,511
7	$183,538	$1,015,557	$252,422	$1,396,703
8	$172,867	$956,509	$222,339	$1,230,251
9	$161,575	$894,032	$212,185	$1,174,067
10	$157,812	$873,209	$193,255	$1,069,324
11	$155,349	$859,581	$180,876	$1,000,824
12	$156,582	$866,400	$180,909	$1,001,007
13	$156,547	$866,206	$181,027	$1,001,662
Avg. Gain	$184,234	$1,019,405	$246,744	$1,365,288

Notes: In Panel A, we present the earnings impacts per classroom of a policy that deselects the bottom 5% of teachers after *N* years and replaces them with a teacher of median quality, where we vary *N* from 1 to 10. We calculate these values using the methods described in Section 7. The first column presents estimates of the NPV earnings gains of deselection based on teacher value-added that is estimated from *N* years of observing a single average-sized (28.2 students) classroom per year of student scores. The third column shows the theoretical gain from deselecting teachers based on current true value-added; this value does not vary across years. Panel B presents the per class impacts of deselecting teachers (after 3 years of observation) in subsequent school years. Column 1 reports the present value of earnings gains in the ten years (i.e., years 4-13) after deselecting teachers based on their VA estimate in year 4, constructed using the past three years of data. The first number in Column 1 of Panel B matches the 3rd number in Column 1 of Panel A. Column 3 presents analogous values, instead deselecting teachers based on true value-added in year 4, so that the gains in the year 4 match the gains reported in Column 3 of Panel A. Columns 2 and 4 in each panel replicate Columns 1 and 3, presenting the undiscounted sum of future earnings impacts instead of present values. The bottom row in the table reports the unweighted means of the estimates from years 4-13 in Panel B for each column; these are the values reported in the introduction of the paper.

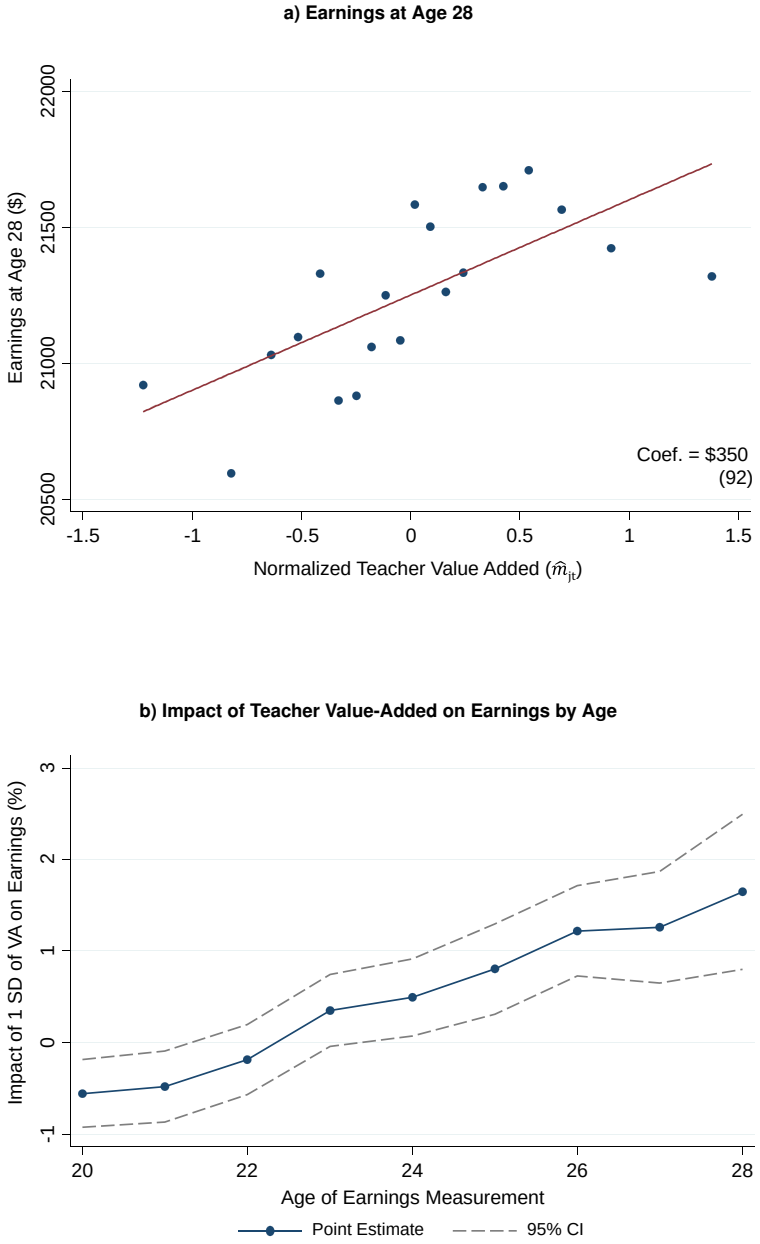
FIGURE 1
Effects of Teacher Value-Added on College Outcomes



Notes: These figures are drawn using the linked analysis sample, pooling all grades and subjects, with one observation per student-subject-school year. Panels A and B are binned scatter plots of college attendance rates and college quality vs. normalized teacher VA $\hat{m}_{jt}$. These plots correspond to the regressions in Columns 1 and 4 of Table 2 and use the same sample restrictions and variable definitions. To construct these binned scatter plots, we first residualize the y-axis variable with respect to the baseline class-level control vector (defined in the notes to Table 2) separately within each subject by school-level cell, using within-teacher variation to estimate the coefficients on the controls as described in Section 2.1. We then divide the VA estimates $\hat{m}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the y-variable residuals within each bin against the mean value of $\hat{m}_{jt}$ within each bin. Finally, we add back the unconditional mean of the y variable in the estimation sample to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data 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. In Panel C, we replicate the regression in Column 1 of Table 2 (depicted in Panel A), varying the age of college attendance from 18 to 28, and plot the resulting coefficients. The dashed lines show the boundaries of the 95% confidence intervals for the effect of value-added on college attendance at each age, with standard errors clustered by school-cohort. The coefficients and standard errors from the regressions underlying Panel C are reported in Appendix Table 9.

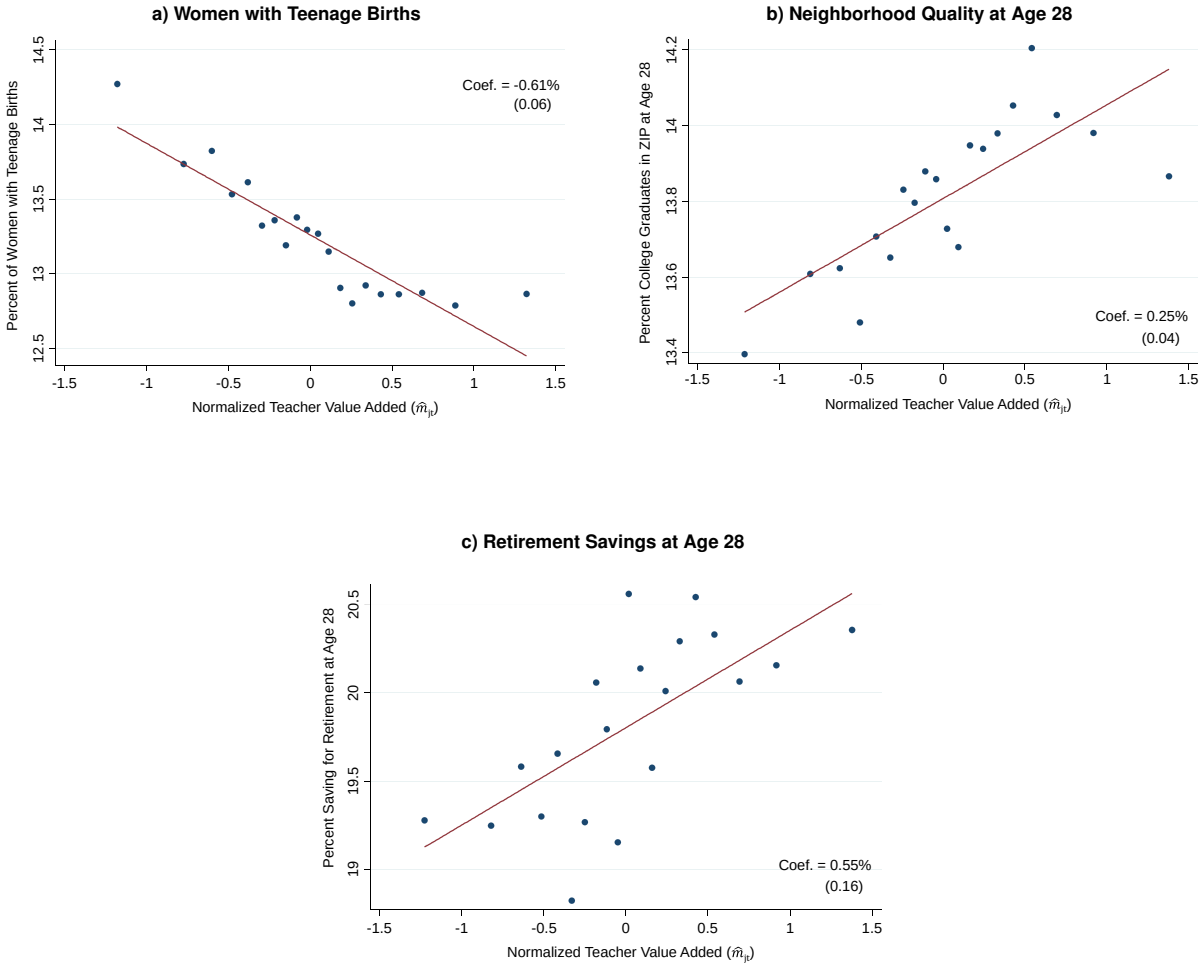
FIGURE 2

Effect of Teacher Value-Added on Earnings



Notes: Panel A is a binned scatter plot of earnings at age 28 vs. normalized teacher VA $\hat{m}_{jt}$. This plot corresponds to the regression in Column 1 of Table 3 and uses the same sample restrictions and variable definitions. See notes to Figure 1 for details on the construction of binned scatter plots. In Panel B, we replicate the regression in Column 1 of Table 3 (depicted in Panel A), varying the age at which earnings are measured from 20 to 28. We then plot the resulting coefficients expressed as a percentage of mean wage earnings in the regression sample at each age. The dashed lines show the boundaries of the 95% confidence intervals for the effect of value-added on earnings at each age, with standard errors clustered by school-cohort. The coefficients and standard errors from the regressions underlying Panel B are reported in Appendix Table 9.

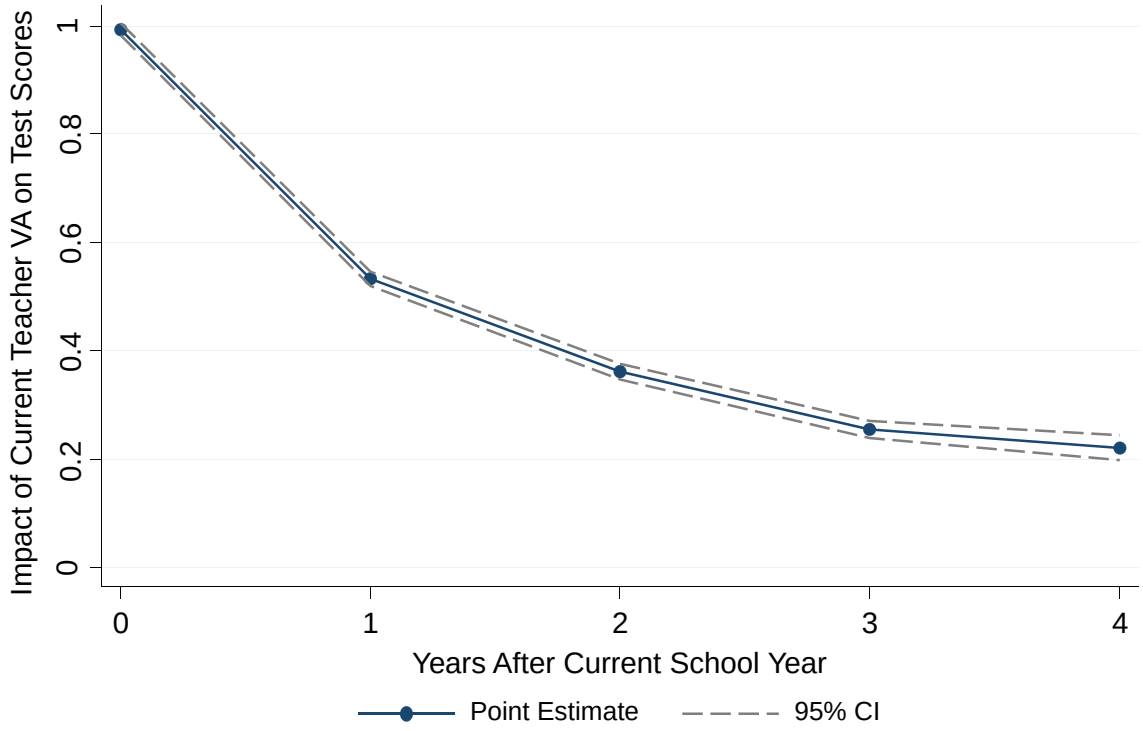
FIGURE 3
Effects of Teacher Value-Added on Other Outcomes in Adulthood



Notes: These three figures are binned scatter plots corresponding to Columns 1-3 of Table 4 and use the same sample restrictions and variable definitions. See notes to Figure 1 for details on the construction of these binned scatter plots.

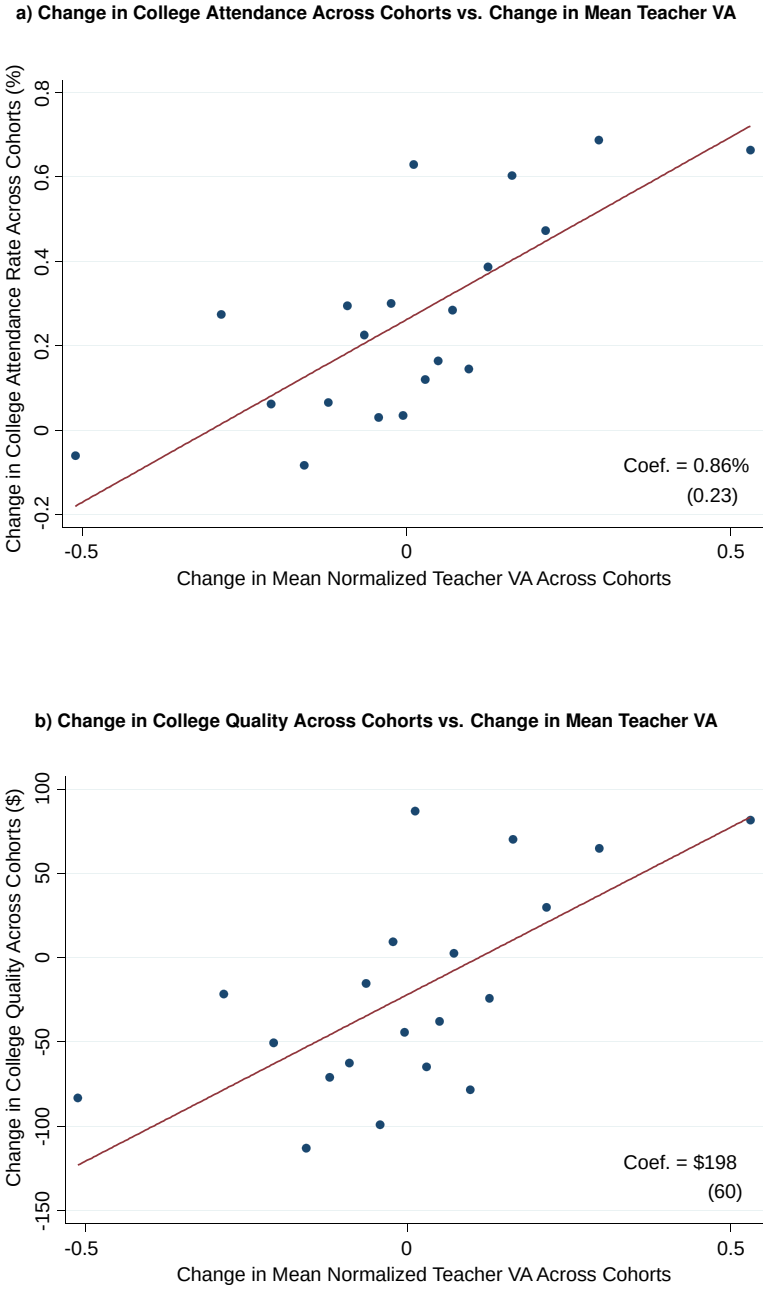
FIGURE 4

Effects of Teacher Value-Added on Future Test Scores



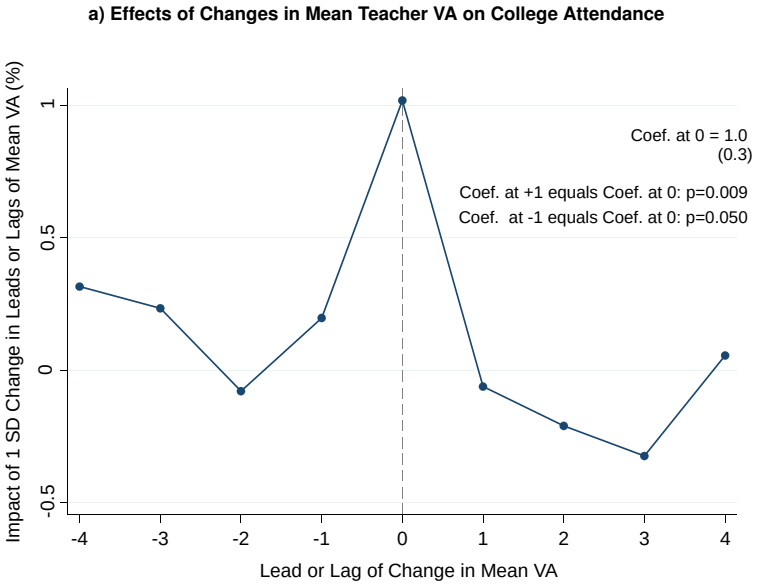
Notes: This figure shows the effect of current teacher VA on test scores at the end of the current and subsequent school years. To construct this figure, we regress end-of-grade test scores in year $t+s$ on teacher VA $\hat{\mu}_{jt}$ in year t, varying s from 0 to 4. As in our companion paper, we scale teacher VA in units of student test-score SD's and include all students in the school district data in these regressions, without restricting to the older cohorts that we use to study outcomes in adulthood. We control for the baseline class-level control vector (defined in the notes to Table 2), using within-teacher variation to identify the coefficients on controls as described in Section 2.1. The dashed lines depict 95% confidence intervals on each regression coefficient, with standard errors clustered by school-cohort. The coefficients and standard errors from the underlying regressions are reported in Appendix Table 10.

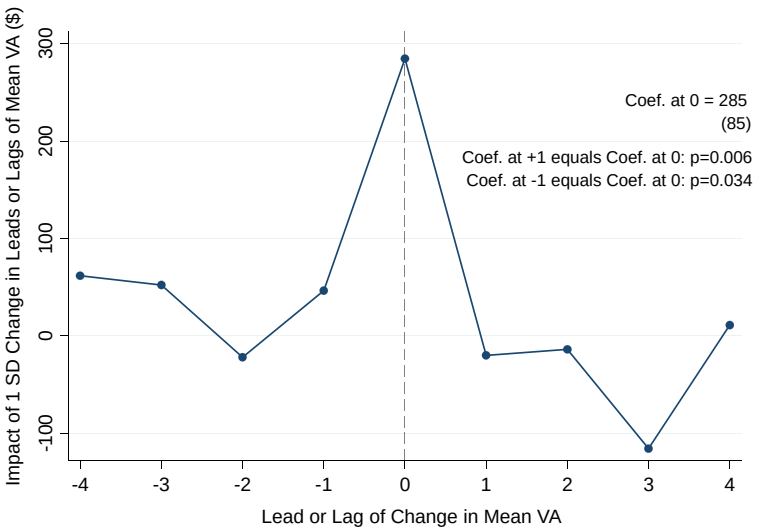
FIGURE 5
Effects of Changes in Teaching Staff Across Cohorts on College Outcomes



Notes: These two figures plot changes in mean college attendance rates (measured in percentage points) and college quality across adjacent cohorts within a school-grade-subject cell against changes in mean teacher VA across those cohorts. These plots correspond to the regressions in Column 1 of Tables 5A and Table 5B and use the same sample restrictions and variable definitions. To construct these binned scatter plots, we first demean both the x- and y-axis variables by school year to eliminate any secular time trends. We then divide the observations into twenty equal-size groups (vingtiles) based on their change in mean VA and plot the means of the y variable within each bin against the mean change in VA within each bin, weighting by the number of students in each school-grade-subject-year cell. Finally, we add back the unconditional (weighted) mean of the x and y variables in the estimation sample.

FIGURE 6
Timing of Changes in Teacher Quality and College Outcomes

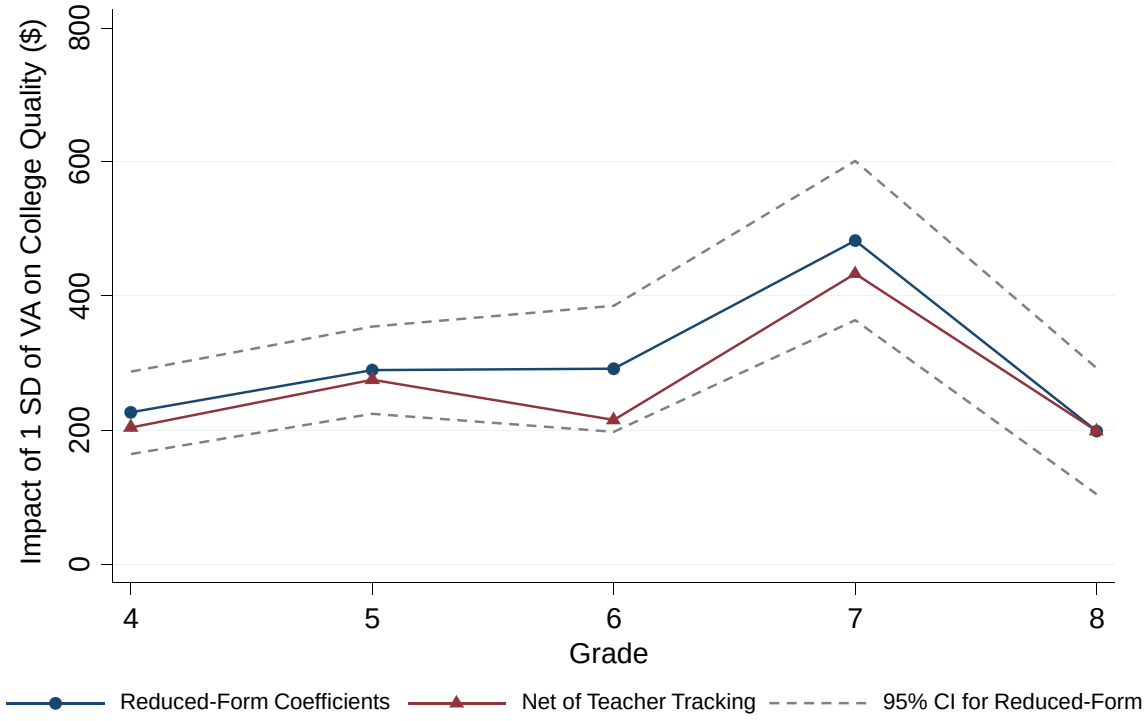




Notes: These figures evaluate whether the timing of changes in teacher quality across cohorts aligns with the timing of changes in college outcomes. The point at 0 represents the "treatment effect" of changes in teacher quality on changes in college outcomes for a given group of students; the other points are "placebo tests" that show the impacts of changes in teacher quality for previous and subsequent cohorts on the same set of students. To construct Panel A, we regress the change in mean college attendance between adjacent cohorts within a school-grade-subject cell on the change in mean teacher quality across those cohorts as well as four lags and four leads of the change in mean teacher quality within the same school-grade-subject. The regression also includes year fixed effects. Panel A plots the coefficients from this regression. We report the point estimate and standard error on the own-year change in mean teacher quality (corresponding to the value at 0). We also report p-values from hypothesis tests for the equality of the own-year coefficient and the one-year lead or one-year lag coefficients. These standard errors and p-values account for clustering at the school-cohort level. Panel B replicates Panel A, replacing the dependent variable with changes in mean college quality across adjacent cohorts.

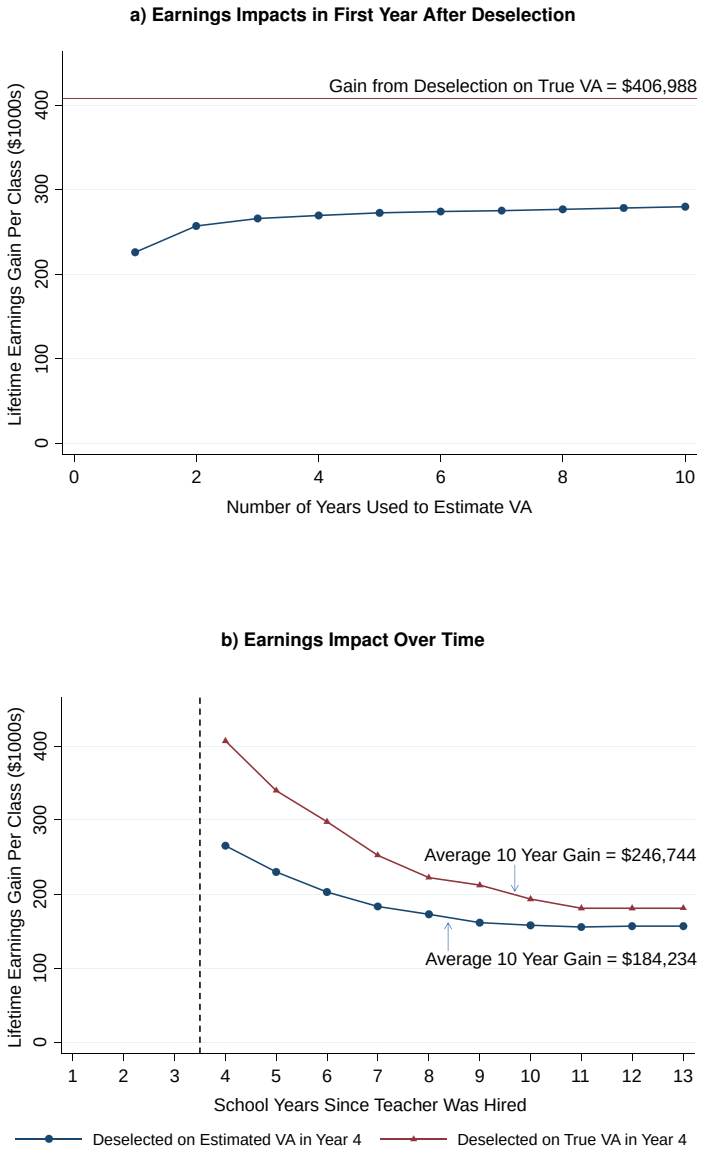
FIGURE 7

Impacts of Teacher Value-Added on College Quality by Grade



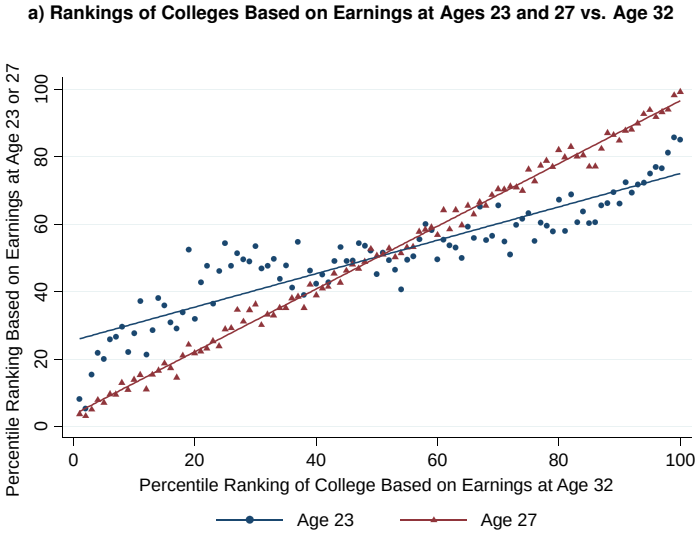
Notes: This figure plots the impact of a 1 SD increase in teacher VA in each grade from 4-8 on our earnings-based index of college quality (defined in the notes to Table 1). The upper (circle) series shows the reduced-form effect of improved teacher quality in each grade, including both the direct impact of the teacher on earnings and the indirect effect through improved teacher quality in future years. To generate this series, we replicate Column 5 of Table 2, interacting VA with grade dummies. We restrict the sample to cohorts who would have been in 4th grade during or after 1994 to obtain a balanced sample across grades. The dots in the series plot the coefficients on each grade interaction. The dashed lines show the boundaries of the 95% confidence intervals for the reduced-form effects, clustering the standard errors by school-cohort. The lower (triangle) 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 12 and solving the system of equations in Section 6.3. Appendix Table 11 reports the reduced-form effects and net-of-future-teachers effects plotted in this figure.

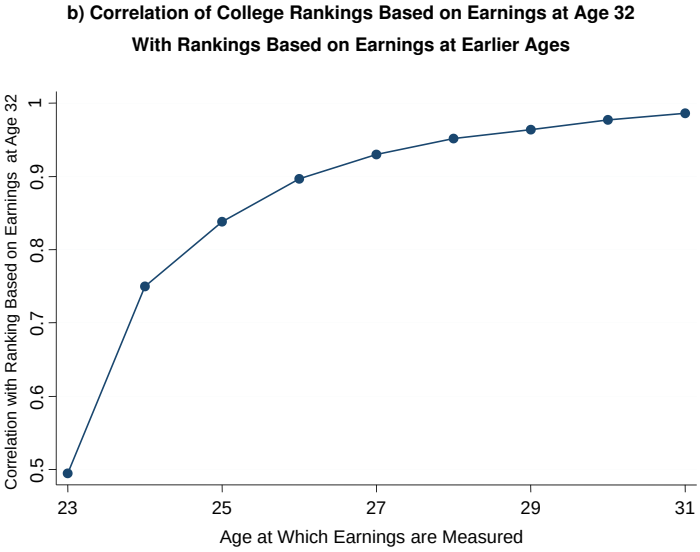
FIGURE 8
Earnings Impacts of Deselecting Low Value-Added Teachers



Notes: This figure analyzes the impacts of replacing teachers with VA in the bottom 5% with teachers of average quality on the present value of lifetime earnings for a single classroom of average size (28.2 students). In Panel A, the horizontal line shows the hypothetical gain from deselecting the bottom 5% of teachers based on their true VA in the current school year. The series in circles plots the gains from deselecting teachers on estimated VA vs. the number of years of prior test score data used to estimate VA. Panel A shows gains for the school year immediately after deselection; Panel B shows the gains in subsequent school years, which decay over time due to drift in teacher quality. The lower series in Panel B (in circles) plots the earnings gains in subsequent school years from deselecting teachers based on their VA estimate at $t = 4$, constructed using the past three years of data. The first point in this series (at $t = 4$) corresponds to the third point in Panel A by construction. The upper series (in triangles) shows the hypothetical gains obtained from deselecting the bottom 5% of teachers based on their true VA at $t = 4$; the first dot in this series matches the value in the horizontal line in Panel A. For both series in Panel B, we also report the unweighted mean gain over the first ten years after deselection. All values in these figures are based on our estimate that a 1 SD increase in true teacher VA increases earnings by 1.34% (Column 2 of Table 3). All calculations assume that teachers teach one class per year and report mean values for math and English teachers, which are calculated separately. We calculate earnings gains using Monte Carlo simulations based on our estimates of the teacher VA process as described in Section 7. All values in these figures and their undiscounted equivalents are reported in Appendix Table 13.

APPENDIX FIGURE 1
Stability of College Rankings by Age of Earnings Measurement

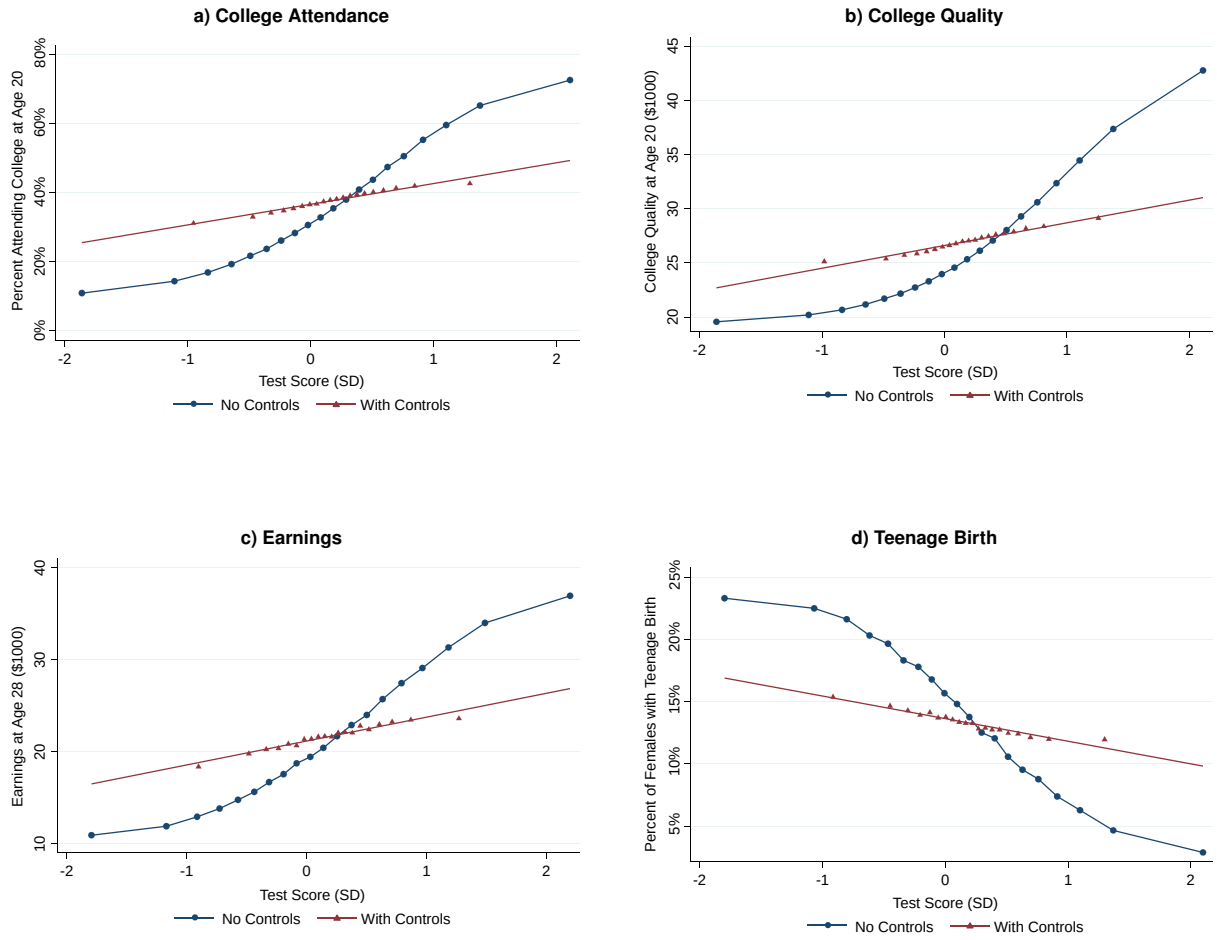




Notes: In Panel A, we take all college attendees in 1999 at age 20, as recorded by 1098-T forms, and construct three separate college quality indices by averaging W-2 earnings by college at ages 23, 27, and 32. We convert each college quality measure into a percentile rank based on the within-age distribution of college quality. We then bin colleges into 100 equal-sized (percentile) bins using the college quality measure based on age 32 earnings and plot the mean percentile rank of colleges in each bin using the age 23 (in circles) and age 27 (in triangles) measures. The best-fit lines are estimated from an unweighted OLS regression of percentile ranks run at the college level. In Panel B, we take the same college attendees and calculate ten separate college quality measures by averaging W-2 earnings by college at each age from 23-32. We then plot the Spearman rank correlation between each college quality measure based on earnings at ages 23-31 and the college quality measure based on earnings at age 32.

APPENDIX FIGURE 2

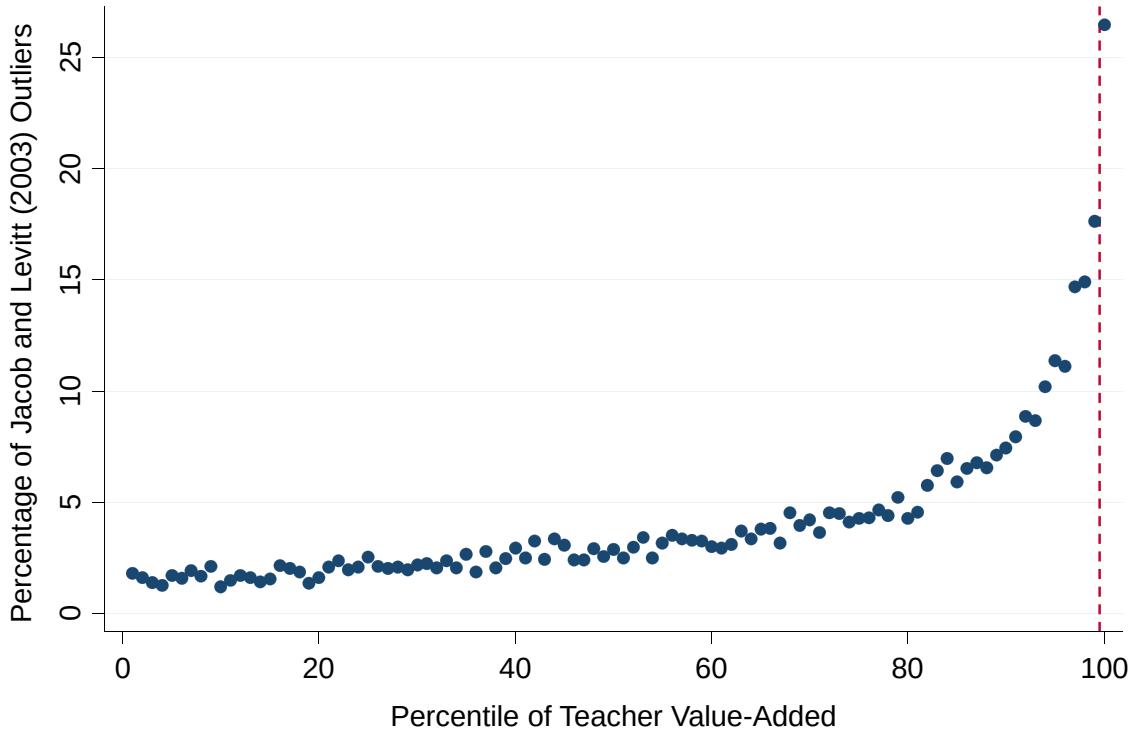
Correlations Between Outcomes in Adulthood and Test Scores



Notes: These figures present binned scatter plots corresponding to the cross-sectional regressions of outcomes in adulthood on test scores presented in Columns 1-4 of Appendix Table 3. See notes to Table 1 and Appendix Table 3 for further information on the variable definitions and sample specification. In each panel, the series in circles corresponds to the first row of estimates, without controls. The series in triangles corresponds to the second row of estimates, which includes the full control vector used to estimate the value-added model. To construct the series in circles, we bin raw test scores into twenty equal-sized groups (vingtiles) and plot the means of the outcome within each bin against the mean test score within each bin. To construct the series in triangles, we first regress both the test scores and adult outcomes on the individual and class controls and teacher fixed effects and compute residuals of both variables. We then divide the test score residuals into twenty equal-sized groups and plot the means of the outcome residuals within each bin against the mean test score residuals within each bin. Finally, we add back the unconditional mean of both test scores and the adult outcome in the estimation sample to facilitate interpretation of the scale. We connect the dots in the non-linear series without controls and show a best-fit line for the series with controls, estimated using an OLS regression on the microdata.

APPENDIX FIGURE 3

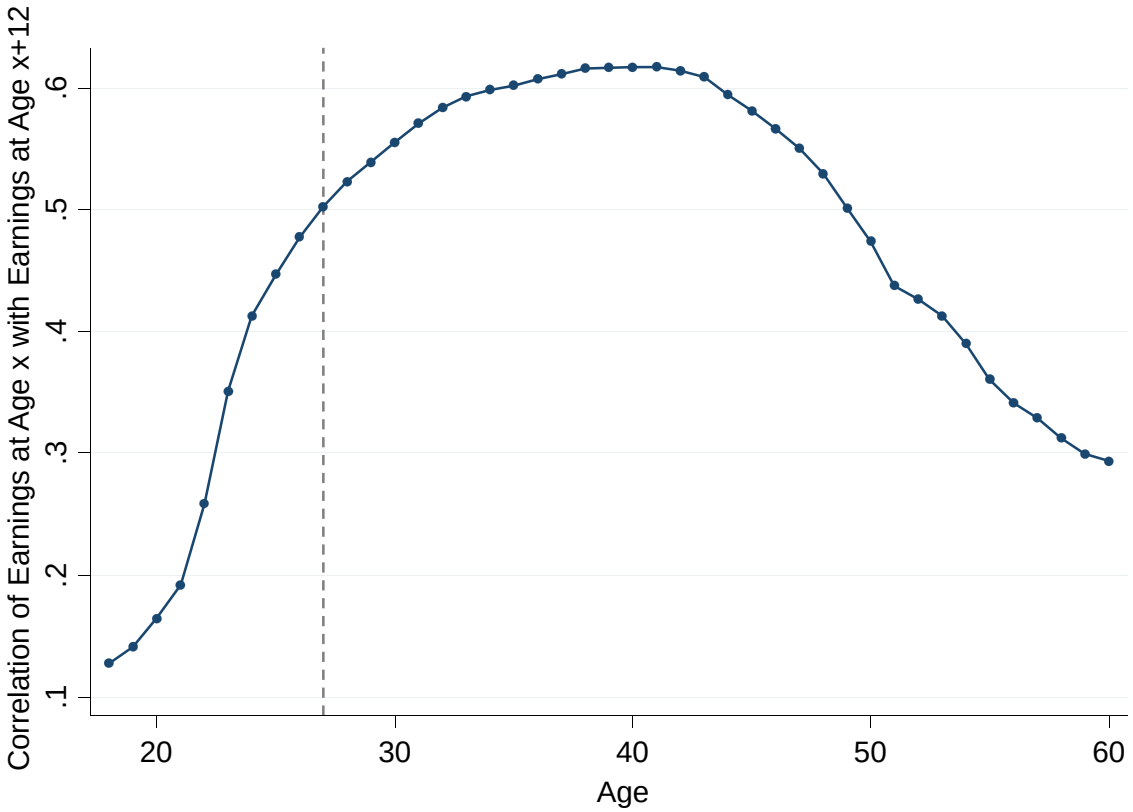
Jacob and Levitt (2003) Proxy for Test Manipulation vs. Value-Added Estimates



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 regressions are estimated on the linked analysis sample (as described in the notes to Table 1). Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure described in Section 4.1. 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 $t-1$ and 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 to be 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 classrooms separately by school-level and subject. We then compute the percentage 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 percentile of teacher VA. The dashed vertical line depicts the 99th 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.

APPENDIX FIGURE 4

Correlation of Earnings Over the Lifecyle



Notes: This figure plots the correlation of wage earnings at each age x with wage earnings at age $x + 12$. We calculate wage earnings as the sum of earnings reported on all W-2 forms for an individual in a given year. Individuals with no W-2 are assigned 0 wage earnings. Earnings at age x are calculated in 1999, the first year in which we have W-2 data, and earnings at age $x+12$ are calculated in 2011, the last year of our data. We calculate these correlations using the population of current U.S. citizens. The dashed vertical line denotes age 28, the age at which we measure earnings in our analysis of teachers' impacts.