# Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates

*By* RAJ CHETTY, JOHN N. FRIEDMAN, AND JONAH E. ROCKOFF*

*Are teachers' impacts on students' test scores ("value-added") a good measure of their quality? One reason this question has sparked debate is disagreement about whether value-added (VA) measures provide unbiased estimates of teachers' causal impacts on student achievement. We test for bias in VA using previously unobserved parent characteristics and a quasi-experimental design based on changes in teaching staff. Using school district and tax records for more than one million children, we find that VA models which control for a student's prior test scores exhibit little bias in forecasting teachers' impacts on student achievement.*

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 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?

* Chetty: Harvard University, Littauer Center 226, Cambridge MA 02138 (e-mail: chetty@fas.harvard.edu); Friedman: Harvard University, Taubman Center 356, Cambridge MA 02138 (e-mail: john_friedman@harvard.edu); Rockoff: Columbia University, Uris 603, New York NY 10027 (e-mail: jonah.rockoff@columbia.edu). We are indebted to Gary Chamberlain, Michal Kolesar, and Jesse Rothstein for many valuable discussions. We also thank Joseph Altonji, Josh Angrist, David Card, David Deming, Caroline Hoxby, Guido Imbens, Brian Jacob, Thomas Kane, Lawrence Katz, Adam Looney, Phil Oreopoulos, Douglas Staiger, Danny Yagan, anonymous referees, the editor, and numerous seminar participants for helpful comments. This paper is the first 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 by Gibson, Dunn, and Crutcher LLP to testify about the importance of teacher effectiveness for student learning in Vergara v. California based on the findings in NBER Working Paper No. 17699. John Friedman is currently on leave from Harvard, working at the National Economic Council; this work does not represent the views of the NEC. 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, Michael Stepner, and Evan Storms 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/va_bias_code.zip

[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).

This paper addresses the first of these two questions.[2] Prior work has reached conflicting conclusions about the degree of bias in VA estimates (Kane and Staiger 2008, Rothstein 2010, Kane et al. 2013). Resolving this debate is critical for policy because biased VA measures will systematically reward or penalize teachers based on the mix of students in their classrooms.

We develop new methods to estimate the degree of bias in VA estimates and implement them using information from two administrative databases. The first is a dataset on test scores and teacher assignments in grades 3-8 from a large urban school district in the U.S. These data cover more than 2.5 million students and include over 18 million tests of math and English achievement spanning 1989-2009. We match 90% of the observations in the school district data to selected data from United States tax records spanning 1996-2011. These data contain information on parent characteristics such as household income, retirement savings, and mother's age at child's birth.

We begin our analysis by constructing VA estimates for the teachers in our data. We predict each teacher's VA in a given school year based on the mean test scores of students she taught in other years. We control for student characteristics such as prior test scores and demographic variables when constructing this prediction to separate the teacher's impact from (observable) student selection. Our approach to estimating VA closely parallels that currently used by school districts, except in one respect. Existing value-added models typically assume that each teacher's quality is fixed over time and thus place equal weight on test scores in all classes taught by the teacher when forecasting teacher quality. In practice, test scores from more recent classes are better predictors of current teacher quality, indicating that teacher quality fluctuates over time. We account for such "drift" in teacher quality by estimating the autocovariance of scores across classrooms taught by a given teacher non-parametrically; intuitively, we regress scores in year $t$ on average scores in other years, allowing the coefficients to vary across different lags. Our VA model implies that a 1 standard deviation (SD) improvement in teacher VA raises normalized test scores by approximately 0.14 SD in math and 0.1 SD in English, slightly larger than the estimates in prior studies that do not account for drift.

Next, we turn to our central question: are the VA measures we construct "unbiased" predictors of teacher quality? To define "bias" formally, consider a hypothetical experiment in which we randomly assign students to teachers, as in Kane and Staiger (2008). Let $\lambda$ denote the mean test score impact of being randomly assigned to a teacher who is rated one unit higher in VA based on observational data from prior school years. We define the degree of "forecast bias" in a VA model as $B = 1 - \lambda$. We say that VA estimates are "forecast unbiased" if $B = 0$, i.e. if teachers whose estimated VA is one unit higher do in fact cause students' test scores to increase by one unit on average.

We develop two methods to estimate the degree of forecast bias in VA estimates. First, we estimate forecast bias based on the degree of selection on observable characteristics excluded from the VA model. We generate predicted test scores for each student based

[2] We address the second question in our companion paper (Chetty, Friedman, and Rockoff 2014). There are also other important concerns about VA. Most importantly, as with other measures of labor productivity, the signal in value-added measures may be degraded by behavioral responses if high-stakes incentives are put in place (Barlevy and Neal 2012).

on parent characteristics from the tax data, such as family income, and regress the predicted scores on teacher VA. For our baseline VA model – which controls for a rich set of prior student, class, and school level scores and demographics – we find that forecast bias from omitting parent characteristics is at most 0.3% at the top of the 95% confidence interval. Using a similar approach, we find that forecast bias from omitting twice-lagged scores from the VA model is at most 2.6%.

In interpreting these results, it is important to note that children of higher-income parents do get higher VA teachers on average. However, such sorting does not lead to biased estimates of teacher VA for two reasons. First, and most importantly, the correlation between VA estimates and parent characteristics vanishes once we control for test scores in the prior school year. Second, even the unconditional correlation between parent income and VA estimates is small: we estimate that a $10,000 increase in parent income is associated with less than a 0.0001 SD improvement in teacher VA (measured in student test-score SD's).[3] One explanation for why sorting is so limited is that 85% of the variation in teacher VA is within rather than between schools. Since most sorting occurs through the choice of schools, parents may have little scope to steer their children toward higher VA teachers.

While our first approach shows that forecast bias due to sorting on certain observable predictors of student achievement is minimal, bias due to other unobservable characteristics could still be substantial. To obtain a more definitive estimate of forecast bias that accounts for unobservables, we develop a quasi-experimental analog to the ideal experiment of random student assignment. Our quasi-experimental design exploits teacher turnover at the school-grade level for identification. To understand the design, suppose a high-VA 4th grade teacher moves from school *A* to another school in 1995. Because of this staff change, 4th graders in school *A* in 1995 will have lower VA teachers on average than the previous cohort of students in school *A*. If VA estimates have predictive content, we would expect 4th grade test scores for the 1995 cohort to be lower on average than the previous cohort.

Using event studies of teacher arrivals and departures, we find that mean test scores change sharply across cohorts as predicted when very high or low VA teachers enter or exit a school-grade cell. We estimate the amount of forecast bias by comparing changes in average test scores across consecutive cohorts of children within a school to changes in the mean value-added of the teaching staff. The forecasted changes in mean scores closely match observed changes. The point estimate of forecast bias in our preferred specification is 2.6% and is not statistically distinguishable from 0. The upper bound on the 95% confidence interval for the degree of bias is 9.1%.

Our quasi-experimental design rests on the identification assumption that high-frequency teacher turnover within school-grade cells is uncorrelated with student and school characteristics. This assumption is plausible insofar as parents are unlikely to immediately switch their children to a different school simply because a single teacher leaves or ar-

[3] An auxiliary implication of this result is that differences in teacher quality explain a small share of the achievement gap between high- and low-SES students. This is not because teachers are unimportant – one could close most of the achievement gap by assigning highly effective teachers to low-SES students – but rather because teacher VA does not differ substantially across schools in the district we study.

rives. Moreover, we show that changes in mean teacher quality in a given subject (e.g., math) are uncorrelated with both prior scores in that subject and *contemporaneous* scores in the other subject (e.g., English), supporting the validity of the research design.

We investigate which of the controls in our baseline VA model are most important to account for student sorting by estimating forecast bias using our quasi-experimental design for several commonly used VA specifications. We find that simply controlling for a student's own lagged test scores generates a point estimate of forecast bias of 5% that is not significantly different from 0. In contrast, models that omit lagged test score controls generate forecast bias exceeding 40%. Thus, most of the sorting of students to teachers that is relevant for future test achievement is captured by prior test scores. This result is reassuring for the application of VA models because virtually all value-added models used in practice control for prior scores.

Our quasi-experimental method provides a simple, low-cost tool for assessing bias in various settings.[4] For instance, Kane, Staiger, and Bacher-Hicks (2014) apply our method to data from the Los Angeles Unified School District (LAUSD). They find that VA estimates that control for lagged scores also exhibit no forecast bias in LAUSD, even though the dispersion in teacher VA is much greater in LAUSD than in the district we study. More generally, the methods developed here could be applied to assess the accuracy of personnel evaluation metrics in a variety of professions beyond teaching.

Our results reconcile the findings of experimental studies (Kane and Staiger 2008, Kane et al. 2013) with Rothstein's (2010) findings on bias in VA estimates. We replicate Rothstein's finding that there is small but statistically significant grouping of students on lagged test score gains and show that this particular source of selection generates minimal forecast bias. Based on his findings, Rothstein warns that selection on unobservables could *potentially* generate substantial bias. We directly evaluate the degree of forecast bias due to unobservables using a quasi-experimental analog of Kane and Staiger's (2008) experiment. Like Kane and Staiger, we find no evidence of forecast bias due to unobservables. Hence, we conclude that VA estimates which control for prior test scores exhibit little bias despite the grouping of students on lagged gains documented by Rothstein.

The paper is organized as follows. In Section I, we formalize how we construct VA estimates and define concepts of bias in VA estimates. Section II describes the data sources and provides summary statistics. We construct teacher VA estimates in Section III. Sections IV and V present estimates of forecast bias for our baseline VA model using the two methods described above. Section VI compares the forecast bias of alternative VA models. Section VII explains how our results relate to findings in the prior literature. Section VIII concludes.

## I. Conceptual Framework and Methods

In this section, we develop an estimator for teacher VA and formally define our measure of bias in VA estimates. We begin by setting up a simple statistical model of test

[4] Stata code to implement our technique is available at http://obs.rc.fas.harvard.edu/chetty/va_bias_code.zip

scores.

### A. Statistical Model

School principals assign each student $i$ in school year $t$ to a classroom $c = c(i,t)$. Principals then assign a teacher $j(c)$ to each classroom $c$. For simplicity, assume that each teacher teaches one class per year, as in elementary schools. Let $j = j(c(i,t))$ denote student $i$'s teacher in year $t$ and $\mu_{jt}$ represent the teacher's "value-added" in year $t$, i.e. teacher $j$'s impact on test scores. We scale teacher VA so that the average teacher has value-added $\mu_{jt} = 0$ and the effect of a 1 unit increase in teacher VA on end-of-year test scores is 1.

Student $i$'s test score in year $t$, $A^*_{it}$, is given by

$$A^*_{it} = \beta \mathbf{X}_{it} + \nu_{it} \tag{1}$$

$$\text{where } \nu_{it} = \mu_{jt} + \theta_c + \widetilde{\varepsilon}_{it} \tag{2}$$

Here, $\mathbf{X}_{it}$ denotes observable determinants of student achievement, such as lagged test scores and family characteristics. We decompose the error term $\nu_{it}$ into three components: teacher value-added $\mu_{jt}$, exogenous class shocks $\theta_c$, and idiosyncratic student-level variation $\widetilde{\varepsilon}_{it}$. Let $\varepsilon_{it} = \theta_c + \widetilde{\varepsilon}_{it}$ denote the unobserved error in scores unrelated to teacher quality. Student characteristics $\mathbf{X}_{it}$ and $\varepsilon_{it}$ may be correlated with $\mu_{jt}$. Accounting for such selection is the key challenge in obtaining unbiased estimates of $\mu_{jt}$.

The model in (1) permits teacher quality $\mu_{jt}$ to fluctuate stochastically over time. We do not place any restrictions on the stochastic processes that $\mu_{jt}$ and $\varepsilon_{it}$ follow except for the following assumption.

**Assumption 1 [Stationarity]** Teacher VA and student achievement follow a stationary process:

$$\mathbb{E}[\mu_{jt}|t] = \mathbb{E}[\varepsilon_{it}|t] = 0, Cov(\mu_{jt}, \mu_{j,t+s}) = \sigma_{\mu s}, Cov(\varepsilon_{it}, \varepsilon_{i,t+s}) = \sigma_{\varepsilon s} \text{ for all } t \tag{3}$$

Assumption 1 requires that (1) mean teacher quality does not vary across calendar years and (2) the correlation of teacher quality, class shocks, and student shocks across any pair of years depends only on the amount of time that elapses between those years. This assumption simplifies the estimation of teacher VA by reducing the number of parameters to be estimated. Note that the variance of teacher effects, $\sigma^2_\mu = Var(\mu_{jt})$, is constant across periods under stationarity.

### B. Estimating Teacher Value-Added

We develop an estimator for teacher value-added in year $t$ ($\mu_{jt}$) based on mean test scores in prior classes taught by teacher $j$.[5] Our approach closely parallels existing

[5]To maximize statistical precision, we use data from all other years – both in the past and future – to predict VA in year $t$ in our empirical implementation. To simplify notation, we present the derivation in this section for the case in which we only use prior data to predict VA.

estimators for value-added (e.g., Kane and Staiger 2008), except that it accounts for drift in teacher quality over time. To simplify exposition, we derive the estimator for the case in which data on test scores is available for $t$ years for all teachers, where all classes have $n$ students, and where each teacher teaches one class per year. In Appendix A, we provide a step-by-step guide to implementation (along with corresponding Stata code) that accounts for differences in class size, multiple classrooms per year, and other technical issues that arise in practice.

We construct our estimator in three steps. First, we regress test scores $A^*_{it}$ on $\mathbf{X}_{it}$ and compute test score residuals adjusting for observables. Next, we estimate the best linear predictor of mean test score residuals in classrooms in year $t$ based on mean test score residuals in prior years, using a technique analogous to an OLS regression. Finally, we use the coefficients of the best linear predictor to predict each teacher's VA in year $t$. We now describe these steps formally.

Let the residual student test score after removing the effect of observable characteristics be denoted by

$$A_{it} = A^*_{it} - \boldsymbol{\beta}\mathbf{X}_{it} = \mu_{jt} + \varepsilon_{it}. \tag{4}$$

We estimate $\boldsymbol{\beta}$ using variation across students taught by the same teacher using an OLS regression of the form

$$A^*_{it} = \alpha_j + \boldsymbol{\beta}\mathbf{X}_{it}, \tag{5}$$

where $\alpha_j$ is a teacher fixed effect. Our approach of estimating $\boldsymbol{\beta}$ using within teacher variation differs from prior studies, which typically use both within- and between-teacher variation to estimate $\boldsymbol{\beta}$ (e.g., Kane, Rockoff, and Staiger 2008, Jacob, Lefgren and Sims 2010). If teacher VA is correlated with $X_{it}$, estimates of $\boldsymbol{\beta}$ 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.[6] For example, suppose $\mathbf{X}$ includes school fixed effects. Estimating $\boldsymbol{\beta}$ without teacher fixed effects would attribute all the test score differences across schools to the school fixed effects, leaving mean teacher quality normalized to be the same across all schools. With school fixed effects, estimating $\boldsymbol{\beta}$ within teacher requires a set of teachers to teach in multiple schools, as in Mansfield (2013). These switchers allow us to identify the school fixed effects independent of teacher effects and obtain a cardinal global ranking of teachers across schools.

Let $\bar{A}_{jt} = \frac{1}{n}\sum_{i\in\{i:j(i,t)=j\}} A_{it}$ denote the mean residual test score in the class teacher $j$ teaches in year $t$. Let $\mathbf{A}_j^{-t} = (\bar{A}_{j1}, ..., \bar{A}_{j,t-1})'$ denote the vector of mean residual scores prior to year $t$ in classes taught by teacher $j$. Our estimator for teacher $j$'s VA in year $t$

[6] Teacher fixed effects account for correlation between $X_{it}$ and mean teacher VA. If $X_{it}$ is correlated with fluctuations in teacher VA across years due to drift, then one may still understate teachers' effects even with fixed effects. We show in Table 6 below that dropping teacher fixed effects when estimating (5) yields VA estimates that have a correlation of 0.98 with our baseline estimates because most of the variation in $X_{it}$ is within classrooms. Since sorting to teachers based on their average impacts turns out to be quantitatively unimportant in practice, sorting based on fluctuations in those impacts is likely to have negligible effects on VA estimates.

($\mathbb{E}\left[\mu_{jt} \mid \mathbf{A}_j^{-t}\right]$) is the best linear predictor of $\bar{A}_{jt}$ based on prior scores ($\mathbb{E}^*\left[\bar{A}_{jt} \mid \mathbf{A}_j^{-t}\right]$), which we write as

$$\widehat{\mu}_{jt} = \sum_{s=1}^{t-1} \psi_s \bar{A}_{js}.$$

We choose the vector of coefficients $\psi = (\psi_1, ..., \psi_{t-1})'$ to minimize the mean-squared error of the forecasts of test scores:

$$\psi = \arg \min_{\{\psi_1,...,\psi_{t-1}\}} \sum_j \left( \bar{A}_{jt} - \sum_{s=1}^{t-1} \psi_s \bar{A}_{js} \right)^2. \tag{6}$$

The resulting coefficients $\psi$ are equivalent to those obtained from an OLS regression of $\bar{A}_{jt}$ on $\mathbf{A}_j^{-t}$. In particular, $\psi = \Sigma_A^{-1}\gamma$ where $\gamma = (Cov(\bar{A}_{jt}, \bar{A}_{j1}), ..., Cov(\bar{A}_{jt}, \bar{A}_{j,t-1}))'$ is the vector of auto-covariances of mean test scores for classes taught by a given teacher and $\Sigma_A$ is the variance-covariance (VCV) matrix of $\mathbf{A}_j^{-t}$. The diagonal elements of $\Sigma_A$ are the variance of mean class scores $\sigma_A^2$. The off-diagonal elements are the covariance of mean test scores of different classes taught by a given teacher $Cov(\bar{A}_{jt}, \bar{A}_{j,t-s})$. Note that $Cov(\bar{A}_{jt}, \bar{A}_{j,t-s}) = \sigma_{As}$ depends only on the time lag $s$ between the two periods under the stationarity assumption in (3).

Finally, using the estimates of $\psi$, we predict teacher $j$'s VA in period $t$ as

$$\widehat{\mu}_{jt} = \psi' \mathbf{A}_j^{-t}. \tag{7}$$

Note that the VA estimates in (7) are leave-year-out (jackknife) measures of teacher quality, similar to those in Jacob, Lefgren and Sims (2010), but with an adjustment for drift. This is important for our analysis below because regressing student outcomes in year $t$ on teacher VA estimates without leaving out the data from year $t$ when estimating VA would introduce the same estimation errors on both the left and right hand side of the regression and produce biased estimates of teachers' causal impacts.[7]

*Special Cases.* The formula for $\widehat{\mu}_{jt}$ in (7) nests two special cases that help build intuition for the general case. First, consider predicting a teacher's impact in year $t$ using data from only the previous year $t-1$. In this case, $\gamma = \sigma_{A,1}$ and $\Sigma_A^{-1} = \frac{1}{\sigma_A^2}$. Hence, $\psi$ simplifies to $\frac{\sigma_{A,1}}{\sigma_A^2}$ and

$$\widehat{\mu}_{jt} = \psi \bar{A}_{j,t-1}. \tag{8}$$

The shrinkage factor $\psi$ in this equation incorporates two forces that determine $\psi$ in the general case. First, because past test scores are a noisy signal of teacher quality, the VA estimate is shrunk toward the sample mean ($\mu_{jt} = 0$) to reduce mean-squared error.

[7]This is the reason that Rothstein (2010) finds that "fifth grade teachers whose students have above average fourth grade gains have systematically lower estimated value-added than teachers whose students underperformed in the prior year." Students who had unusually high fourth grade gains due to idiosyncratic, non-persistent factors (e.g., measurement error) will tend to have lower than expected fifth grade gains, making their fifth grade teacher have a lower VA estimate.

Second, because teacher quality drifts over time, the predicted effect differs from past performance. For instance, if teacher quality follows a mean-reverting process, past test scores are further shrunk toward the mean to reduce the influence of transitory shocks to teacher quality.

Second, consider the case where teacher quality is fixed over time ($\mu_{jt} = \mu_j$ for all $t$) and the student and class level errors are i.i.d. This is the case considered by most prior studies of value-added. Here, $Cov(\bar{A}_{jt}, \bar{A}_{j,t-s}) = Cov(\mu_j, \mu_j) = \sigma^2_\mu$ for all $s \neq t$ and $\sigma^2_A = \sigma^2_\mu + \sigma^2_\theta + \sigma^2_{\bar{\varepsilon}}/n$. In this case, (7) simplifies to

$$\widehat{\mu}_{jt} = \bar{A}_j^{-t} \frac{\sigma^2_\mu}{\sigma^2_\mu + (\sigma^2_\theta + \sigma^2_{\bar{\varepsilon}}/n)/(t-1)}. \tag{9}$$

where $\bar{A}_j^{-t}$ is the mean residual test score in classes taught by teacher $j$ in years prior to $t$ and $\psi = \frac{\sigma^2_\mu}{\sigma^2_\mu + (\sigma^2_\theta + \sigma^2_{\bar{\varepsilon}}/n)/(t-1)}$ is the "reliability" of the VA estimate. This formula coincides with Equation 5 in Kane and Staiger (2008) in the case with constant class size.[8] Here, the signal to noise ratio $\psi$ does not vary across years because teacher performance in any year is equally predictive of performance in year $t$. Because years are interchangeable, VA depends purely on mean test scores over prior years, again shrunk toward the sample mean to reduce mean-squared error.[9]

Importantly, $\widehat{\mu}_{jt} = \mathbb{E}^*\left[\bar{A}_{jt} \mid \mathbf{A}_j^{-t}\right]$ simply represents the best linear predictor of the future test scores of students assigned to teacher $j$ in observational data. This prediction does not necessarily measure the expected causal effect of teacher $j$ on students' scores in year $t$, $\mathbb{E}\left[\mu_{jt} \mid \mathbf{A}_j^{-t}\right]$, because part of the prediction could be driven by systematic sorting of students to teachers. We now turn to evaluating the degree to which $\widehat{\mu}_{jt}$ measures a teacher's causal impact.

### *C. Definition of Bias*

An intuitive definition of bias is to ask whether VA estimates $\widehat{\mu}_{jt}$ accurately predict differences in the mean test scores of students who are randomly assigned to teachers in year $t$, as in Kane and Staiger (2008). Consider an OLS regression of residual test scores $A_{it}$ in year $t$ on $\widehat{\mu}_{jt}$ (constructed from observational data in prior years) in such an experiment:

$$A_{it} = \alpha_t + \lambda \widehat{\mu}_{jt} + \zeta_{it} \tag{10}$$

[8] Kane and Staiger (2008) derive (9) using an Empirical Bayes approach instead of a best linear predictor. If teacher VA, class shocks, and student errors follow independent Normal distributions, the posterior mean of $\mu_{jt}$ coincides with (9). Analogously, (7) can be interpreted as the posterior expectation of $\mu_{jt}$ when teacher VA follows a multivariate Normal distribution whose variance-covariance matrix controls the drift process.

[9] In our original working paper (Chetty, Friedman, and Rockoff 2011b), we estimated value-added using (9). Our qualitative conclusions were similar, but our out-of-sample forecasts of teachers' impacts were attenuated because we did not account for drift.

Because $\mathbb{E}\left[\varepsilon_{it} \mid \widehat{\mu}_{jt}\right] = 0$ under random assignment of students in year $t$, the coefficient $\lambda$ measures the relationship between true teacher effects $\mu_{jt}$ and estimated teacher effects $\widehat{\mu}_{jt}$:

$$\lambda \equiv \frac{Cov\left(A_{it}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)} = \frac{Cov\left(\mu_{jt}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)}. \tag{11}$$

We define the degree of bias in VA estimates based on this regression coefficient as follows.

**Definition 1.** The amount of *forecast bias* in a VA estimator $\widehat{\mu}_{jt}$ is $B\left(\widehat{\mu}_{jt}\right) = 1 - \lambda$.

Forecast bias determines the mean impact of changes in the estimated VA of the teaching staff. A policy that increases estimated teacher VA $\widehat{\mu}_{jt}$ by 1 standard deviation raises student test scores by $(1 - B)\,\sigma(\widehat{\mu}_{jt})$, where $\sigma(\widehat{\mu}_{jt})$ is the standard deviation of VA estimates scaled in units of student test scores. If $B = 0$, $\widehat{\mu}_{jt}$ provides an unbiased forecast of teacher quality in the sense that an improvement in estimated VA of $\Delta\widehat{\mu}_{jt}$ has the same causal impact on test scores as an increase in true teacher VA $\Delta\mu_{jt}$ of the same magnitude.[10]

Two issues should be kept in mind when interpreting estimates of forecast bias. First, VA estimates can be forecast-unbiased even if children with certain characteristics (e.g., those with higher ability) are systematically assigned to certain teachers. Forecast unbiasedness only requires that the observable characteristics $\mathbf{X}_{it}$ used to construct test score residuals $A_{it}$ are sufficiently rich that the remaining unobserved heterogeneity in test scores $\varepsilon_{it}$ is balanced across teachers with different VA estimates. Second, even if VA estimates $\widehat{\mu}_{jt}$ are forecast unbiased, one can potentially improve forecasts of $\mu_{jt}$ by incorporating other information beyond past test scores. For example, data on principal ratings or teacher characteristics could potentially reduce the mean-squared error of forecasts of teacher quality. Hence, the fact that a VA estimate is forecast unbiased does not necessarily imply that it is the optimal forecast of teacher quality given all the information that may be available to school districts.

An alternative definition of bias, which we term "teacher-level bias," asks whether the VA estimate for each teacher converges to her true quality as estimation error vanishes, as in Rothstein (2009). We define teacher-level bias and formalize its connection to forecast bias in Appendix B. Teacher-level unbiasedness is more challenging to estimate and is a stronger requirement than forecast unbiasedness: some teachers could be systematically overrated relative to others even if forecasts based on VA estimates are accurate on average. However, forecast-unbiased VA estimates can be biased at the teacher level only in the knife-edge case in which the covariance between the bias in each teacher's VA estimate and true VA perfectly offsets the variance in true VA (see Appendix B). Hence,

[10] Ex-post, any estimate of VA can be made forecast unbiased by redefining VA as $\widehat{\mu}'_{jt} = (1 - B)\widehat{\mu}_{jt}$. However, the causal effect of a policy that raises the estimated VA of teachers is unaffected by such a rescaling, as the effect of a 1 SD improvement in $\widehat{\mu}'_{jt}$ is still $(1 - B)\,\sigma(\widehat{\mu}_{jt})$. Hence, given the standard deviation of VA estimates, the degree of forecast bias is informative about the potential impacts of improving estimated VA.

if VA estimates obtained from a pre-specified value-added model turn out to be forecast unbiased, they are unlikely to be biased at the teacher level.[11] We therefore focus on estimating $B$ and testing if VA estimates from existing models are forecast unbiased in the remainder of the paper.

## II. Data

We draw information from two databases: administrative school district records and federal income tax records. This section describes the two data sources and the structure of the linked analysis dataset and then provides descriptive statistics.

### *A. School District Data*

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

*Test Scores.* The data include approximately 18 million test scores. Test scores are available for English language arts and math for students in grades 3-8 in every year from the spring of 1989 to 2009, with the exception of 7th grade English scores in 2002.[12]

The testing regime varies over the 20 years we study. In the early and mid 1990s, all tests were specific to the district. Starting at the end of the 1990s, the tests in grades 4 and 8 were administered as part of a statewide testing system, and all tests in grades 3-8 became statewide in 2006 as required under the No Child Left Behind law. All tests were administered in late April or May during the early and mid 1990s. Statewide tests were sometimes given earlier in the school year (e.g., February) during the latter years of our data.

Because of this variation in testing regimes, we follow prior work by normalizing the official scale scores from each exam to have mean zero and standard deviation one by year and grade. The within-grade variation in achievement in the district we study is comparable to the within-grade variation nationwide, so our results can be compared to estimates from other samples.[13]

*Demographics.* The dataset contains information on ethnicity, gender, age, receipt of special education services, and limited English proficiency for the school years 1989

[11] This logic relies on having a pre-specified VA model that is not changed ex-post after observing estimates of forecast bias. If we were to redefine our VA estimates after estimating forecast bias as $\widehat{\mu}'_{jt} = (1 - B)\widehat{\mu}_{jt}$, then $\widehat{\mu}'_{jt}$ would be forecast unbiased by construction. However, teacher-level bias in $\widehat{\mu}'_{jt}$ is no longer a knife-edge case because we have introduced a free parameter into the VA estimation procedure that is mechanically chosen to offset the excess variance created by teacher-level bias.

[12] We also have data on math and English test scores in grade 2 from 1991-1994, which we use only when evaluating sorting on lagged test score gains. Because these observations constitute a very small fraction of our sample, excluding them has little impact on our results.

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

through 2009. The database used to code special education services and limited English proficiency changed in 1999, creating a break in these series that we account for in our analysis by interacting these two measures with a post-1999 indicator. Information on free and reduced price lunch is available starting in school year 1999. These missing data issues are not a serious problem because our estimates of forecast bias are insensitive to excluding demographic characteristics from the VA model entirely.

*Teachers.* The dataset links students in grades 3-8 to classrooms and teachers from 1991 through 2009.[14] This information is derived from a data management system which was phased in over the early 1990s, so not all schools are included in the first few years of our sample. In addition, data on course teachers for middle and junior high school students—who, unlike students in elementary schools, are assigned different teachers for math and English—are more limited. Course teacher data are unavailable prior to the school year 1994, then grow in coverage to roughly 60% by 1998, and stabilize at approximately 85% after 2003. To ensure that our estimates are not biased by the missing data, we show that our conclusions remain very similar in a subsample of school-grade-subject cells with no missing data (see Table 5 below).

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

### *B. Tax Data*

We obtain information on parent characteristics from U.S. federal income tax returns spanning 1996-2011.[15] The school district records were linked to the tax data using an algorithm based on standard identifiers (date of birth, state of birth, gender, and names) described in Appendix C, after which individual identifiers were removed to protect confidentiality. 88.6% of the students and 89.8% of student-subject-year observations in the

[14] 5% of students switch classrooms or schools in the middle of a school year. We assign these students to the classrooms in which they took the test to obtain an analysis dataset with one observation per student-year-subject. However, when defining class and school-level means of student characteristics (such as fraction eligible for free lunch), we account for such switching by weighting students by the fraction of the year they spent in that class or school.

[15] 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 $t + 1$.

sample used to estimate value-added were matched to the tax data, i.e. uniquely linked to a social security record or tax form as described in Appendix C. Students were then linked to parents based on the earliest 1040 form filed between tax years 1996 and 2011 on which the student was claimed as a dependent. We identify parents for 97.6% of the observations in the analysis dataset conditional on being matched to the tax data.[16] We are able to link almost all children to their parents because almost all parents file a tax return at some point between 1996 and 2012 to obtain a tax refund on their withheld taxes and the Earned Income Tax Credit.

In this paper, we use the tax data to obtain information on five time-invariant parent characteristics, defined as follows. We define parental household income as mean Adjusted Gross Income (capped at $117,000, the 95th percentile in our sample) between 2005 and 2007 for the primary filer who first claimed the child; measuring parent income in other years yields very similar results (not reported). For years in which parents did not file a tax return, they are assigned an income of 0. We measure income in 2010 dollars, adjusting for inflation using the Consumer Price Index.

We define marital status, home ownership, and 401(k) saving as indicators for whether the first primary filer who claims the child ever files a joint tax return, makes a mortgage interest payment (based on data from 1040's for filers and 1099's for non-filers), or makes a 401(k) contribution (based on data from W-2's) between 2005 and 2007.

We define mother's age at child's birth using data from Social Security Administration records on birth dates for parents and children. For single parents, we define the mother's age at child's birth using the age of the filer who first claimed the child, who is typically the mother but is sometimes the father or another relative.[17] When a child cannot be matched to a parent, we define all parental characteristics as zero, and we always include a dummy for missing parents in regressions that include parent characteristics.

### *C. Summary Statistics*

The linked school district and tax record analysis dataset has one row per student per subject (math or English) per school year, as illustrated in Appendix Table 1. Each observation in the analysis dataset contains the student's test score in the relevant subject test, demographic information, teacher assignment, and time-invariant parent characteristics. We organize the data in this format so that each row contains information on a treatment by a single teacher conditional on pre-determined characteristics. We account for the fact that each student appears multiple times in the dataset by clustering standard errors as described in Section III.

After imposing the sample restrictions described above, the linked dataset contains 10.7 million student-year-subject observations. We use this "core sample" of 10.7 million observations to construct quasi-experimental estimates of forecast bias, which do not

[16] Consistent with this statistic, Chetty et al. (2014, Appendix Table I) show that virtually all children in the U.S. population are claimed as dependents at some point during their childhood. Note that our definition of parents is based on who claims the child as a dependent, and thus may not reflect the biological parent of the child.

[17] We set the mother's age at child's birth to missing for 78,007 observations in which the implied mother's age at birth based on the claiming parent's date of birth is below 13 or above 65, or where the date of birth is missing entirely from SSA records.

require any additional controls. 9.1 million records in the core sample have information on teacher assignment and 7.6 million have information on teachers, lagged test scores, and the other controls needed to estimate our baseline VA model.

Table 1 reports summary statistics for the 7.6 million observation sample used to estimate value-added (which includes those with missing parent characteristics). 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 1.37 million students in this sample and each student has 5.6 subject-school year observations on average.

The mean test score in the sample used to estimate VA 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). 80% of students are eligible for free or reduced price lunches. For students whom we match to parents, mean parent household income is $40,800, while the median is $31,700. 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 $34,300, with 10% of parents earning more than $100,000.

## III. Value-Added Estimates

We estimate teacher VA using the methodology in Section I.B in three steps: (1) construct student test score residuals, (2) estimate the autocovariance of scores across classes taught by a given teacher, and (3) predict VA for each teacher in each year using test score data from other years.

*Test Score Residuals.* Within each subject (math and English) and school-level (elementary and middle), we construct test score residuals $A_{it}$ by regressing raw standardized test scores $A^*_{it}$ on a vector of covariates $\mathbf{X}_{it}$ and teacher fixed effects, as in (5). Our baseline control vector $\mathbf{X}_{it}$ is similar to Kane, Rockoff, and Staiger (2008). We control for prior test scores using a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English, and we interact these cubics with the student's grade level to permit flexibility in the persistence of test scores as students age.[18] We also control for students' ethnicity, gender, age, lagged suspensions and absences, and indicators for special education, limited English proficiency, and grade repetition. We also include the following class- and school-level controls: (1) cubics in class and school-grade means of prior-year test scores in math and English (defined based on those with non-missing prior scores) each interacted with grade, (2) class and school-year means of all the other individual covariates, (3) class size and class-type indicators (honors, remedial), and (4)

[18] We exclude observations with missing data on current or prior scores in the subject for which we are estimating VA. We also exclude classrooms that have fewer than 7 observations with current and lagged scores in the relevant subject (2% of observations) to avoid estimating VA based on very few observations. When prior test scores in the other subject are missing, we set the other subject prior score to 0 and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores.

grade and year dummies. Importantly, our control vector $\mathbf{X}_{it}$ consists entirely of variables from the school district dataset.[19]

*Auto-Covariance Vector.* Next, we estimate the auto-covariance of mean test score residuals across classes taught in different years by a given teacher, again separately for each subject and school-level. Exploiting the stationarity assumption in (3), we use all available classrooms with a time span of $s$ years between them to estimate $\sigma_{As} = Cov(\bar{A}_{jt}, \bar{A}_{jt-s})$, weighting by the total number of students in each pair of classes. In middle school, where teachers teach multiple classes per year, we estimate $\sigma_{As}$ after collapsing the data to the teacher-year level by calculating precision-weighted means across the classrooms as described in Appendix A.

Figure 1 plots the autocorrelations $r_s = \frac{\sigma_{As}}{\sigma_A^2}$ for each subject and school level; the values underlying this figure along with the associated covariances are reported in Panel A of Table 2. If one were using data only from year $s$, the VA estimate for teacher $j$ in year $t$ would simply be $\widehat{\mu}_{jt} = r_s A_{j,t-s}$, since $r_s = \psi = \frac{\sigma_{As}}{\sigma_A^2}$, as shown in (8). Hence, $r_s$ represents the "reliability" of mean class test scores for predicting teacher quality $s$ years later. The reliability of VA estimates decays over time: more recent test scores are better predictors of current teacher performance. For example, in elementary school, reliability declines from $r_1 = 0.43$ in the first year to $r_7 = 0.25$ in math after seven years, and remains roughly stable for $s > 7$. The decay in $r_s$ over a period of 5-7 years followed by stability beyond that point implies that teacher quality includes both a permanent component and a transitory component. The transitory component may arise from variation in teaching assignments, curriculum, or skill (Goldhaber and Hansen 2010, Jackson 2010).[20]

In middle school, we can estimate the within-year covariance of test score residuals $\sigma_{A0}$ because teachers teach multiple classes per year. Under the assumption that class and student level shocks are i.i.d., this within-year covariance corresponds to the variance of teacher effects. We estimate the standard deviation of teacher effects, $\sigma_\mu = \sqrt{\sigma_{A0}}$, to be 0.098 in English and 0.134 in math in middle school (Table 2, Panel B).[21] In elementary school, $\sigma_{A0}$ is unidentified because teachers teach only one class per year. Conceptually, we cannot distinguish the variance of idiosyncratic class shocks from unforecasted innovations in teacher effects when teachers teach only one class per year. However, we can obtain a lower bound on $\sigma_\mu$ of $\sqrt{\sigma_{A1}}$ because $\sigma_{A0} \geq \sigma_{A1}$. This bound implies that $\sigma_\mu$ is at least 0.113 in English and at least 0.149 in math in elementary school (second-to-last row of Panel B of Table 2). To obtain a point estimate of $\sigma_\mu$, we

[19] We do not control for teacher experience in our baseline specification because doing so complicates our quasi-experimental teacher switching analysis, as we would have to track changes in both teacher experience and VA net of experience. However, we show in Table 6 below that the correlation between our baseline VA estimates and VA estimates that control for experience is 0.99.

[20] Prior studies, which do not account for drift, typically estimate reliability as the correlation between mean scores for a random pair of classes in different years (e.g., Kane and Staiger 2008, Chetty, Friedman, and Rockoff 2011b). That method yields an estimate of the mean value of $r_s$ over the years available in the dataset. A recent summary (McCaffrey et al. 2009) finds reliability in the range of 0.2–0.5 for elementary school and 0.3–0.7 for middle school teachers, consistent with the estimates in Figure 1.

[21] As is standard in the literature, in this paper we scale value-added in units of *student* test scores, i.e., a 1 unit increase in value-added refers to a teacher whose VA is predicted to raise student test scores by 1 SD.

fit a quadratic function to the log of the first seven covariances within each subject in elementary school (listed in Table 2, Panel A) and extrapolate to 0 to estimate $\sigma_{A0}$. This method yields estimates of $\sigma_{\mu}$ of 0.124 in English and 0.163 in math, as shown in the final row of Table 2, Panel B.[22] The point estimates are close to the lower bounds because the rate of drift across one year is small. Our estimates of the standard deviations of teacher effects are slightly larger than those in prior studies (e.g., Kane, Rockoff, and Staiger 2008, Chetty, Friedman, and Rockoff 2011b) because the earlier estimates were attenuated by drift.

*Prediction of VA.* We use the estimated autocovariance vectors in Table 2 to predict teacher VA. Since there is no trend in reliability after 7 periods and because the precision of the estimates falls beyond that point, we fix $\sigma_{As} = \sigma_{A7}$ for $s > 7$ in all subject and school levels when estimating VA. We predict each teacher's VA in each year $t$ using test score residuals from all years (both past and the future) *except* year $t$. For example, when predicting teachers' effects on student outcomes in year $t = 1995$ ($\widehat{\mu}_{j,1995}$), we estimate VA based on all years excluding 1995. We construct these estimates using the formula in (7) with an adjustment for the variation in the number of students per class and the number of classes per year (for middle school) as described in Appendix A.[23]

The empirical distributions of our teacher VA estimates are plotted in Appendix Figure 1. The standard deviation of $\widehat{\mu}_{jt}$ is 0.116 in math and 0.080 in English in elementary school; the corresponding SD's are 0.092 and 0.042 in middle school. The standard deviations are smaller than the true SD of teacher effects because the best linear predictor shrinks VA estimates toward the sample mean to minimize mean-squared-error.

*Out-of-Sample Forecasts.* Under the stationarity assumption in (3), an OLS regression of $A_{it}$ on $\widehat{\mu}_{jt}$ – the best-linear predictor of $A_{it}$ – should yield a coefficient of 1 by construction. We confirm that this is the case in Column 1 of Table 3, which reports estimates from a univariate OLS regression of test score residuals $A_{it}$ on $\widehat{\mu}_{jt}$ in the sample used to estimate the VA model. We include fixed effects for subject (math vs. English) by school-level (elementary vs. middle) in this and all subsequent regressions to obtain a single regression coefficient that is identified purely from variation within the subject-by-school-level cells. We cluster standard errors at the school-by-cohort level (where cohort is defined as the year in which a child entered kindergarten) to adjust for correlated errors across students within classrooms and the multiple observations for each student in different grades and subjects.[24] The point estimate of the coefficient on $\widehat{\mu}_{jt}$ is 0.998 and the 95% confidence interval is (0.986, 1.010).

Figure 2a plots the relationship between $A_{it}$ and $\widehat{\mu}_{jt}$ non-parametrically, dividing the

[22] Applying the same quadratic extrapolation to the middle school data yields estimates of $\sigma_{\mu}$ of 0.134 in math and 0.079 in English. These estimates are fairly close to the true values estimated from the within-year covariance, supporting the extrapolation in elementary school.

[23] Among the classrooms with the requisite controls to estimate value-added (e.g., lagged test scores), we are unable to predict teacher VA for 9% of student-subject-year observations because their teachers are observed in the data for only one year.

[24] In our original working paper, we evaluated the robustness of our results to alternative forms of clustering (Chetty, Friedman, and Rockoff 2011b, Appendix Table 7). 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 (Cameron, Gelbach, and Miller 2011).

VA estimates $\widehat{\mu}_{jt}$ into twenty equal-size groups (vingtiles) and plotting the mean value of $A_{it}$ in each bin.[25] 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 micro data (not the binned averages), with standard errors clustered by school-cohort. The conditional expectation function in Figure 2a is almost perfectly linear. Teacher VA has a 1-1 relationship with test score residuals throughout the distribution, showing that the linear prediction model fits the data well.

The relationship between $\widehat{\mu}_{jt}$ and students' test scores in Figure 2a could be driven by the causal impact of teachers on achievement ($\mu_{jt}$) or persistent differences in student characteristics across teachers ($\varepsilon_{it}$). For instance, $\widehat{\mu}_{jt}$ may forecast students' test scores in other years simply because some teachers are always assigned students with higher or lower income parents. In the next two sections, we estimate the degree to which the relationship in Figure 2a reflects teachers' causal effects vs. bias due to student sorting.

## IV. Estimating Bias Using Observable Characteristics

We defined forecast bias in Section I.C under the assumption that students were randomly assigned to teachers in the forecast year $t$. In this section, we develop a method of estimating forecast bias using observational data, i.e., when the school follows the same (non-random) assignment rule for teachers and students in the forecast period $t$ as in previous periods. We first derive an estimator for forecast bias based on selection on observable characteristics and then implement it using data on parent characteristics and lagged test score gains.

### *A. Methodology*

In this subsection, we show that forecast bias can be estimated by regressing predicted test scores based on observable characteristics excluded from the VA model (such as parent income) on VA estimates. To begin, observe that regressing test score residuals $A_{it}$ on $\widehat{\mu}_{jt}$ in observational data yields a coefficient of 1 because $\widehat{\mu}_{jt}$ is the best linear predictor of $A_{it}$:

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

It follows from the definition of forecast bias that in observational data,

$$B\left(\widehat{\mu}_{jt}\right) = \frac{Cov\left(\varepsilon_{it}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)}. \tag{12}$$

[25] In this and all subsequent scatter plots, we first demean the x and y variables within subject by school level groups to isolate variation within these cells, as in the regressions.

Intuitively, the degree of forecast bias can be quantified by the extent to which students are sorted to teachers based on unobserved determinants of achievement $\varepsilon_{it}$.

Although we cannot observe $\varepsilon_{it}$, we can obtain information on components of $\varepsilon_{it}$ using variables that predict test score residuals $A_{it}$ but were omitted from the VA model, such as parent income. Let $\mathbf{P}^*_{it}$ denote a vector of such characteristics and $\mathbf{P}_{it}$ denote the residuals obtained after regressing the elements of $\mathbf{P}^*_{it}$ on the baseline controls $\mathbf{X}_{it}$ in a specification with teacher fixed effects, as in (5). Decompose the error in score residuals $\varepsilon_{it} = \boldsymbol{\rho}\mathbf{P}_{it} + \varepsilon'_{it}$ into the component that projects onto $\mathbf{P}_{it}$ and the remaining (unobservable) error $\varepsilon'_{it}$. To estimate forecast bias using $\mathbf{P}$, we make the following assumption.

**Assumption 2 [Selection on Excluded Observables]** Students are sorted to teachers purely on excluded observables $\mathbf{P}$:

$$\mathbb{E}\left[\varepsilon'_{it} \mid j\right] = \mathbb{E}\left[\varepsilon'_{it}\right]$$

Under this assumption, $B = \frac{Cov\left(\boldsymbol{\rho}\mathbf{P}_{it}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)}$. As in (5), we estimate the coefficient vector $\boldsymbol{\rho}$ using an OLS regression of $A_{it}$ on $\mathbf{P}_{it}$ with teacher fixed effects:

$$A_{it} = \alpha_j + \boldsymbol{\rho}\mathbf{P}_{it}. \tag{13}$$

This leads to the feasible estimator

$$B_p = \frac{Cov\left(A^p_{it}, \widehat{\mu}_{jt}\right)}{Var\left(\widehat{\mu}_{jt}\right)}, \tag{14}$$

where $A^p_{it} = \widehat{\boldsymbol{\rho}}\mathbf{P}_{it}$ is estimated using (13). Equation (14) shows that forecast bias $B_p$ can be estimated from an OLS regression of predicted scores $A^p_{it}$ on VA estimates under Assumption 2.

### *B. Estimates of Forecast Bias*

We estimate forecast bias using two variables that are excluded from standard VA models: parent characteristics and lagged test score gains.

*Parent Characteristics.* We define a vector of parent characteristics $\mathbf{P}^*_{it}$ that consists 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 marital status interacted with a quartic in parent household income.[26] We construct residual parent characteristics $\mathbf{P}_{it}$ by regressing each element of $\mathbf{P}^*_{it}$ on the baseline control vector $\mathbf{X}_{it}$ and teacher fixed effects, as in (5). We then regress $A_{it}$ on $\mathbf{P}_{it}$, again including teacher fixed effects,

[26] We code the parent characteristics as 0 for the 12.3% of students whom we are unable to match to a parent either because we could not match the student to the tax data (10.1%) or because we could not find a parent for a matched student (2.2%). We include indicators for missing parent data in both cases. We also code mother's age at child's birth as 0 for observations where we match parents but do not have valid data on the parent's age, and include an indicator for such cases.

and calculate predicted values $A_{it}^{p} = \widehat{\boldsymbol{\rho}}\mathbf{P}_{it}$. We fit separate models for each subject and school level (elementary and middle) as above when constructing the residuals $\mathbf{P}_{it}$ and predicted test scores $A_{it}^{p}$.

In Column 2 of Table 3, we regress $A_{it}^{p}$ on $\widehat{\mu}_{jt}$, including subject-by-school-level fixed effects as above. The coefficient in this regression is $B_p = 0.002$, i.e. the degree of forecast bias due to selection on parent characteristics is 0.2%. The upper bound on the 95% confidence interval for $B_p$ is 0.25%. Figure 2b presents a non-parametric analog of this linear regression. It plots $A_{it}^{p}$ vs. teacher VA $\widehat{\mu}_{jt}$ using a binned scatter on the same scale as in Figure 2a. The relationship between predicted scores and teacher VA is nearly flat throughout the distribution.

Another intuitive way to assess the degree of selection on parent characteristics – which corresponds to familiar methods of assessing omitted variable bias – is to control for $\mathbf{P}_{it}^{*}$ when estimating the impact of $\widehat{\mu}_{jt}$ on test scores, as in Kane and Staiger (2008, Table 6). To implement this approach using within-teacher variation to construct residuals, we first regress raw test scores $A_{it}^{*}$ on the baseline control vector used in the VA model $\mathbf{X}_{it}$, parent characteristics $\mathbf{P}_{it}^{*}$, and teacher fixed effects, as in (5). Again, we fit a separate model for each subject and school level (elementary and middle). We then regress the residuals from this regression (adding back the teacher fixed effects) on $\widehat{\mu}_{jt}$, including subject-by-school-level fixed effects as in Column 1. Column 3 of Table 3 shows that the coefficient on VA falls to 0.996 after controlling for parent characteristics. The difference between the point estimates in Columns 1 and 3 is 0.002. This difference coincides exactly with our estimate of $B_p$ in Column 2 because $A_{it}^{p}$ is simply the difference between test score residuals with and without controlling for parent characteristics.

The magnitude of forecast bias due to selection on parent characteristics is very small for two reasons. First, a large fraction of the variation in test scores that project onto parent characteristics is captured by lagged test scores and other controls in the school district data. The standard deviation of class-average predicted score residuals based on parent characteristics is 0.014. Intuitively, students from high income families have higher test scores not just in the current grade but in previous grades as well, and thus previous scores capture a large portion of the variation in family income. However, because lagged test scores are noisy measures of latent ability, parent characteristics still have significant predictive power for test scores even conditional on $\mathbf{X}_{it}$. The F-statistic on the parent characteristics in the regression of test score residuals $A_{it}$ on $\mathbf{P}_{it}$ is 84, when run at the classroom level (which is the variation relevant for detecting class-level sorting to teachers). This leads to the second reason that forecast bias is so small: the remaining variation in parent characteristics after conditioning on $\mathbf{X}_{it}$ is essentially unrelated to teacher VA. The correlation between $A_{it}^{p}$ and $\widehat{\mu}_{jt}$ is 0.014. If this correlation were 1 instead, the bias due to sorting on parent characteristics would be $B_p = 0.149$. This shows that there is substantial scope to detect sorting of students to teachers based on parent characteristics even conditional on the school district controls. In practice, such sorting turns out to be quite small in magnitude, leading to minimal forecast bias.

*Prior Test Scores*. Another natural set of variables to evaluate bias is prior test scores

(Rothstein 2010). Value-added models typically control for $A_{i,t-1}$, but one can evaluate sorting on $A_{i,t-2}$ (or, equivalently, on lagged gains, $A_{i,t-1} - A_{i,t-2}$). The question here is effectively whether controlling for additional lags substantially affects VA estimates once one controls for $A_{i,t-1}$. For this analysis, we restrict attention to the subsample of students with data on both lagged and twice-lagged scores, essentially dropping 4th grade from our sample. We re-estimate VA $\widehat{\mu}_{jt}$ on this sample to obtain VA estimates on exactly the sample used to evaluate bias.

We assess forecast bias due to sorting on lagged score gains using the same approach as with parent characteristics. Column 4 of Table 3 replicates Column 2, using predicted score residuals based on $A_{i,t-2}$, which we denote by $A^l_{it}$, as the dependent variable. The coefficient on $\widehat{\mu}_{jt}$ is 0.022, with a standard error of 0.002. The upper bound on the 95% confidence interval for forecast bias due to twice lagged scores is 0.026. Figure 2c plots predicted scores based on twice-lagged scores ($A^l_{it}$) against teacher VA, following the same methodology used to construct Figure 2b. Consistent with the regression estimates, there is a slight upward-sloping relationship between VA and $A^l_{it}$ that is small relative to the relationship between VA and test score residuals in year $t$.

Forecast bias due to omitting $A_{i,t-2}$ is small for the same two reasons as with parent characteristics. The baseline control vector $\mathbf{X}_{it}$ captures much of the variation in $A_{i,t-2}$: the variation in class-average predicted score residuals based on $A_{i,t-2}$ is 0.048. The remaining variation in $A_{i,t-2}$ is not strongly related to teacher VA: the correlation between $A_{i,t-2}$ and $\widehat{\mu}_{jt}$ is 0.037. If this correlation were 1, forecast bias would be 0.601, again implying that sorting on lagged gains is quite minimal relative to what it could be.

We conclude that selection on two important predictors of test scores excluded from standard VA models – parent characteristics and lagged test score gains – generates negligible forecast bias in our baseline VA estimates.

### *C. Unconditional Sorting*

To obtain further insight into the sorting process, we analyze unconditional correlations (without controlling for school-district observables $\mathbf{X}_{it}$) between teacher VA and excluded observables such as parent income in Appendix D. We find that sorting of students to teachers based on parent characteristics does not generate significant forecast bias in VA estimates for two reasons.

First, although higher socio-economic status children have higher VA teachers on average, the magnitude of such sorting is quite small. For example, a $10,000 (0.3 SD) increase in parent income raises teacher VA by 0.00084 (Appendix Table 2). One explanation for why even unconditional sorting of students to teacher VA based on family background is small is that 85% of the variation in teacher VA is within rather than between schools. The fact that parents sort primarily by choosing schools rather than teachers within schools limits the scope for sorting on unobservables.

Second, and more importantly, controlling for a student's lagged test score entirely eliminates the correlation between teacher VA and parent income. This shows that the unconditional relationship between parent income and teacher VA runs through lagged test scores and explains why VA estimates that control for lagged scores are unbiased

despite sorting on parent income.

An auxiliary implication of these results is that differences in teacher quality are responsible for only a small share of the gap in achievement by family income. In the cross-section, a $10,000 increase in parental income is associated with a 0.065 SD increase in 8th grade test scores (on average across math and English). Based on estimates of the persistence of teachers' impacts and the correlation between teacher VA and parent income, we estimate that only 4% of the cross-sectional correlation between test scores and parent income in 8th grade can be attributed to differences in teacher VA from grades K-8 (see Appendix D). This finding is consistent with evidence on the emergence of achievement gaps at very young ages (Fryer and Levitt 2004, Heckman et al. 2006), which also suggests that existing achievement gaps are largely driven by factors other than teacher quality. However, it is important to note that good teachers in primary school can close a significant portion of the achievement gap created by these other factors. If teacher quality for low income students were improved by 0.1 SD in all grades from K-8, 8th grade scores would rise by 0.34 SD, enough to offset more than a $50,000 difference in family income.

## V. Estimating Bias Using Teacher Switching Quasi-Experiments

The evidence in Section IV does not rule out the possibility that students are sorted to teachers based on unobservable characteristics orthogonal to parent characteristics and lagged score gains. The ideal method to assess bias due to unobservables would be to estimate the relationship between test scores and VA estimates (from pre-existing observational data) in an experiment where students are randomly assigned to teachers. In this section, we develop a quasi-experimental analog to this experiment that exploits naturally occurring teacher turnover to estimate forecast bias. Our approach yields more precise estimates of the degree of bias than experimental studies (Kane and Staiger 2008, Kane et al. 2013) at the cost of stronger identification assumptions, which we describe and evaluate in detail below.

### *A. Methodology*

Adjacent cohorts of students within a school are frequently exposed to different teachers. In our core sample, 30.1% of teachers switch to a different grade within the same school the following year, 6.1% of teachers switch to a different school within the same district, and another 5.8% switch out of the district entirely.[27]

To understand how we use such turnover to estimate forecast bias, 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 in math is 0.3 higher. Assume that the distribution of unobserved determinants of scores $\varepsilon_{it}$ does not change between 1994 and 1995. If forecast bias $B = 0$, this change in teaching staff should raise average 4th grade math scores in the school by $0.3/3 = 0.1$. More generally, we can estimate $B$ by

[27] Although teachers switch grades frequently, only 4.5% of students have the same teacher for two consecutive years.

comparing the change in mean scores across cohorts to the change in mean VA driven by teacher turnover provided that student quality is stable over time.

To formalize this idea, let $\widehat{\mu}_{jt}^{-\{t,t-1\}}$ denote the VA estimate for teacher $j$ in year $t$ constructed as in Section III using data from all years except $t-1$ and $t$. Similarly, let $\widehat{\mu}_{j,t-1}^{-\{t,t-1\}}$ denote the VA estimate for teacher $j$ in year $t-1$ based on data from all years except $t-1$ and $t$.[28] Let $Q_{sgt}$ denote the (student-weighted) mean of $\widehat{\mu}_{jt}^{-\{t,t-1\}}$ across teachers in school $s$ in grade $g$. We define the change in mean teacher value-added from year $t-1$ to year $t$ in grade $g$ in school $s$ as $\Delta Q_{sgt} = Q_{sgt} - Q_{sg,t-1}$. By leaving out both years $t$ and $t-1$ when estimating VA, we ensure that the variation in $\Delta Q_{sgt}$ is driven by changes in the teaching staff rather than changes in VA estimates.[29] Leaving out two years eliminates the correlation between changes in mean test scores across cohorts $t$ and $t-1$ and estimation error in $\Delta Q_{sgt}$.[30]

Let $A_{sgt}$ denote the mean value of $A_{it}$ for students in school $s$ in grade $g$ in year $t$ and define the change in mean residual scores as $\Delta A_{sgt} = A_{sgt} - A_{sg,t-1}$. We estimate the degree of forecast bias $B$ by regressing changes in mean test scores across cohorts on changes in mean teacher VA:

$$\Delta A_{sgt} = a + b\Delta Q_{sgt} + \Delta \chi_{sgt} \tag{15}$$

The coefficient $b$ in (15) identifies the degree of forecast bias as defined in (10) under the following identification assumption.

**Assumption 3 [Teacher Switching as a Quasi-Experiment]** Changes in teacher VA across cohorts within a school-grade are orthogonal to changes in other determinants of student scores:

$$Cov\left(\Delta Q_{sgt}, \Delta \chi_{sgt}\right) = 0. \tag{16}$$

Under Assumption 3, the regression coefficient in (15) measures the degree of forecast bias: $b = \lambda = 1 - B(\widehat{\mu}_{jt}^{-\{t,t-1\}})$. We present a simple derivation of this result in Appendix E. Intuitively, (15) is a quasi-experimental analog of the regression in (10) used to define forecast bias, differencing and averaging across cohorts. Importantly, because we analyze student outcomes at the school-grade-cohort level in (15), we do not exploit information on classroom assignment, thus overcoming the non-random assignment of students to classrooms within each school-grade-cohort.

Assumption 3 could potentially be violated by endogenous student or teacher sorting to schools over time. 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

[28] As above, we estimate VA separately for each subject (math and English) and school-level (elementary and middle), and hence only identify forecast bias from teacher switches within subject by school-level cells.

[29] Part of the variation in $\Delta \bar{\mu}_{sgt}$ comes from drift. Even for a given teacher, predicted VA will change because our forecast of VA varies across years. Because the degree of drift is small across a single year, drift accounts for 5.5% of the variance in $\Delta \bar{\mu}_{sgt}$. As a result, isolating the variation due purely to teacher switching using an instrumental variables specification yields very similar results (not reported).

[30] Formally, not using a two-year leave out would immediately violate Assumption 3 below, because unobserved determinants of scores ($\varepsilon_{sgt}$ and $\varepsilon_{sg,t-1}$) would appear directly in $\Delta Q_{sgt}$.

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 high-frequency changes we analyze are likely driven by idiosyncratic shocks such as changes in staffing needs, maternity leaves, or the relocation of spouses. Hence, we believe that (16) is a plausible assumption and we present evidence supporting its validity below.

The estimate of $\lambda$ obtained from (15) is analogous to a local average treatment effect (LATE) because it applies to the teacher switches that occur in our data rather than all potential teacher switches in the population. For example, if teachers of honors math classes never teach regular math classes, we would be unable to determine if VA estimates for honors teachers are biased relative to those of regular teachers. This is unlikely to be a serious limitation in practice because we observe a wide variety of staff changes in the district we study (including switches from honors to regular classes), as is typical in large school districts (Cook and Mansfield 2013). Moreover, as long as the policies under consideration induce changes similar to those that occur already – for example, policies do not switch teachers from one type of course to another if such changes have never occurred in the past – this limitation does not affect the policy relevance of our estimates.

If observable characteristics $\mathbf{X}_{it}$ are also orthogonal to changes in teacher quality across cohorts (i.e., satisfy Assumption 3), we can implement (15) simply by regressing the change in raw test scores $\Delta A^*_{sgt}$ on $\Delta Q_{sgt}$. We therefore begin by analyzing changes in raw test scores $\Delta A^*_{sgt}$ across cohorts and then confirm that changes in control variables across cohorts are uncorrelated with $\Delta Q_{sgt}$. Because we do not need controls, in this section we use the core sample of 10.7 million observations described in Section II.C rather than the subset of 7.6 million observations that have lagged scores and the other controls needed to estimate the VA model.

Our research design is related to recent work analyzing the impacts of teacher turnover on student achievement, but is the first to use turnover to validate VA models. Rivkin, Hanushek, and Kain (2005) identify the variance of teacher effects from differences in variances of test score gains across schools with low vs. high teacher turnover. We directly identify the causal impacts of teachers from first moments – the relationship between changes in mean scores across cohorts and mean teacher value-added – rather than second moments. Jackson and Bruegmann (2009) analyze whether the VA of teachers who enter or exit affects the test scores of *other* teachers' students in their school-grade cell, but do not compare changes in mean test scores by cohort to the predictions of VA models.[31]

[31]The peer effects documented by Jackson and Bruegmann could affect our estimates of forecast bias using the switcher design. However, peer learning effects are likely to be smaller with teacher exits than entry, provided that knowledge does not deteriorate very rapidly. We find that teacher entry and exit yield broadly similar results, suggesting that spillovers across teachers are not a first-order source of bias for our technique.

### B. Event Studies

To illustrate our research design, we begin with event studies of test scores around the entry and exit of high and low VA teachers (Figure 3). Let year 0 denote the school year that a teacher enters or exits a school-grade-subject cell and define all other school years relative to that year (e.g., if the teacher enters in 1995, year 1992 is -3 and year 1997 is +2). We define an entry event as the arrival of a teacher who did not teach in that school-grade-subject cell for the three preceding years; analogously, we define an exit event as the departure of a teacher who does not return to the same school-grade-subject cell for at least three years. To obtain a balanced sample, we analyze events for which we have data on average test scores at the school-grade-subject level for at least three years before and after the event.[32]

We define a teacher as "high VA" if her estimated VA in her year of entry or exit is in the top 5% of the distribution of all entrants or leavers in her subject. We define "low VA" teachers analogously as those who have estimated VA in the bottom 5% of the distribution of those who enter or exit in their subjects.[33] We estimate VA for each teacher in the year of entry or exit using data *outside* the six-year window used for the event studies.[34] Because VA is measured with error, some teachers who are classified as "high VA" are those who had very good students by chance. Thus, if we were to define a "high VA" teacher as one whose students scored highly within the event window, we would spuriously find a relationship between test scores and the entry of such a teacher even if she had no causal impact on student performance.

Figure 3a plots the impact of the entry of a high-VA teacher on mean test scores. The solid series plots school-grade-subject-year means of test scores in the three years before and after a high-VA teacher enters the school-grade-subject cell. The dashed line in Figure 3a plots test scores in the previous school year (i.e., the previous grade) for the same cohorts of students. To eliminate any secular trends, we remove year effects by regressing mean test scores on year dummies and plotting the residuals. We normalize both current and previous scores to 0 in the first year of the event study to facilitate interpretation of the scale. We do not condition on any other covariates in this figure: each point simply shows average test scores for different cohorts of students within a school-grade-subject cell adjusted for year effects.

When a high-VA teacher arrives, end-of-year test scores in the subject and grade taught by that teacher rise immediately. But test scores in the prior grade remain stable, as one would expect: the entry of a high-VA teacher in grade 5 should have no impact on the same cohort's 4th grade test scores. The stability of prior test scores supports the identification assumption in (16) that school quality and student attributes are not

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

[33] In cases where multiple teachers enter or exit at the same time, we use the teachers' mean VA to decide whether the event falls in the top or bottom 5% of the relevant VA distribution.

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

changing sharply around the entry of a high-VA teacher. We also find that class size does not change significantly around the entry and exit events we study.

The mean test score in the grade in which the high VA teacher enters rises by 0.042 SD from year -1 to 0, while the mean lagged test score changes by 0.008. Hence, the entry of a high VA teacher increases mean test score gains by 0.035. The null hypothesis that this change is 0 is rejected with $p < 0.001$, with standard errors clustered by school-cohort as above. More importantly, the magnitude of the increase in mean test score gains is very similar to the change in mean teacher VA in the school-grade-subject cell, which is 0.042.[35] The hypothesis that the observed impact on mean score gains equals the increase in mean VA is not rejected ($p = 0.34$), consistent with the hypothesis that the VA estimates are forecast unbiased.

The remaining panels of Figure 3 repeat the event study in Panel A for other types of arrivals and departures. Figure 3b examines current and lagged test scores around the departure of a high-VA teacher. In this figure, we normalize both current and lagged mean scores to 0 in the final year of the event study to facilitate interpretation. There is a smooth negative trend in both current and lagged scores, suggesting that high-VA teachers leave schools with declining scores. However, scores in the grade taught by the teacher who leaves drop sharply relative to prior scores in the event year, showing that the departure of the high VA teacher lowers the achievement of subsequent cohorts of students. Figures 3c and 3d analyze the arrival and departure of low VA teachers. Test scores in the grade taught by the teacher fall relative to prior scores when low VA teachers enter a school-grade cell and rise when low VA teachers leave.

In every case, the change in test score gains is significantly different from 0 with $p < 0.01$ but is not significantly different from what one would forecast based on the change in mean teacher VA. Together, these event studies provide direct evidence that deselecting low VA teachers and retaining high-VA teachers improves the academic achievement of students.

### *C. Estimates of Forecast Bias*

The event studies focus on the tails of the teacher VA distribution and thus exploit only a small fraction of the variation arising from teacher turnover in the data. We now exploit all the variation in teacher VA across cohorts due to teaching staff changes to obtain a broader and more precise estimate of forecast bias. To do so, we first construct a leave-two-year-out estimate of VA $\widehat{\mu}_{jt}^{-\{t,t-1\}}$ for each teacher using data for all years except $t-1$ and $t$. We then calculate the change in mean teacher VA for each school-grade-subject-year cell $\Delta Q_{sgt}$ as described in Section V.A.

Figure 4a presents a binned scatter plot of the changes in mean raw test scores across cohorts $\Delta A^*_{sgt}$ against changes in mean teacher value-added $\Delta Q_{sgt}$, weighting by the number of students in each cell. We include year fixed effects so that the estimate is identified purely from differential changes in teacher VA across school-grade-subject

[35] When computing this change in mean VA, we weight teachers by the number of students they teach. For teachers who do not have any VA measures from classrooms outside the leave-out window, we impute VA as 0 (the sample mean). We discuss the robustness of our results to this imputation below.

cells over time and restrict the sample to classrooms with non-missing teacher VA estimates. Changes in the quality of the teaching staff strongly predict changes in test scores across cohorts in a school-grade-subject cell. The estimated coefficient on $\Delta Q_{sgt}$ is 0.974, with a standard error of 0.033 (Table 4, Column 1). The implied forecast bias of 2.6% is not statistically distinguishable from 0 and the upper bound of the 95% confidence interval is 9.1%.

The conclusion that VA measures exhibit little or no forecast bias rests on the validity of the identification assumption in (16). One natural concern is that improvements in teacher quality may be correlated with other improvements in a school that also increase test scores and thus lead us to underestimate forecast bias. To address this concern, Column 2 of Table 4 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 $\Delta Q_{sgt}$ remains virtually unchanged.

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. Including these additional controls for trends has little impact on our estimate of forecast bias.

### *D. Placebo Tests*

We further evaluate (16) using a series of placebo tests. Column 4 of Table 4 replicates the specification in Column 2, replacing the change in actual scores with the change in predicted scores based on parent characteristics. We predict scores using an OLS regression of $A^*_{it}$ on the five parent characteristics $\mathbf{P}^*_{it}$ used in Section IV, with no other control variables. The estimated effect on predicted test scores is not significantly different from 0, and the upper bound on the 95% confidence interval is 0.013. Figure 4b presents a binned scatter plot corresponding to this regression. There is little relationship between changes in mean teacher VA and mean predicted scores throughout the distribution, supporting the assumption that changes in the quality of the teaching staff are uncorrelated with changes in student characteristics.

Next, we use contemporaneous test scores in the *other* subject (math or English) to evaluate changes in student quality. In this placebo test, it is important to distinguish between elementary and middle schools. In middle school, students have different teachers for math and English. Therefore, if (16) holds, we would expect changes in mean math teacher VA to have little impact on English test scores, holding fixed the quality of English teachers.[36] We test this hypothesis in Column 5 of Table 4 by regressing changes in mean scores on $\Delta Q_{sgt}$ in both the same subject and the other subject. The estimated effect on test scores in the other subject is 0.038 and is not statistically distinguishable from 0 ($p = 0.64$).

[36] If math teacher quality directly affects English scores (or vice versa), one will obtain a positive coefficient in this regression even if (16) holds. Hence, this test is a sufficient but not necessary condition for (16).

Figure 5a presents a non-parametric analog of this regression by plotting changes in mean test scores across cohorts vs. changes in mean teacher VA in the other subject. To partial out the effect of changes in mean teacher VA in the same subject, we regress both the x and y variables on changes in mean teacher VA in the own subject and compute residuals. We then bin the x-axis residuals into 20 equal-sized groups (vingtiles) and plot the mean of the x and y residuals within each bin. Consistent with the regression estimate, there is no relationship between changes in teacher VA and scores in the other subject throughout the distribution. We view this result as strong evidence supporting (16), as it is unlikely that changes in student unobservables $\varepsilon_{sgt}$ would have no impact on scores in the other subject.

In elementary school, students have one teacher for both math and English. Because elementary school teachers' math and English VA are highly correlated ($\rho = 0.6$) and because VA is measured with error, changes in teacher VA in one subject have signal content for teaching quality in the other subject. Therefore, in elementary school, we should expect mean teacher VA to have non-zero effects on scores in the other subject. Figure 5b replicates Figure 5a for elementary school and shows that this is indeed the case. A 1 unit improvement in mean teacher VA raises scores in the other subject by 0.237 (Column 6 of Table 4).

Together, these tests imply that any violation of (16) would have to be driven by unobserved determinants of test scores $\varepsilon_{sgt}$ that (1) are uncorrelated with parent characteristics, (2) are unrelated to prior test scores and contemporaneous test scores in the other subject, and (3) change differentially across grades within schools at an annual frequency. We believe that such sorting on unobservables is implausible given the information available to teachers and students and the constraints they face in switching across schools at high frequencies.

### *E. Additional Robustness Checks*

In Table 5, we further assess the robustness of our estimates to alternative specifications and sample selection criteria. Our preceding specifications pool all sources of variation in teacher quality, including switches across grades within a school as well as entry into and exit from schools. In Column 1 of Table 5, we isolate the variation in teacher quality due to departures from schools, which are perhaps least likely to be correlated with high-frequency fluctuations in student quality across cohorts. We instrument for the change in mean teacher VA $\Delta Q_{sgt}$ with the fraction of students in the prior cohort taught by teachers who leave the school, multiplied by the mean VA among school-leavers (with the instrument defined as 0 for cells with no school-leavers).[37] We include year fixed effects in this specification, as in Column 1 of Table 4. This specification yields a coefficient on $\Delta Q_{sgt}$ of 1.045, similar to our baseline estimate of 0.974.

The preceding estimates all use the subsample of students for whom we have estimates

[37] For example, if the previous cohort had four equally sized classrooms, with one taught by a school-leaver with $\widehat{\mu}_{jt}^{-\{t,t-1\}} = 0.2$, the value of our instrument would be $0.25 * 0.2 = 0.05$. This is the expected loss in mean VA across cohorts due to school-leavers. The first stage coefficient on this variable is -0.97 (t-statistic $= -53$).

of teacher VA. While this corresponds directly to the sample used to estimate VA, restricting the sample to classrooms with non-missing VA estimates could lead to violations of the identification assumption in (16) because we do not use the entire school-grade-subject cohort for identification.[38] To evaluate how missing data on teacher VA affects our results, in Column 2 of Table 5, we replicate the baseline specification (Column 1 of Table 4) using all students in the core sample. We impute teacher VA as the sample mean (0) for the 16% of observations for whom we have no leave-two-year-out VA estimate, either because we have no teacher information or because the teacher taught for only those two years in the district. The coefficient on $\Delta Q_{sgt}$ with this imputation procedure in the full sample is 0.877. The estimate falls because the imputation of teacher VA generates measurement error in average teaching quality, leading to attenuation bias. For example, if a highly effective teacher enters the district for only a single year, so that we are not able to calculate VA from other years of data, our imputation procedure will treat this teacher as being average, leading to measurement error in mean VA in the school-grade-subject cell.[39]

To assess whether measurement error due to imputation is responsible for the smaller coefficient, we restrict the sample to school-grade-subject-year cells in which the fraction of observations with imputed teacher VA estimates is less than 25% in both the current and preceding year. Unlike the sample restrictions imposed in Table 4, restrictions at the school-grade-subject level cannot generate violations of (16) due to selection bias because we exclude entire cohorts rather than individual classrooms. In this subsample, the fraction of observations with non-missing teacher VA information is 94%. Accordingly, the coefficient on $\Delta Q_{sgt}$ rises to 0.952, as shown in Column 3 of Table 5. In Column 4, we further restrict the sample to school-grade-subject-year cells with no missing teacher VA observations in the current and preceding year. In this sample, we obtain an estimate on $\Delta Q_{sgt}$ of 0.990, implying forecast bias of 1.0%. Hence, missing data on teacher VA does not appear to affect our conclusions significantly.

## VI. Comparing Bias Across VA Models

Our analysis thus far has evaluated bias in a value-added specification that includes a very rich set of controls. Which of these controls are most important in obtaining unbiased estimates of VA? To answer this question, we report estimates of bias for various VA models in Table 6. Each row of the table considers a different VA specification. Column 1 reports correlations between the VA estimates obtained from each model and the baseline estimates. Column 2 reports estimates of forecast bias using our preferred quasi-experimental specification (Column 1 of Table 4).

[38] To take an extreme example, suppose teacher information is reported for only one of 5 classrooms in a given school-grade-subject cell. In this case, comparisons of mean scores across students in two cohorts with non-missing teacher information is equivalent to comparing mean scores across a single classroom in two different years. This could violate the identifying assumption of our quasi-experimental design if assignment to classrooms is non-random.

[39] We measure $\Delta Q_{sgt}$ with error because we code VA as 0 for teachers with missing VA before we define $\Delta Q_{sgt}$. If we were to instead set $\Delta Q_{sgt}$ to 0 in all cells where VA is missing for any teacher, we would mechanically obtain the same coefficient as in Column 1 of Table 4.

The first row of the table replicates our baseline VA model as a reference. For comparability, we estimate all the remaining models on the sample used in this row.

In row 2, we follow the traditional approach used in prior work of constructing student test-score residuals by regressing raw scores $A^*_{it}$ on controls $\mathbf{X}_{it}$ in a specification without teacher fixed-effects, unlike in (5). As discussed in Section I.B, this method exploits variation both within and across teachers to identify the coefficients on the control vector and thus can understate teacher effects by over-attributing test score growth to covariates if there is sorting. In practice, this turns out to be a minor concern: VA estimates constructed from test score residuals that are identified using the traditional approach have a correlation of 0.979 with our baseline VA estimates. Correspondingly, they exhibit forecast bias of 2.2%, very similar to our baseline model. This is because unconditional sorting is relatively minimal in practice, as shown in Section IV.C, and thus most of the variation in the controls $\mathbf{X}_{it}$ is within rather than between teachers.

In row 3, we replicate our baseline specification adding indicators for teacher experience to the control vector $\mathbf{X}_{it}$. We include indicators for years of experience from 0 to 5, with the omitted group being teachers with 6 or more years of experience. The resulting VA estimates have a correlation of 0.989 with our baseline estimates because teacher experience has small effects on student achievement (Kane, Rockoff, and Staiger 2008).

The remaining rows of the table strip out elements of the baseline control vector to determine which controls matter most for forecast bias. In row 4, we include only year fixed effects and the controls based on prior-year test scores: cubic polynomials in student, classroom, and school-grade math and English scores interacted with grade level. These VA estimates remain highly correlated with the baseline estimates and continue to exhibit small and statistically insignificant forecast bias of 3.8%.

In row 5, we further reduce the controls to include only the cubics in student-level prior math and English test scores (without grade interactions) along with grade and year fixed effects. Forecast bias is 4.8%, somewhat higher than with our much richer baseline control vector. Nevertheless, even with this very parsimonious control vector, we do not reject the hypothesis that VA estimates are forecast unbiased. In row 6, we replicate row 5, dropping the controls for test scores in the other subject. Forecast bias rises to 10.2% in this specification, showing that prior test scores in the other subject are useful in accounting for sorting.

In row 7, we control for all variables in the baseline model *except* those based on prior test scores. Here, the degree of forecast bias jumps to 45.4%. Finally, row 8 estimates VA without any controls except year and grade fixed effects, i.e., using raw mean test scores by teacher. These VA estimates are very poorly correlated with the other VA measures and are biased by nearly 66%.

The lesson of this analysis is that controlling for prior student-level test scores is the key to obtaining unbiased VA estimates. One potential explanation for this result is that classroom assignments in large schools are made primarily on the basis of prior-year test performance and its correlates.

## VII. Relationship to Prior Work

Prior research on bias in value-added models has reached conflicting conclusions (e.g., Kane and Staiger 2008, Rothstein 2010, Koedel and Betts 2011, Kinsler 2012, Goldhaber and Chaplin 2012, Kane et al. 2013). Our findings help reconcile these results and show that there is actually consensus in the literature on the central empirical issues.

Rothstein (2009, 2010) initiated the recent body of research on bias in VA models by reporting two results. First, there is significant grouping of students into classrooms based on twice-lagged scores (lagged gains), even conditional on once-lagged scores (Rothstein 2010, Table 4). Second, this grouping on lagged gains generates minimal bias in VA estimates: controlling for twice-lagged scores does not have a significant effect on VA estimates (Rothstein 2010, Table 6).[40] We replicate these results in our data (see Table 3), as do Kane and Staiger (2008, Table 6). Therefore, the literature is in agreement that VA measures do not suffer from substantial bias due to selection on lagged score gains.

Rothstein (2010) emphasizes that his findings raise concerns about the *potential* for bias due to selection on unobservable student characteristics.[41] To address this concern, Kane and Staiger (2008) and Kane et al. (2013) report estimates of bias from experiments in which students were randomly assigned to teachers. Their point estimates suggest that selection on unobservables is small, but their 95% confidence intervals are consistent with a large amount of bias – up to 50% in some cases – because of constraints on sample size and compliance. Our quasi-experimental estimates show that the degree of bias due to selection on unobservables turns out to be negligible with much greater precision in a more representative sample. The recent replication of our quasi-experimental methodology by Kane, Staiger, and Bacher-Hicks (2014) in the Los Angeles Unified School District also reaches the same conclusion. Hence, studies that have directly tested for selection on unobservables are in agreement that VA estimates are not significantly biased by such selection.

In future research, it may be interesting to explore why the grouping of students on lagged test score gains originally documented by Rothstein (2010) does not ultimately lead to significant forecast bias in VA estimates. However, the findings in this paper and the related literature are sufficient to conclude that standard estimates of teacher VA provide unbiased forecasts of teachers' impacts on students' test scores.

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

[41] In personal correspondence, Rothstein notes that his findings are "neither necessary nor sufficient for there to be bias in a VA estimate" and that "if the selection is just on observables, the bias is too small to matter. The worrying scenario is selection on unobservables."

## VIII. Conclusion

The main lesson of this study is that value-added models that control for a student's prior-year test scores provide unbiased forecasts of teachers' causal impacts on student achievement. Because the dispersion in teacher effects is substantial, this result implies that improvements in teacher quality can raise students' test scores significantly.

Although our analysis shows that test-score based VA measures are useful predictors of teacher quality, one can potentially improve such predictions in at least two dimensions. First, incorporating other measures of teacher quality – such as principal evaluations or information on teacher characteristics – may yield more precise forecasts of teachers' impacts on test scores. One could use the quasi-experimental methodology developed here to validate such measures of teacher quality. For example, when a teacher with good evaluations switches into a school, do outcomes improve? Second, it would be valuable to develop richer measures of teacher quality that go beyond the mean test score impacts that we analyzed here. One could develop VA measures at various percentiles of the test score distribution and for specific demographic subgroups. For instance, are some teachers better with boys rather than girls or high achievers rather than low achievers?

In this paper, we established that value-added measures can help us identify which teachers have the greatest ability to raise students' test scores. One cannot conclude from this result that teacher VA is a good measure of teacher "quality," however, as test scores are not the ultimate outcome of interest. Do high VA teachers also improve students' long-term outcomes, or are they simply better at teaching to the test? We turn to this question in the next paper in this volume.

## REFERENCES

**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.

**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.

**Barlevy, Gadi and Derek Neal.** 2012. "Pay for Percentile." *American Economic Review* 102(5): 1805-31.

**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.

**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.

**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.

**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.

**Chetty, Raj, John N. Friedman, and Jonah E. Rockoff.** 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* (forthcoming).

**Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez.** 2014. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States" NBER Working Paper 19843.

**Cook, Jason and Richard Mansfield.** 2013. "Task-Specific Experience and Task-Specific Talent: Decomposing the Productivity of High School Teachers." Cornell University mimeo.

**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.

**Fryer, Roland G. and Steven D. Levitt.** 2004. "Understanding the Black-White Test Score Gap in the First Two Years of School." *The Review of Economics and Statistics* 86(2): 447-464.

**Goldhaber, Dan, and Duncan Chaplin.** 2012. "Assessing the 'Rothstein Falsification Test': Does It Really Show Teacher Value-Added Models Are Biased?" CEDR Working Paper 2011-5. University of Washington, Seattle, WA.

**Goldhaber, Dan and Michael Hansen.** 2010. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review Papers and Proceedings*

100(2): 250-255.

**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.

**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.

**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.

**Heckman, James J., Jora Stixrud and Sergio Urzua.** 2006. "The Effects Of Cognitive and Noncognitive Abilities On Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24(3): 411-482.

**Jackson, C. Kirabo.** 2009. "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation." *Journal of Labor Economics* 27(2): 213-256.

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

**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.

**Kane, Thomas J., and Douglas O. Staiger.** 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper 14607.

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

**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.

**Kane, Thomas J., Douglas O. Staiger, and Andrew Bacher-Hicks.** 2014. "Validating Teacher Effect Estimates using Between School Movers: A Replication and Extension of Chetty et al." Harvard University Working Paper.

**Kinsler, Josh.** 2012. "Assessing Rothstein's Critique of Teacher Value-Added Models." *Quantitative Economics* 3: 333-362

**Koedel, Cory and Julian R. Betts.** 2011. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy* 6(1): 18-42.

**Mansfield, Richard K.** 2013. "Teacher Quality and Student Inequality." Unpublished Working Paper.

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

**Murnane, Richard J.** 1975. *The Impact of School Resources on Learning of Inner City*

*Children*. Cambridge, MA: Ballinger Publishing.

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

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

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

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

**Online Appendix A: Value-Added Estimation Methods**

In this appendix, we provide a step-by-step guide to implementing our method of estimating VA in the presence of drift. In practice, we cannot follow exactly the method described in Section I.B because data availability varies across teachers. For instance, there are different numbers of students per class and teachers have a different number of past and future classes from which to construct value-added in any given year. We calculate value-added in three steps, separately for each subject (math and English) and school level (elementary and middle).

Step 1 [Residualization of Test Scores]: We begin by residualizing student scores $A^*_{it}$ with respect to controls $\mathbf{X}_{it}$ by running an OLS regression with teacher fixed effects of the form

$$A^*_{it} = \alpha_j + \boldsymbol{\beta}\mathbf{X}_{it}$$

and constructing residuals

$$A_{it} = A^*_{it} - \widehat{\boldsymbol{\beta}}\mathbf{X}_{it}.$$

Step 2 [Estimation of Variance Components]: Next, we estimate the individual-level variance of residual test scores, $\sigma^2_\varepsilon = Var(\varepsilon_{it})$, as

$$\hat{\sigma}^2_\varepsilon = MSE * \left(\frac{N-1}{N-K-C+1}\right)$$

where $MSE$ is the variance of the within-classroom deviations of $A_{it}$, $N$ is the total number of students, $C$ is the total number of classrooms, and $K$ is the number of control variables in the $X_{it}$ control vector. The scaling term is required to correct the degrees of freedom for the fact that we have already estimated $K$ parameters to form the residual $A_{it}$. We also estimate $Var(A_{it})$, the total variance of $A_{it}$, again accounting for the prior estimation of $\widehat{\boldsymbol{\beta}}$ when calculating the degrees of freedom correction.

At this point, we collapse the data to the classroom-level by constructing the average residualized score $\bar{A}_{ct}$ for each classroom $c$ and proceed to use class-level means for the remaining steps. In middle school, teachers teach more than one class per year. We handle such cases by collapsing the data to the teacher-year level. We do so by constructing precision-weighted averages of classroom-average scores within a teacher-year. The weight for classroom $c$ in year $t$ is

$$h_{ct} = \frac{1}{\hat{\sigma}^2_\theta + \frac{\hat{\sigma}^2_\varepsilon}{n_{ct}}},$$

where $\hat{\sigma}^2_\theta$ is an estimate of the class-level variance component and $n_{ct}$ denotes the number of students in the classroom. We construct this estimate as $\hat{\sigma}^2_\theta = Var(A_{it}) - \hat{\sigma}^2_\varepsilon - \hat{\sigma}_{A0}$, where $\hat{\sigma}_{A0}$ is our estimate of the within-teacher-year between-class covariance in average scores, reflecting the teacher-level component of the variance. To simplify computation, we follow Kane and Staiger (2008) and randomly sort classrooms within each teacher-

year cell; we then estimate the covariance $\hat{\sigma}_{A0}$ based on the covariance of the test scores of adjacent classrooms in each teacher-year cell, weighting each pair of classrooms by the sum of students taught.

We next estimate the covariances between mean scores across years within teacher, denoted $\hat{\sigma}_{As}$, in both elementary and middle schools. We allow a separate covariance for each possible time lag $s \in \{1, 2, \ldots\}$ denoting the separation between the two years in which the classes were taught. We weight each teacher-year pair by the sum of students taught. We set all covariances for lags greater than 7 to $\hat{\sigma}_{A7}$, the estimated covariance for the 7th lag.

Step 3: [Construction of VA Estimates] In this step, we use the parameter estimates to construct a VA estimate for each teacher $j$ in each year $t$ that she appears in the data. We depart from the method described in Section I.B by using data from all other years – not just years before year $t$ – to increase the precision of our VA estimates.[42] Let $\mathbf{A}_j^{-t}$ denote the vector of teacher-year-mean scores used to predict teacher $j$'s VA in year $t$. Let $N_{jt}$ denote the length of this vector, so that we are using $N_{jt}$ other years to project scores in year $t$. We construct the best linear predictor of teacher quality in year $t$ as

$$\widehat{\mu}_{jt} = \left(\Sigma_{A_{jt}}^{-1}\gamma_{jt}\right)' \mathbf{A}_j^{-t}$$

where $\gamma_{jt}$ is a $N_{jt} \times 1$ vector and $\Sigma_{A_{jt}}$ is a $N_{jt} \times N_{jt}$ matrix. We denote the weights on scores $\mathbf{A}_j^{-t}$ by $\psi_{jt} = \Sigma_{A_{jt}}^{-1}\gamma_{jt}$. If the $m$th and $n$th element of the scores vector $\mathbf{A}_j^{-t}$ are $A_{js}$ and $A_{js'}$, the $m$th element of the diagonal of $\Sigma_{A_{jt}}$ in middle school is

$$\left[\Sigma_{A_{jt}}\right]_{mm} = \hat{\sigma}_{A0} + \frac{1}{\sum_{c\in\{c:j(c)=j\}} h_{cs}},$$

where the denominator of the second term is the sum of precisions for the classes taught by a teacher in year $s$, which is the precision of the teacher-year mean in year $s$. In elementary school, where teachers teach one class per year, we cannot estimate $\hat{\sigma}_{A0}$ but we can estimate $\hat{\sigma}_{A0} + \hat{\sigma}_{\theta}^2 = Var(A_{it}) - \hat{\sigma}_{\varepsilon}^2$. Here, the $m$th element of the diagonal of $\Sigma_{A_{jt}}$ is

$$\left[\Sigma_{A_{jt}}\right]_{mm} = (\hat{\sigma}_{A0} + \hat{\sigma}_{\theta}^2) + \frac{\hat{\sigma}_{\varepsilon}^2}{n_{ct}}.$$

In both elementary and middle school, the $mn$th off-diagonal element of $\Sigma_{A_{jt}}$ is

$$\left[\Sigma_{A_{jt}}\right]_{mn} = \hat{\sigma}_{A,|s-s'|}$$

and the $m$th element of $\gamma_{jt}$ is

$$\left[\gamma_{jt}\right]_{m} = \hat{\sigma}_{A,|t-s|}.$$

[42] Using data from other years increases precision not just by increasing sample size but also because we have more data from nearby years. For example, data from year $t+1$ are more informative for VA in year $t$ than data from year $t-2$ in the presence of drift.

Because the distribution of other years in which data are available varies both across teachers $j$ and across the years $t$ within a teacher, both the matrix $\Sigma_{A_{jt}}$ and the vector $\gamma_{jt}$ will vary across $j$ and $t$. We therefore construct these elements separately for each teacher-year in the data. Note that we can use this algorithm even if data on test scores for teacher $j$'s students are missing in year $t$, since those data are not required to estimate $\widehat{\mu}_{jt}$.

**Online Appendix B: Teacher-Level Bias**

In this appendix, we define an alternative notion of bias in VA estimates, which we term "teacher-level bias," and characterize its relationship to the concept of forecast bias that we focus on in the text.[43] For simplicity, we follow Rothstein (2009) and focus on the case without drift in teacher value-added.[44] In this case, $\mu_{jt} = \mu_j$ in all periods and our estimator for teacher VA simplifies to

$$\widehat{\mu}_{jt} = \bar{A}_j^{-t} \frac{\sigma_\mu^2}{\sigma_\mu^2 + (\sigma_\theta^2 + \sigma_{\bar{\varepsilon}}^2/n)/(t-1)}, \tag{17}$$

as shown in (9). Let $\widehat{\mu}_j^* = \lim_{t\to\infty} \widehat{\mu}_{jt}$ denote the value to which the VA estimate for teacher $j$ converges as the number of classrooms observed approaches infinity. The asymptotic bias in the estimate of teacher $j$'s VA is

$$\omega_j = \widehat{\mu}_j^* - \mu_j. \tag{18}$$

**Definition 2.** Value-added estimates are *unbiased at the teacher-level* if $Var(\omega_j) = 0$.

VA estimates are biased at the teacher-level if they are inconsistent, i.e. if we systematically mispredict a given teacher's performance when estimation error in VA vanishes. Such teacher-level bias is relevant for determining whether a value-added model treats all teachers equitably.

*Forecast vs. Teacher-Level Bias.* To see the connection between teacher-level bias and forecast bias, consider an experiment in which students are randomly assigned to teachers in year $t$. By the definition of forecast bias,

$$\begin{aligned} 1 - B\left(\widehat{\mu}_j^*\right) &= \frac{Cov\left(A_{it}, \widehat{\mu}_j^*\right)}{Var(\widehat{\mu}_j^*)} \\ &= \frac{Var\left(\mu_j\right) + Cov\left(\mu_j, \omega_j\right)}{Var\left(\mu_j\right) + Var\left(\omega_j\right) + 2Cov\left(\mu_j, \omega_j\right)} \end{aligned}$$

where the second step follows because $Cov\left(A_{it} - \mu_j, \widehat{\mu}_j^*\right) = 0$ under random assign-

[43] We thank Jesse Rothstein for drawing our attention to the distinction between teacher-level bias and forecast bias.

[44] Drift complicates the asymptotics because additional information from prior years does not eliminate estimation error in expected VA.

ment in year $t$. Hence,

$$B\left(\widehat{\mu}_j^*\right) = 0 \Leftrightarrow Var(\omega_j) + Cov\left(\mu_j, \omega_j\right) = 0.$$

This identity has two implications. First, if VA estimates are unbiased at the teacher-level, they must also be forecast-unbiased: $Var(\omega_j) = 0 \Rightarrow B(\widehat{\mu}_j^*) = 0$. Second, and more importantly for our application, forecast-unbiased VA estimates can be biased at the teacher level only if $Cov\left(\mu_j, \omega_j\right) = -Var(\omega_j)$. Intuitively, if the teacher-level bias $\omega_j$ is negatively correlated with true value-added, then the covariance of VA estimates with true scores is reduced, but the variance of VA estimates also falls. If the two forces happen to cancel out exactly, $B(\widehat{\mu}_j^*)$ could be 0 even if $Var(\omega_j) > 0$. In this sense, if a pre-specified value-added model produces VA estimates $\widehat{\mu}_{jt}$ that exhibit no forecast bias, the existence of teacher-level bias is a measure-zero (knife-edge) case.

Note that estimating the degree of forecast bias is simpler than teacher-level bias because forecast bias can be directly estimated using finite-sample estimates of $\widehat{\mu}_{jt}$ without any additional inputs. In contrast, estimating teacher-level bias requires accounting for the impacts of estimation error on $\widehat{\mu}_{jt}$ to construct the limit $\widehat{\mu}_j^*$, which is a non-trivial problem, particularly in the presence of drift.

**Online Appendix C: Matching Algorithm**

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

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

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

for each individual from the school files. We keep a potential match if any of these three last names match any of the last names present in the SSA file.

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

Step 3 [Dependent ZIP code]: For each potential match that remains, we find the household that claimed the individual as a dependent (if the individual was claimed at all) in each year. We then match the location of the claiming household, identified by the 5-digit ZIP code, to the home address ZIP code recorded in the school files. We classify potential matches based on the best ZIP code match across all years using the following tiers: exact match, match within 10 (e.g., 02139 and 02146 would qualify as a match), match within 100, and non-match. We retain potential matches only in the highest available tier of ZIP code match quality. For example, suppose there are 5 potential matches for a given individual, and that there are no exact matches on ZIP code, two matches within 10, two matches within 100, and one non-match. We would retain only the two that matched within 10. After this procedure, we isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 2$.

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

Step 5 [Rule In on First Name]: After exhausting other available information, we return to the first name. In step 2 we retained potential matches that either matched on

first name or for which there was no first name available. In this step, we retain only potential matches that match on first name, if such a potential match exists for a given student. We also use information on first name present on 1040 forms filed by potential matches as adults to identify matches at this stage. We then isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 4$.

Step 6 [Fuzzy Date-of Birth]: In previous work (Chetty et al. 2011), we found that 2-3% of individuals had a reported date of birth that was incorrect. In some cases the date was incorrect only by a few days; in others the month or year was off by one, or the transcriber transposed the month and day. To account for this possibility, we take all individuals for whom no eligible matches remained after step 2. Note that if any potential matches remained after step 2, then we would either settle on a unique best match in the steps that follow or find multiple potential matches even after step 5. We then repeat step 1, matching on gender, first four letters of last name, and fuzzy date-of-birth. We define a fuzzy DOB match as one where the absolute value of the difference between the DOB reported in the SSA and school data was in the set $\{1, 2, 3, 4, 5, 9, 10, 18, 27\}$ in days, the set $\{1, 2\}$ in months, or the set $\{1\}$ in years. We then repeat steps 2 through 5 exactly as above to find additional matches. We classify matches found using this fuzzy-DOB algorithm as $MQ = 5.X$, where $X$ is the corresponding $MQ$ from the non-fuzzy DOB algorithm. For instance, if we find a unique fuzzy-DOB match in step 3 using dependent ZIP codes, then $MQ = 5.2$.

The following table shows the distribution of match qualities for all students. We match 88.6% of students and 89.8% of student-subject observations in the analysis sample used to calculate VA. Unmatched students are split roughly evenly among those for whom we found multiple matches and those for whom we found no match.

| Match Quality (MQ) | Frequency | Percent | Cumulative Match Rate |
|---|---|---|---|
| 1 | 650002 | 47.55% | 47.55% |
| 2 | 511363 | 37.41% | 84.95% |
| 3 | 24296 | 1.78% | 86.73% |
| 4 | 10502 | 0.77% | 87.50% |
| 5.1 | 14626 | 1.07% | 88.57% |
| 5.2 | 779 | 0.06% | 88.63% |
| 5.3 | 96 | 0.01% | 88.63% |
| 5.4 | 31 | 0.01% | 88.64% |
| Multiple Matches | 75010 | 5.49% | |
| No Matches | 80346 | 5.88% | |
| Total | 1367051 | | 88.64% |

**Online Appendix D: Unconditional Sorting of Students to Teachers**

In this appendix, we assess the unconditional relationship between teacher VA and student observables to determine whether high VA teachers are systematically assigned to certain types of students (see Section IV.C). We are able to study such unconditional sorting because our method of constructing student test score residuals in (5) only exploits within-teacher variation. Prior studies that estimate VA typically construct test score residuals using both between- and within-teacher variation and thus do not necessarily obtain a global ranking. For example, suppose schools with higher SES students have better teachers. By residualizing test scores with respect to student SES before computing teacher VA, one would attribute the differences in outcomes across these schools to differences in student SES rather than teacher quality. As a result, one only obtains a relative ranking of teachers conditional on student SES and cannot compare teacher quality across students with different characteristics. Using within-teacher variation to estimate the coefficients on the control vector $\mathbf{X}_{it}$ resolves this problem and yields a global ranking of teachers across the school district.

To estimate unconditional sorting of students to teachers based on observable characteristics $\mathbf{X}_{it}$, one would ideally regress teacher VA $\mu_{jt}$ on $\mathbf{X}_{it}$:

$$\mu_{jt} = \alpha + \boldsymbol{\rho}\mathbf{X}_{it} + \eta_{it}. \tag{19}$$

Since true VA is unobserved, we substitute VA estimates $\widehat{\mu}_{jt}$ for $\mu_{jt}$ on the left hand side of (19). This yields an attenuated estimate of $\boldsymbol{\rho}$ because $\widehat{\mu}_{jt}$ is shrunk toward 0 to account for estimation error (see Section I.B). If all teachers taught the same number of classes and had the same number of students, the shrinkage factor would not vary across observations. In this case, we could identify $\rho$ by using $\widehat{\mu}_{jt}$ as the dependent variable in (19) and multiplying the estimate of $\boldsymbol{\rho}$ by $SD(\mu_{jt})/SD(\widehat{\mu}_{jt})$. In the sample for which we observe lagged test scores, the standard deviation of teacher VA estimates is $SD(\mu_{jt})/SD(\widehat{\mu}_{jt}) = 1.56$. We therefore multiply the estimate of $\boldsymbol{\rho}$ obtained from estimating (19) with $\widehat{\mu}_{jt}$ as the dependent variable by 1.56. This simple approach to correcting for the attenuation bias is an approximation because the shrinkage factor does vary across observations. However, our estimates of the magnitudes of unconditional sorting are small and hence further adjusting for the variation in shrinkage factors is unlikely to affect our conclusions.

We report estimates of unconditional sorting in Appendix Table 2. Each column reports estimates of an OLS regression of VA estimates $\widehat{\mu}_{jt}$ on various observables (multiplied by 1.56), with standard errors clustered at the teacher level to account for correlated errors in the assignment process of classrooms to teachers.

We begin in Column 1 by regressing $\widehat{\mu}_{jt}$ on lagged test scores $A^*_{i,t-1}$. Better students are assigned slightly better teachers: students who score 1 unit higher in the previous grade get a teacher whose VA is 0.0122 better on average. The tracking of better students to better teachers magnifies gaps in achievement, although the magnitude of this amplification effect is small relative to other determinants of the variance in student achievement.

Column 2 shows that special education students are assigned teachers with 0.003 lower VA on average. Again, this effect is statistically different from zero, but is quantitatively

small. Relative to other students with similar prior test scores, special education students receive slightly *higher* VA teachers (not reported).

In Column 3, we regress $\widehat{\mu}_{jt}$ on parent income. A \$10,000 (0.3 SD) increase in parent income raises teacher VA by 0.00084, with the null hypothesis of 0 correlation rejected with $p < 0.0001$. Column 4 demonstrates that controlling for a student's lagged test score $A^*_{i,t-1}$ entirely eliminates the correlation between teacher VA and parent income.

Column 5 analyzes the correlation between teacher VA and ethnicity. Mean teacher quality is no different on average across minority (Hispanic or Black) vs. non-minority students.

Finally, Columns 6 and 7 analyze the relationship between teacher value-added and school-level demographics. The relationship between mean parent income in a school and teacher quality remains quite small (Column 6) and there is no relationship between fraction minority and school quality.

Finally, we assess the extent to which differences in teacher quality contribute to the gap in achievement by family income. In the sample used to estimate VA, a \$10,000 increase in parental income is associated with a 0.065 SD increase in 8th grade test scores (averaging across math and English). To calculate how much smaller this gradient would be if teacher VA did not vary with parent income, we must take a stance on how teachers' impacts cumulate over time. In our companion paper, we estimate that 1 unit improvement in teacher VA in a given grade raises achievement by approximately 0.53 units after 1 year, 0.36 after 2 years, and stabilizes at approximately 0.25 after 3 years (Chetty, Friedman, and Rockoff 2014, Appendix Table 10). Under the assumption that teacher effects are additive across years, these estimates of fade-out imply that a 1 unit improvement in teacher quality in all grades K-8 would raise 8th grade test scores by 3.4 units. Using the estimate in Column 3 of Appendix Table 2, it follows that only $\frac{3.4 \times 0.00084}{0.065} = 4\%$ of the income-score gradient can be attributed to differences in teacher quality from grades K-8. However, a sequence of good teachers can close a significant portion of the achievement gap. If teacher quality for low income students were improved by 0.1 units in all grades from K-8, 8th grade scores would rise by 0.34, enough to offset more than a \$50,000 difference in family income.

**Online Appendix E: Quasi-Experimental Estimator of Forecast Bias**

This appendix shows that estimating (15) using OLS identifies the degree of forecast bias under Assumption 3. Recall that we define the degree of forecast bias $B = 1 - \lambda$ based on the best linear predictor of $A_{it}$ in a randomized experiment in year $t$:

$$E^*[A_{it}|1, \widehat{\mu}_{jt}] = \alpha_t + \lambda \widehat{\mu}_{jt}.$$

In the observational data, we can decompose test scores in year $t$ into the effect of teacher VA $E^*[A_{it}|1, \widehat{\mu}_{jt}]$ and student-level errors $\chi_{it}$:

$$A_{it} = \alpha_t + \lambda \widehat{\mu}_{jt} + \chi_{it},$$

where $\chi_{it}$ may be correlated with $\widehat{\mu}_{jt}$ because of non-random assignment. Taking averages over all the students in a school-grade cell and first-differencing gives the quasi-experimental specification in (15):

$$\Delta \overline{A}_{sgt} = \alpha + \lambda \Delta Q_{sgt} + \Delta \chi_{sgt},$$

where $Q_{sgt}$ is the mean of $\widehat{\mu}_{jt}$ in school $s$ in grade $g$ in year $t$. It follows immediately that estimating (15) using OLS yields an unbiased estimate of $\lambda$ under Assumption 3 ($\Delta \chi_{sgt}$ orthogonal to $\Delta Q_{sgt}$).

TABLE 1
Summary Statistics for Sample Used to Estimate Value-Added Model

| Variable | Mean (1) | Std. Dev. (2) | Observations (3) |
|---|---|---|---|
| Student Data: | | | |
| Class size (not student-weighted) | 27.3 | 5.6 | 391,487 |
| Number of subject-school years per student | 5.6 | 3.0 | 1,367,051 |
| Test score (SD) | 0.2 | 0.9 | 7,639,288 |
| Female | 50.8% | | 7,639,288 |
| Age (years) | 11.4 | 1.5 | 7,639,288 |
| Free lunch eligible (1999-2009) | 79.6% | | 5,021,163 |
| Minority (black or hispanic) | 71.6% | | 7,639,288 |
| English language learner | 4.8% | | 7,639,288 |
| Special education | 1.9% | | 7,639,288 |
| Repeating grade | 1.7% | | 7,639,288 |
| Matched to parents in tax data | 87.7% | | 7,639,288 |
| Parent Characteristics: | | | |
| Household income | 40,773 | 34,270 | 6,695,982 |
| Owned a house | 32.9% | | 6,695,982 |
| Contributed to a 401k | 30.9% | | 6,695,982 |
| Married | 42.7% | | 6,695,982 |
| Age at child birth | 29.2 | 8.0 | 6,617,975 |
| Predicted score | 0.17 | 0.26 | 7,639,288 |

*Notes:* All statistics reported are for the sample used in estimating the baseline value-added model, as described in Section III.C. This sample includes only students who have non-missing lagged test scores and other requisite controls to estimate the VA model. Number of observations is number of classrooms in the first row, number of students in the second row, and number of student-subject-year observations in all other rows. Student data are from the administrative records of a large urban school district in the U.S. Parent characteristics are measured between 2005-2007 from 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 is based on standardized scale scores, as described in Section II.A. 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. Conditional on being matched to the tax data, we are unable to link 2.4% of students to their parents; the summary statistics for parents exclude these observations. Parent income is average adjusted gross income during the tax-years 2005-2007. For parents who do not file, household income is defined as zero. Home ownership is defined as reporting mortgage interest payments on a 1040 or 1099 form in any year between 2005-2007. Contributed to a 401(k) is an indicator for ever contributing to a 401(k) between 2005-2007. Marital status is measured by whether the claiming parent files a joint return at any point between 2005-2007. Parent age at child birth is the difference between the age of the mother (or father if single father) and the student. Predicted score is predicted from a regression of scores on parent characteristics using the estimating equation in Section IV.

TABLE 2
Teacher Value-Added Model Parameter Estimates

| Sample: | Elem. School English | Elem. School Math | Middle School English | Middle School Math |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | *Panel A: Autocovariance and Autocorrelation Vectors* | | | |
| Lag 1 | 0.013 | 0.022 | 0.005 | 0.013 |
| | (0.0003) | (0.0003) | (0.0002) | (0.0002) |
| | [0.305] | [0.434] | [0.234] | [0.476] |
| Lag 2 | 0.011 | 0.019 | 0.004 | 0.011 |
| | (0.0003) | (0.0003) | (0.0002) | (0.0003) |
| | [0.267] | [0.382] | [0.186] | [0.396] |
| Lag 3 | 0.009 | 0.017 | 0.003 | 0.009 |
| | (0.0003) | (0.0004) | (0.0003) | (0.0003) |
| | [0.223] | [0.334] | [0.156] | [0.339] |
| Lag 4 | 0.008 | 0.015 | 0.002 | 0.007 |
| | (0.0004) | (0.0004) | (0.0003) | (0.0004) |
| | [0.190] | [0.303] | [0.097] | [0.269] |
| Lag 5 | 0.008 | 0.014 | 0.002 | 0.006 |
| | (0.0004) | (0.0005) | (0.0004) | (0.0005) |
| | [0.187] | [0.281] | [0.096] | [0.217] |
| Lag 6 | 0.007 | 0.013 | 0.002 | 0.006 |
| | (0.0004) | (0.0006) | (0.0004) | (0.0006) |
| | [0.163] | [0.265] | [0.084] | [0.221] |
| Lag 7 | 0.006 | 0.013 | 0.001 | 0.005 |
| | (0.0005) | (0.0006) | (0.0005) | (0.0006) |
| | [0.147] | [0.254] | [0.060] | [0.201] |
| Lag 8 | 0.006 | 0.012 | 0.001 | 0.005 |
| | (0.0006) | (0.0007) | (0.0005) | (0.0007) |
| | [0.147] | [0.241] | [0.030] | [0.210] |
| Lag 9 | 0.007 | 0.013 | 0.001 | 0.004 |
| | (0.0007) | (0.0008) | (0.0006) | (0.0008) |
| | [0.165] | [0.248] | [0.051] | [0.162] |
| Lag 10 | 0.007 | 0.012 | 0.001 | 0.005 |
| | (0.0008) | (0.0010) | (0.0007) | (0.0012) |
| | [0.153] | [0.224] | [0.062] | [0.179] |
| | *Panel B: Within-Year Variance Components* | | | |
| Total SD | 0.537 | 0.517 | 0.534 | 0.499 |
| Individual Level SD | 0.506 | 0.473 | 0.513 | 0.466 |
| Class+Teacher Level SD | 0.117 | 0.166 | 0.146 | 0.178 |
| Class-Level SD | | | 0.108 | 0.116 |
| Teacher SD | | | 0.098 | 0.134 |
| Estimates of Teacher SD: | | | | |
| Lower Bound based on Lag 1 | 0.113 | 0.149 | 0.068 | 0.115 |
| Quadratic Estimate | 0.124 | 0.163 | 0.079 | 0.134 |

*Notes:* Panel A reports the estimated autocovariance, the standard error of that covariance estimate clustered at the teacher level (in parentheses), and the autocorrelation (in brackets) of average test score residuals between classrooms taught by the same teacher. We measure these statistics at time lags ranging from one (i.e. two classrooms taught one year apart) to ten years (i.e., two classrooms taught ten years apart), weighting by the sum of the relevant pair of class sizes. Each covariance is estimated separately for English and math and for elementary and middle school classrooms. Panel B reports the raw standard deviation of test score residuals and decomposes this variation into components driven by idiosyncratic student-level variation, classroom shocks, and teacher-level variation. The variances in rows 2 and 3 of Panel B sum to that in row 1; the variances in rows 4 and 5 sum to that in row 3. In middle school, we estimate the standard deviation of teacher effects as the square root of the covariance of mean score residuals across a random pair of classrooms within the same year. In elementary schools, we cannot separately identify class-level and teacher-level standard deviations because we observe only one classroom per year. We use the square root of the autocovariance across classrooms at a one year lag to estimate a lower bound on the within-year standard deviation for elementary schools. We also report an estimate of the standard deviation by regressing the log of first seven autocovariances in Panel A on the time lag and time lag squared and extrapolating to 0 to estimate the within-year covariance.

TABLE 3
Estimates of Forecast Bias Using Parent Characteristics and Lagged Scores

| Dep. Var.: | Score in Year t | Pred. Score using Parent Chars. | Score in Year t | Pred. Score using Year t-2 Score |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Teacher VA | 0.998 | 0.002 | 0.996 | 0.022 |
| | (0.0057) | (0.0003) | (0.0057) | (0.0019) |
| Parent Chars. Controls | | | X | |
| Observations | 6,942,979 | 6,942,979 | 6,942,979 | 5,096,518 |

*Notes:* Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are run on the sample used to estimate the baseline VA model, restricted to observations with a non-missing leave-out teacher VA estimate. There is one observation for each student-subject-school year in all regressions. Teacher VA is scaled in units of student test score standard deviations and is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections I.B and III. Teacher VA is estimated using the baseline control vector, which includes: a cubic in lagged own- and cross-subject scores, interacted with the student's grade level; student-level characteristics including ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, and limited English; 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 all the student-level characteristics; and grade and year dummies. When prior test scores in the other subject are missing, we set the other subject prior score to 0 and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores. In Columns 1 and 3, the dependent variable is the student's test score in a given year and subject. In Column 2, the dependent variable is the predicted value generated from a regression of test score on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent household income, after residualizing all variables with respect to the baseline control vector. In Column 4, the dependent variable is the predicted value generated in the same way from twice-lagged test scores. See Section IV.B for details of the estimating equation for predicted scores.

TABLE 4
Quasi-Experimental Estimates of Forecast Bias

| Dependent Variable: | Δ Score | Δ Score | Δ Score | Δ Predicted Score | Δ Other Subj. Score | Δ Other Subj. Score |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Changes in Mean Teacher VA across Cohorts | 0.974 (0.033) | 0.957 (0.034) | 0.950 (0.023) | 0.004 (0.005) | 0.038 (0.083) | 0.237 (0.028) |
| Year Fixed Effects | X | | | | X | X |
| School x Year Fixed Effects | | X | X | X | | |
| Lagged Score Controls | | | X | | | |
| Lead and Lag Changes in Teacher VA | | | X | | | |
| Other-Subject Change in Mean Teacher VA | | | | | X | X |
| Grades | 4 to 8 | 4 to 8 | 4 to 8 | 4 to 8 | Middle Sch. | Elem. Sch. |
| Number of School x Grade x Subject x Year Cells | 59,770 | 59,770 | 46,577 | 59,323 | 13,087 | 45,646 |

*Notes:* Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on a dataset containing school-grade-subject-year means from the core sample described in Section II.C, excluding classrooms in which we cannot construct the leave-two-year-out VA estimate described below. All regressions are weighted by the number of students in the school-grade-subject-year cell. We calculate changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows. First, we calculate teacher VA for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean VA across all teachers, weighting by the number of students they teach. Finally, we compute the difference in mean teacher VA (year t minus year t-1) to obtain the independent variable. The dependent variables are defined by calculating the change in the mean of the dependent variable (year t minus year t-1) within a school-grade-subject cell. In Columns 1-3, the dependent variable is the change in mean test scores within subject (English or math). In Column 4, it is the change in the predicted score, constructed based on parental characteristics, as described in Section V.D. In Columns 5 and 6, the dependent variable is the change in the score in the other subject (e.g. math scores for English teachers). Column 5 restricts the sample to middle schools, where different teachers teach math and English; Column 6 restricts the sample to elementary schools, where the same teacher teaches the two subjects. Column 1 includes only year fixed effects and no other controls. Columns 2 and 4 include school-by-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. Columns 5 and 6 control for the change in mean teacher VA in the other subject as well as year fixed effects.

TABLE 5
Quasi-Experimental Estimates of Forecast Bias: Robustness Checks

| Specification: | Teacher Exit Only | Full Sample | <25% Imputed VA | 0% Imputed VA |
|---|---|---|---|---|
| Dependent Variable: | Δ Score | Δ Score | Δ Score | Δ Score |
| | (1) | (2) | (3) | (4) |
| Changes in Mean Teacher VA across Cohorts | 1.045 | 0.877 | 0.952 | 0.990 |
| | (0.107) | (0.026) | (0.032) | (0.045) |
| Year Fixed Effects | X | X | X | X |
| Number of School x Grade x Subject x Year Cells | 59,770 | 62,209 | 38,958 | 17,859 |
| Percent of Obs. with Non-Imputed VA | 100.0 | 83.6 | 93.8 | 100.0 |

*Notes:* Each column reports coefficients from a regression with standard errors clustered by school-cohort in parentheses. The regressions are estimated on a dataset containing school-grade-subject-year means from the core sample described in Section II.C. The dependent variable in all specifications is the change in the mean test scores (year t minus year t-1) within a school-grade-subject cell. The independent variable is the change in mean teacher VA across consecutive cohorts within a school-grade-subject cell; see notes to Table 4 for details on the construction of this variable. All regressions are weighted by the number of students in the school-grade-subject-year cell and include year fixed effects. In column 1, we report 2SLS estimates, instrumenting for changes in mean teacher VA with the fraction of students in the prior cohort taught by teachers who leave the school multiplied by the mean-VA among these school-leavers. Columns 2-4 replicate the specification in Column 1 of Table 4, varying the way in which we handle classrooms with missing teacher VA. Column 2 includes all classrooms, imputing the sample mean VA (0) to classrooms with missing teacher VA. Column 3 replicates Column 2, excluding entire school-grade-subject-year cells in which more than 25% of student observations have missing teacher VA. Column 4 restricts to entire school-grade-subject-year cells with no missing teacher VA. The last row reports the percentage of students for whom teacher VA is not imputed in each estimation sample.

TABLE 6
Comparisons of Forecast Bias Across Value-Added Models

| | | Correlation with Baseline VA Estimates | Quasi-Experimental Estimate of Bias (%) |
|---|---|---|---|
| | | (1) | (2) |
| (1) | Baseline | 1.000 | 2.58 |
| | | | (3.34) |
| (2) | Baseline, no teacher FE | 0.979 | 2.23 |
| | | | (3.50) |
| (3) | Baseline, with teacher experience | 0.989 | 6.66 |
| | | | (3.28) |
| (4) | Prior test scores | 0.962 | 3.82 |
| | | | (3.30) |
| (5) | Student's lagged scores in both subjects | 0.868 | 4.83 |
| | | | (3.29) |
| (6) | Student's lagged score in same subject only | 0.787 | 10.25 |
| | | | (3.17) |
| (7) | Non-score controls | 0.662 | 45.39 |
| | | | (2.26) |
| (8) | No controls | 0.409 | 65.58 |
| | | | (3.73) |

*Notes:* In this table, we estimate seven alternative VA models and report correlations of the resulting VA estimates with the baseline VA estimates in Column 1. In Column 2, we report quasi-experimental estimates of forecast bias for each model, defined as 1 minus the coefficient in a regression of the cross-cohort change in scores on the cross-cohort change in mean teacher VA. These coefficients are estimated using exactly the specification in Column 1 of Table 4. All the VA models are estimated on a constant sample of students for whom all the variables in the baseline control vector are non-missing. All models are estimated separately by school level and subject; the correlations and estimates of forecast bias pool VA estimates across all groups. Each model only varies the control vector used to estimate student test score residuals in equation (4); the remaining steps of the procedure used to construct VA estimates (described in Appendix A) are the same for all the models. Model 1 replicates the baseline model as a reference; see notes to Table 3 for definition of the baseline control vector. The estimated forecast bias for this model coincides with that implied by Column 1 of Table 4. Model 2 uses all of the baseline controls but omits teacher fixed effects when estimating equation (5), so that the coefficients on the controls are identified from both within- and between-teacher variation as in traditional VA specifications. Model 3 includes all the controls in the baseline specification as well as indicators for years of teacher experience; teachers with 6 or more years of experience are pooled into a single group and are the omitted category. Model 4 includes only student, class and school level test score controls from the baseline control vector along with grade and year fixed effects. Model 5 includes only cubic polynomials in prior-year scores in math and English along with grade and year fixed effects. Model 6 replicates model 5, dropping the cubic polynomial for prior scores in the other subject. Model 7 removes all controls related to test scores from the baseline specification, leaving only non-score controls at the student, class, and school level (e.g., demographics, free lunch participation, etc.). Model 8 drops all controls except grade and year fixed effects.

APPENDIX TABLE 1
Structure of Analysis Dataset

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

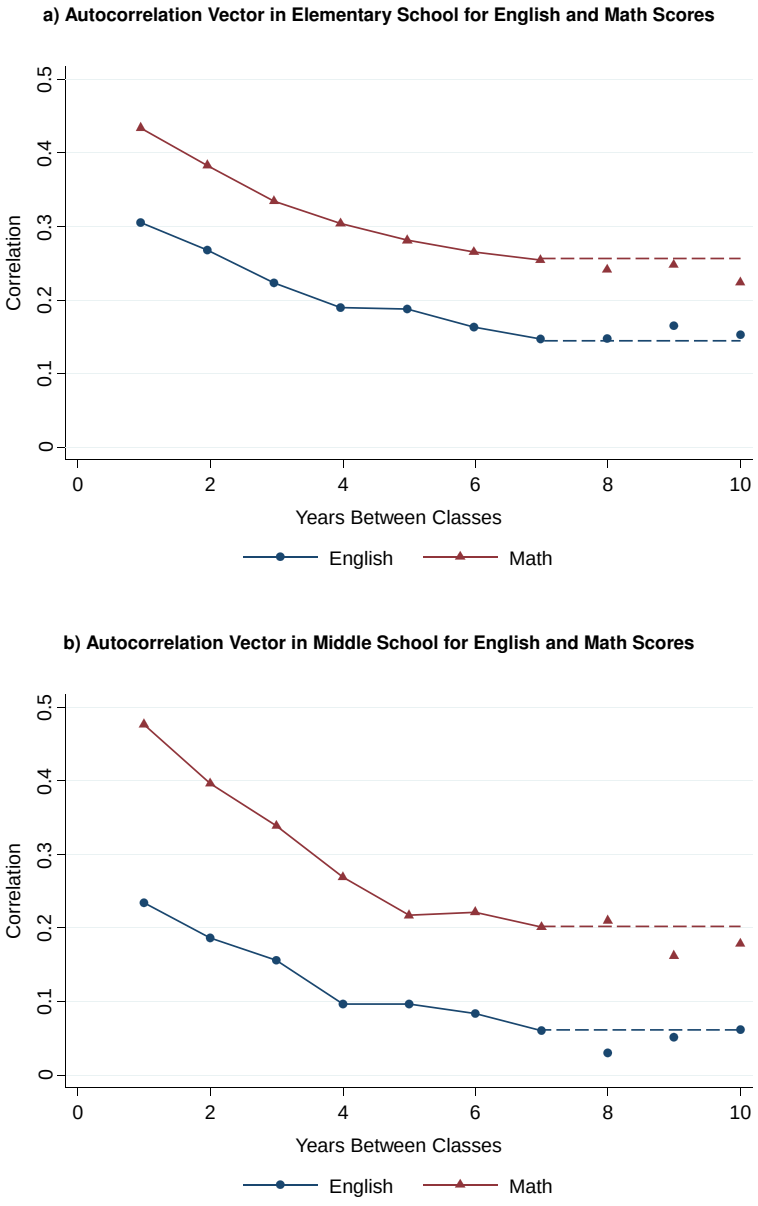
*Notes:* This table illustrates the structure of the core sample, which combines information from the school district database and the tax data. There is one row for each student-subject-school year. Students who were not linked to the tax data have missing data on parent characteristics. The values in this table are not real data and are for illustrative purposes only.

APPENDIX TABLE 2
Differences in Teacher Quality Across Students and Schools

| Dependent Variable: | Teacher Value-Added | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| Lagged Test Score | 0.0122 | | | 0.0123 | | | |
| | (0.0006) | | | (0.0006) | | | |
| Special education student | | -0.003 | | | | | |
| | | (0.001) | | | | | |
| Parent Income ($10,000s) | | | 0.00084 | 0.00001 | | | |
| | | | (0.00013) | (0.00011) | | | |
| Minority (black or hispanic) student | | | | | -0.001 | | |
| | | | | | (0.001) | | |
| School Mean Parent Income ($10,000s) | | | | | | 0.0016 | |
| | | | | | | (0.0007) | |
| School Fraction Minority | | | | | | | 0.003 |
| | | | | | | | (0.003) |
| Observations | 6,942,979 | 6,942,979 | 6,094,498 | 6,094,498 | 6,942,979 | 6,942,979 | 6,942,979 |

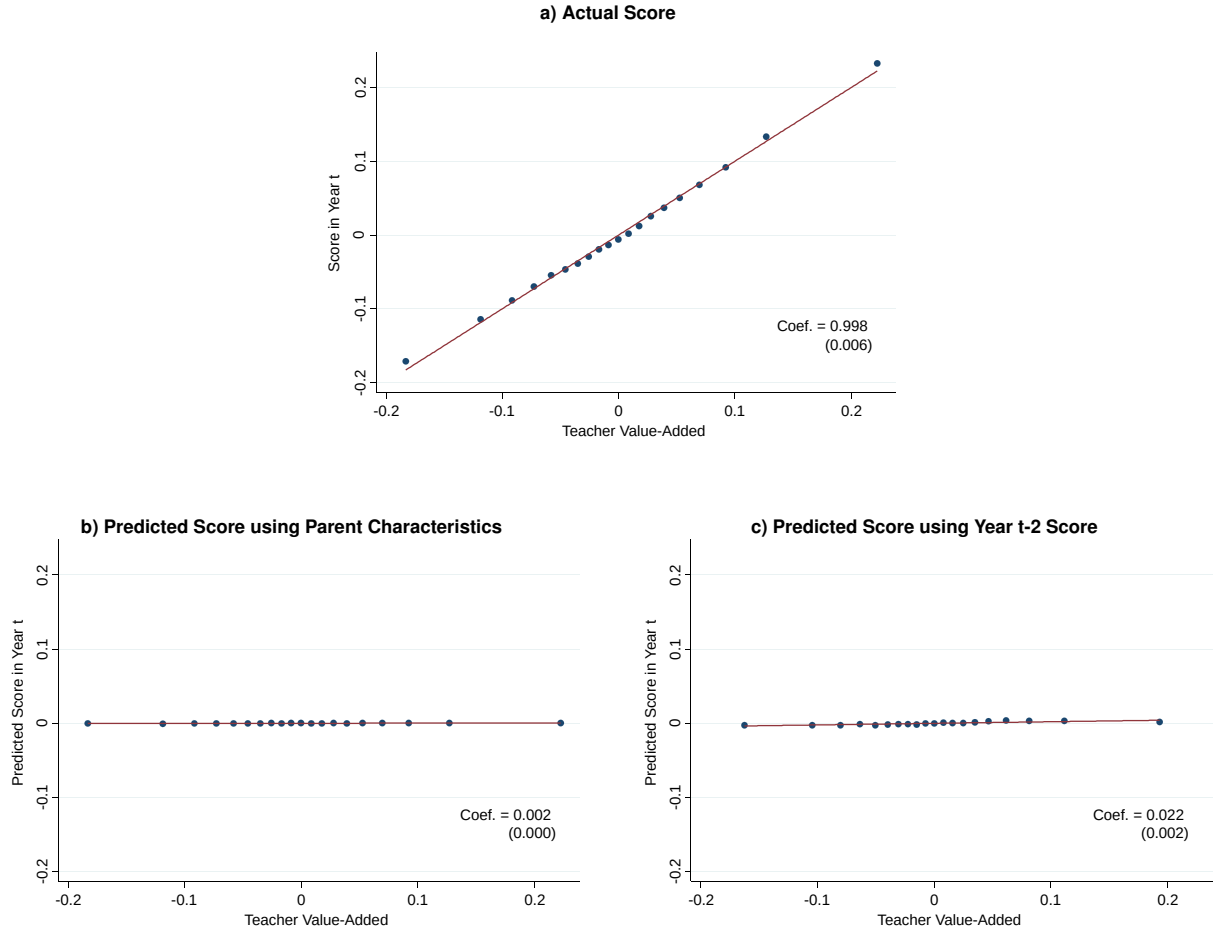
*Notes:* Each column reports coefficients from an OLS regression, with standard errors clustered by teacher in parentheses. Teacher VA, which is the dependent variable in all columns, is scaled in units of student test score standard deviations. Teacher VA is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections II.B and 4 and using the baseline control vector (see notes to Table 3 for more details). The regressions are run at the student-subject-year level on the sample used to estimate the baseline VA model. We multiply the resulting regression coefficients by 1.56 to account for the attenuation bias due to using VA estimates instead of true VA as the dependent variable (see Appendix D for details). Columns 3 and 4 restrict the sample to students whom we are able to link to parents in the tax data. Each specification includes the student-level covariate(s) listed at the left hand side of the table and no additional control variables. See notes to Table 1 for definitions of these independent variables. In Columns 6-7, the independent variable is the school-mean of the independent variables in Columns 3 and 5, respectively. We calculate these means as the unweighted mean across all student-subject-year observations with non-missing data for the relevant variable in each school.

# FIGURE 1
## Drift in Teacher Value-Added Across Years



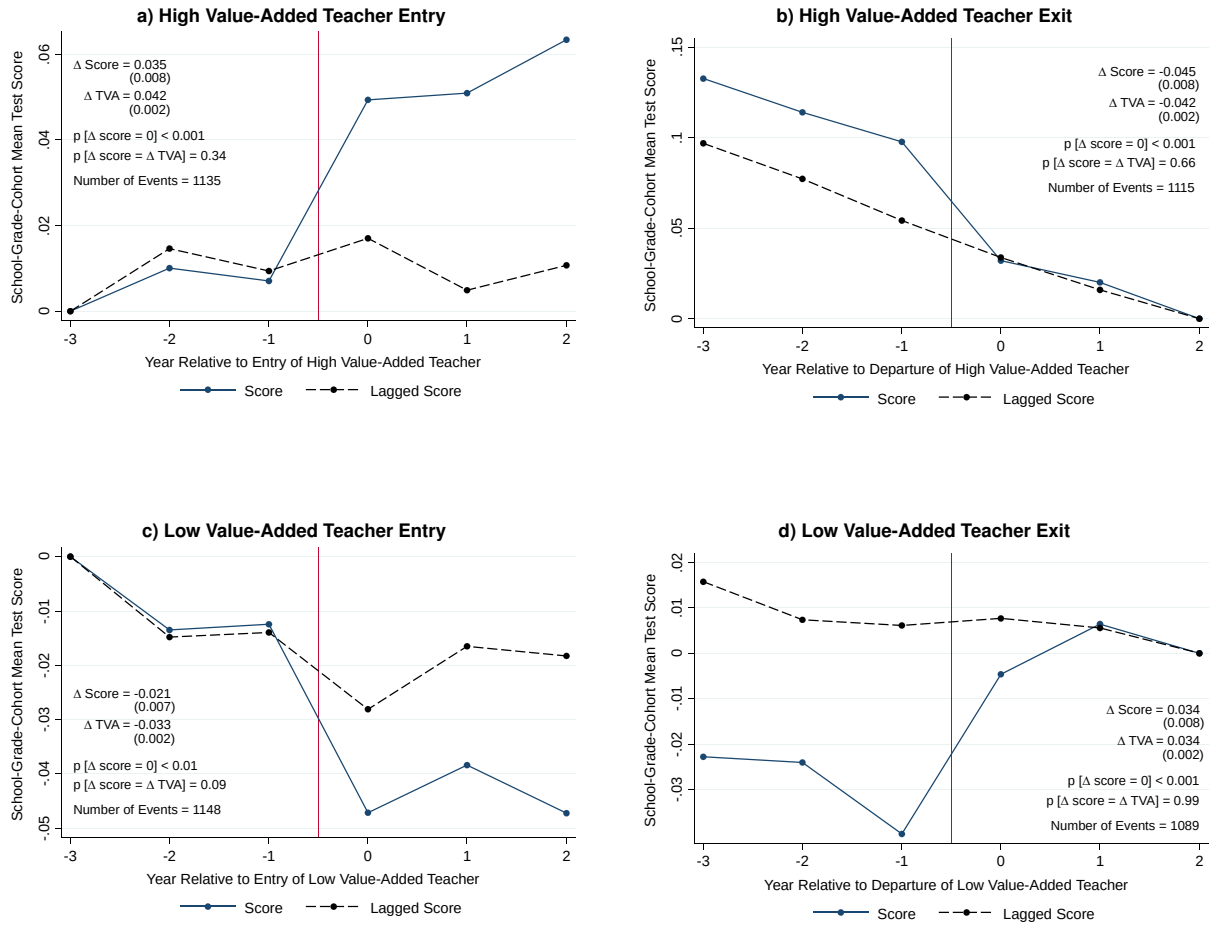

Notes: These figures show the correlation between mean test-score residuals across classes taught by the same teacher in different years. Panels A and B plot autocorrelation vectors for elementary and middle school. To calculate these vectors, we first residualize test scores using within-teacher variation with respect to our baseline control vector (see notes to Table 3). We then calculate a (precision-weighted) mean test score residual across classrooms for each teacher-year. Finally, we calculate the autocorrelation coefficients as the correlation across years for a given teacher, weighting by the sum of students taught in the two years. See Appendix A for more details on the estimation procedure for these and other parameters of the value-added model.

# FIGURE 2
## Effects of Teacher Value-Added on Actual and Predicted Scores




Notes: These figures pool all grades and subjects and are constructed using the sample used to estimate the VA model, which has one observation per student-subject-school year. The three panels are binned scatter plots of actual scores, predicted scores based on parent characteristics, and predicted scores based on twice-lagged test scores vs. teacher VA. These plots correspond to the regressions in Columns 1, 2, and 4 of Table 3 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 control vector (defined in the notes to Table 3) separately within each subject by school-level cell, using within-teacher variation to estimate the coefficients on the controls as described in Section I.B. We then divide the VA estimates $\hat{\mu}_{jt}$ into twenty equal-sized groups (vingtiles) and plot the means of the y-variable residuals within each bin against the mean value of $\hat{\mu}_{jt}$ within each bin. 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.

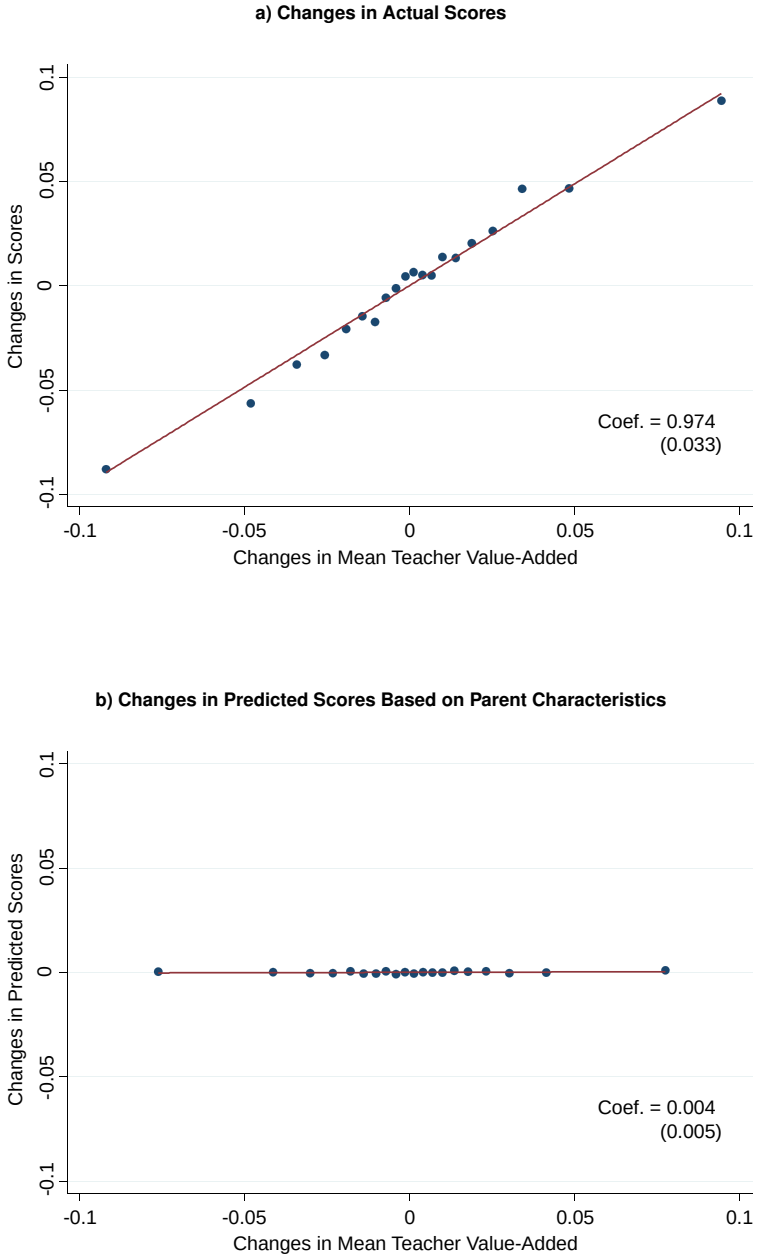
# FIGURE 3
## Impacts of Teacher Entry and Exit on Test Scores




Notes: These figures plot event studies of current scores (solid line) and scores in the previous school year (dashed line) by cohort as teachers enter or leave a school-grade-subject cell in year $t = 0$. Panels A and B analyze the entry and exit of a high-VA teacher (teachers with VA in the top 5% of the distribution); Panels C and D analyze the entry and exit of a low-VA (bottom 5%) teacher. All panels are plotted using the core sample collapsed to school-grade-subject-year means, as described in Section V.B. To construct each panel, we first identify the set of teachers who entered or exited a school-grade-subject cell and define event time as the school year relative to the year of entry or exit. We then estimate each teacher's value-added in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$. In Panel A, we identify the subset of teachers with VA estimates in the top 5% of the distribution among entering teachers. We then plot mean current and lagged scores in the relevant school-grade-subject cell for the event years before and after the entry of such a teacher. Panels B-D are constructed analogously. We demean test scores by school year to eliminate secular time trends and normalize the residual scores to zero at event year $t = -3$ for teacher entry and at event year $t = 2$ for teacher exit. Each panel reports the change in mean score gains (current minus lag scores) from $t = -1$ to $t = 0$ and the change in mean estimated VA. We report p-values from F tests of the hypotheses that the change in score gains from $t = -1$ to $t = 0$ equals the change in mean VA and that the change in score gains equals 0. Mean teacher VA is calculated using a student-weighted average, imputing the sample mean VA (0) for teachers who do not have data outside the $t \in [-3, 2]$ window necessary to compute a leave-out VA estimate.

# FIGURE 4

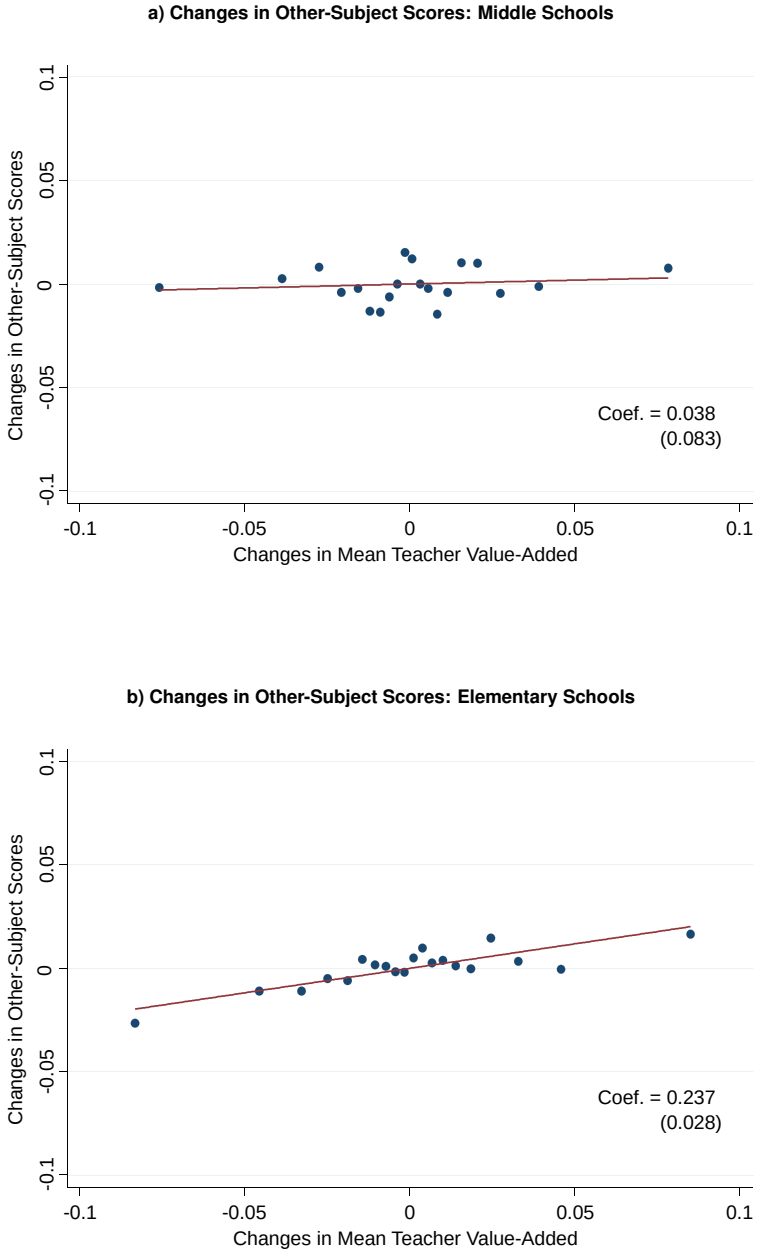
## Effects of Changes in Teaching Staff on Scores Across Cohorts




Notes: This figure plots changes in average test scores across cohorts versus changes in average teacher VA across cohorts, generalizing the event study in Figure 3 to include all changes in teaching staff. Panel A is a binned scatterplot of changes in actual scores vs. changes in mean VA, corresponding to the regression in Column 1 of Table 4. Panel B is a binned scatterplot of changes in predicted scores based on parent characteristics vs. changes in mean VA, corresponding to the regression in Column 4 of Table 4. See notes to Table 4 for details on variable definitions and sample restrictions. Both panels are plotted using the core sample collapsed to school-grade-subject-year means, as described in Section V.C. 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. The solid line shows the best linear fit estimated on the underlying micro data using a weighted OLS regression as in Table 4. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

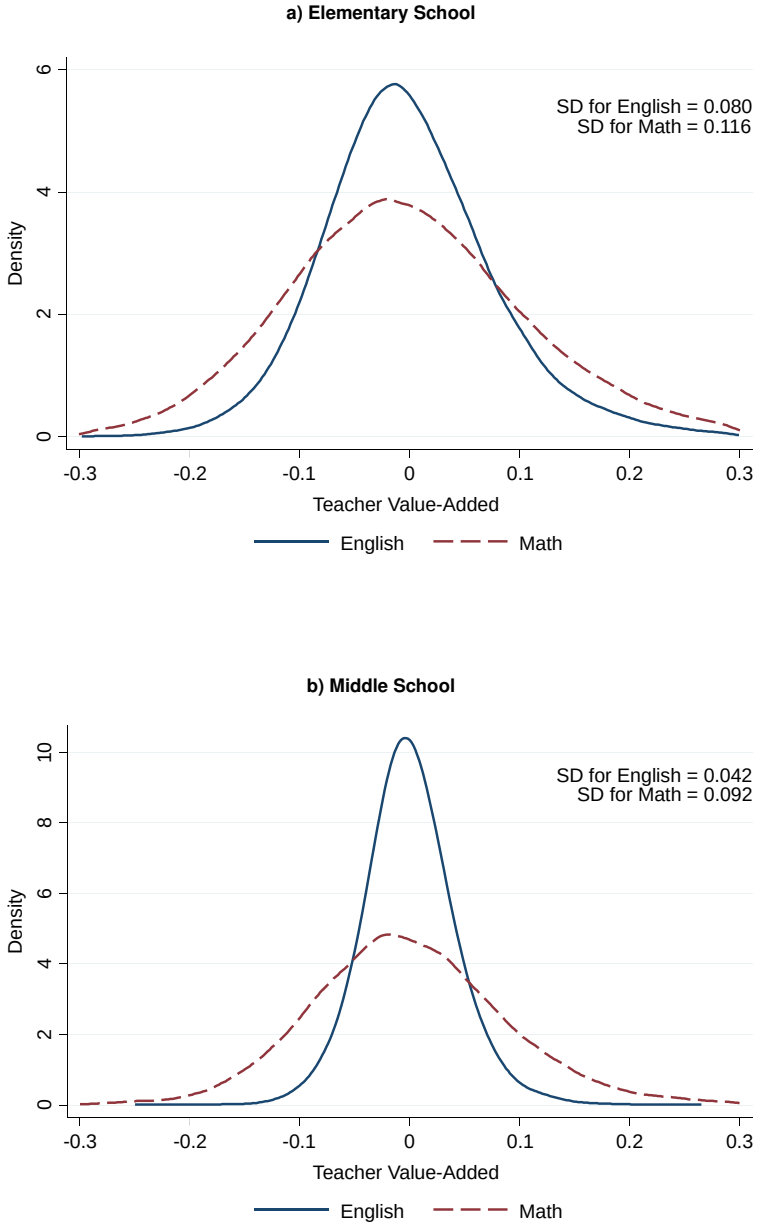
# FIGURE 5

## Effects of Changes in Teaching Staff on Scores in Other Subject




Notes: This figure plots changes in average test scores in the other subject across cohorts versus changes in average teacher VA, controlling for changes in other-subject VA. Panel A restricts the sample to middle schools, corresponding to the regression in Column 5 of Table 4. Panel B restricts the sample to elementary schools, corresponding to the regression in Column 6 of Table 4. See notes to Table 4 for details on variable definitions and sample restrictions. Both panels are plotted using the core sample collapsed to school-grade-subject-year means, as described in Section V.C. To construct these binned scatter plots, we first regress both the x- and y-axis variables on changes in mean teacher VA in the other subject as well as year fixed effects and compute residuals, weighting by the number of students in each school-grade-subject-year cell. We then divide the x residuals into twenty equal-size groups (vingtiles) and plot the means of the y residuals within each bin against the mean of the x residuals within each bin, again weighting by the number of students in each school-grade-subject-year cell. The solid line shows the best linear fit estimated on the underlying micro data using a weighted OLS regression as in Table 4. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

# APPENDIX FIGURE 1
## Empirical Distributions of Teacher VA Estimates




Notes: This figure plots kernel densities of the empirical distribution of teacher VA estimates $\hat{\mu}_{jt}$ for each subject (math and English) and school-level (elementary and middle school). The densities are weighted by the number of student test score observations used to construct the teacher VA estimate and are estimated using a bandwidth of 0.01. We also report the standard deviations of these empirical distributions of VA estimates. Note that these standard deviations are smaller than the standard deviation of true teacher effects reported in Table 2 because VA estimates are shrunk toward the mean to account for noise and obtain unbiased forecasts.