1%(5 :25.,1* 3$3(5 6(5,(6

0($685,1* 7+(,03$&76 2) 7($&+(56 ,
(9$/8$7,1* %,$6 ,1 7($&+(5 9$/8($''(' (67,0$7(6

5DM &KHWW\
-RKQ 1)ULHGPDQ
-RQDK (5RFNRII

:RUNLQJ 3DSHU
KWWS   ZZZ QEHU RUJ SDSHUV Z

1$7,21$/ %85($8 2) (&2120,& 5(6($5&+
0DVVDFKXVHWWV $YHQXH
&DPEULGJH 0$
6HSWHPEHU

:H DUH LQGHEWHG WR *DU\ &KDPEHUODLQ 0LFKDO .ROHVDU DQG
:H DOVR WKDQN -RVHSK $OWRQML -RVK $QJULVW 'DYLG &DUG
%ULDQ -DFRE 7KRPDV .DQH /DZUHQFH .DW] $GDP /RRQH\ 3K
<DJDQ DQRQ\PRXV UHIHUHHV WKH HGLWRU DQG QXPHURX
7KLV SDSHU LV WKH ILUVW RI WZR FRPSDQLRQ SDSHUV RQ WH
SUHYLRXVO\ FRPELQHG LQ 1%(5 :RUNLQJ 3DSHU 1R   HQWLW
7HDFKHU 9DOXH $GGHG DQG 6WXGHQW 2XWFRPHV LQ $GXOW
DQDO\VLV FRGH DUH SRVWHG DW KWWS   REV UF IDV KDUYDU
KHUHLQ DUH WKRVH RI WKH DXWKRUV DQG GR QRW QHFHVVDULO\
5HVHDUFK

1%(5 ZRUNLQJ SDSHUV DUH FLUFXODWHG IRU GLVFXVVLRQ DQ

0HDVXULQJ WKH ,PSDFWV RI 7HDFKHUV , (YDOXDWLQJ %LD
5DM &KHWW\ -RKQ 1)ULHGPDQ DQG -RQDK (5RFNRII
1%(5 :RUNLQJ 3DSHU 1R
6HSWHPEHU 0D\5HYLVHG

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?

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 λ 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 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 SDs).[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-

[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.

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 arrives. 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

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

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 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 .

nous class shocks $_c$, and idiosyncratic student-level variation $\ddot{\varepsilon}_{it}$. Let $\varepsilon_{it} = {}_c + \ddot{\varepsilon}_{it}$ denote the unobserved error in scores unrelated to teacher quality. Student characteristics X_{it} and ε_{it} may be correlated with $_{jt}$. Accounting for such selection is the key challenge in obtaining unbiased estimates of $_{jt}$.

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

Assumption 1 [Stationarity]

form

(5) $$A_{it} = {}_j + X_{it},$$

where ${}_j$ is a teacher xed effect. Our approach of estimating using within teacher variation differs from prior studies, which typically use both within- and between-teacher variation to estimate (e.g., Kane, Rockoff, and Staiger 2008, Jacob, Lefgren and Sims 2010). If teacher VA is correlated with X_{it}, estimates of in a speci cation without teacher xed effects overstate the impact of the X's because part of the teacher effect is attributed to the covariates.[6] For example, suppose X includes school xed effects. Estimating without teacher xed effects would attribute all the test score differences across schools to the school xed effects, leaving mean teacher quality normalized to be the same across all schools. With school xed effects, estimating within teacher requires a set of teachers to teach in multiple schools, as in Mans eld (2013). These switchers allow us to identify the school xed 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 $A_j^{-t} = (\bar{A}_{j1}, \ldots, \bar{A}_{j,t-1})'$ 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 $(E\left[{}_{jt} \mid A_j^{-t}\right]$

assumption in (3).

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

$$b_{jt} \; D \quad {}^{0}A_{j}^{\;t}. \tag{7}$$

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.

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, $\hat{\mu}_{jt} = E\left[\bar{A}_{jt} \mid A_j^{-t}\right]$ 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, $E\left[\mu_{jt} \mid A_j^{-t}\right]$, because part of the prediction could be driven by systematic sorting of students to teachers. We now turn to evaluating the degree to which $\hat{\mu}_{jt}$ measures a teacher's causal impact.

C. Definition of Bias

An intuitive definition of bias is to ask whether VA estimates $\hat{\mu}_{jt}$ 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 $\hat{\mu}_{jt}$ (constructed from observational data in prior years) in such an experiment:

$$A_{it} = \alpha_t + \lambda \hat{\mu}_{jt} + \varepsilon_{it} \tag{10}$$

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

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

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 $\hat{\mu}_{jt}$ is $B\left(\hat{\mu}_{jt}\right) = 1 - \lambda$.

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

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.

[10] Ex-post, any estimate of VA can be made forecast unbiased by redefining VA as $\hat{\mu}'_{jt} = (1 - B)\hat{\mu}_{jt}$. 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 $\hat{\mu}'_{jt}$ is still $(1 - B)\,\sigma\left(\hat{\mu}_{jt}\right)$. Hence, given the standard deviation of VA estimates, the degree of forecast bias is informative about the potential impacts of improving estimated VA.

atically assigned to certain teachers. Forecast unbiasedness only requires that the observable characteristics X_{it} used to construct test score residuals A_{it} are sufficiently rich that the remaining unobserved heterogeneity in test scores ε_{it} is balanced across teachers with different VA estimates. Second, even if VA estimates $\hat{\mu}_{jt}$ are forecast unbiased, one can potentially improve forecasts of μ_{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 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 speci c 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 of cial 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 pro ciency for the school years 1989 through 2009. The database used to code special education services and limited English pro ciency 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 rst 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

[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 signi cantly lower than the national average, as expected given the urban setting of the district.

[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 de ning 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.

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 identi ers (date of birth, state of birth, gender, and names) described in Appendix C, after which individual identi ers were removed to protect con dentiality. 88.6% of the students and 89.8% of student-subject-year observations in the 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 led 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 le 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 ve time-invariant parent characteristics, de ned as follows. We de ne 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 ler who rst claimed the child; measuring parent income in other years yields very similar results (not reported). For years in which parents did not le a tax return, they are assigned an income of 0. We measure income in 2010 dollars, adjusting for in ation using the Consumer Price Index.

We de ne marital status, home ownership, and 401(k) saving as indicators for whether the rst primary

[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 led during the calendar year $t+1$.

[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 de nition of parents is based on who claims the child as a dependent, and thus may not re ect the biological parent of the child.

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 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

[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.

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 X_{it} and teacher fixed effects, as in (5). Our baseline control vector 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) grade and year dummies. Importantly, our control vector 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 $\sigma_{As} = Cov(\bar{A}_{jt}, \bar{A}_{j,t-s})$, weighting by the total number of students in each pair of classes. In middle school, where teachers teach multiple classes per year, we estimate σ_{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 $t-s$, the VA estimate for teacher j in year t would simply be $\hat{\mu}_{jt} = r_s \bar{A}_{j,t-s}$, since $r_s = \frac{\sigma_{As}}{\sigma_A^2}$, as

[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.

[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.

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 $_{A0}$ 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, $= \sqrt{_{A0}}$, to be 0.098 in English and 0.134 in math in middle school (Table 2, Panel B).[21] In elementary school, $_{A0}$ 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 of on $\sqrt{_{A1}}$ because $_{A0}$ $_{A1}$. This bound implies that 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 , we 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 $_{A0}$. This method yields estimates of of 0.124 in English and 0.

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. 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 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 $\hat{\mu}_{jt}$ in observational data yields a coefficient of 1 because $\hat{\mu}_{jt}$ is the best linear predictor of A_{it}:

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

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

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

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 ε_{it}.

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

Assumption 2 [Selection on Excluded Observables] Students are sorted to teachers purely on excluded observables P:

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

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

$$A_{it} = \alpha_j + \rho P_{it}. \quad (13)$$

This leads to the feasible estimator

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

where $\hat{A}^p_{it} = \hat{\beta} P_{it}$ is estimated using (13). Equation (14) shows that forecast bias B_p can be estimated from an OLS regression of predicted scores $\hat{A}^p_{it}$ 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 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 P^*_{it} by regressing each element of P_{it} on the baseline control vector X_{it} and teacher fixed effects, as in (5). We then regress A^*_{it} on P^*_{it}, again including teacher fixed effects, and calculate predicted values $\hat{A}^p_{it} = \hat{\beta} P^*_{it}$. We fit separate models for each subject and school level (elementary and middle) as above when constructing the residuals P^*_{it} and predicted test scores $\hat{A}^p_{it}$.

In Column 2 of Table 3, we regress $\hat{A}^p_{it}$ on $\hat{\mu}_{jt}$, 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 coincides exactly with our estimate of B_p in Column 2 because A^p_{it} 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 X_{it}. The F-statistic on the parent characteristics in the regression of test score residuals A_{it} on 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 X_{it} is essentially unrelated to teacher VA. The correlation between A^p_{it} and $\hat{\mu}_{jt}$ 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 baseline control vector X_{it} captures much of the variation in $\Delta A_{i,t-2}$: the variation in class-average predicted score residuals based on $\Delta A_{i,t-2}$ is 0.048. The remaining variation in $\Delta A_{i,t-2}$ is not strongly related to teacher VA: the correlation between $\Delta A_{i,t-2}$ and $\hat{\mu}_{jt}$ is 0.037. If this correlation were 1, forecast bias would be ≤ 0.01, 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 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 determinayclsinlf995 o]TJ/

in mean test scores across cohorts t and t 1 and estimation error in 1 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 de ne the change in mean residual scores as 1 A_{sgt} D 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:

(15) $$1 A_{sgt} \; D \; a \; C \; b 1 \; Q_{sgt} \; C \; 1 \quad _{sgt}$$

The coef cient b in (15) identi es the degree of forecast bias as de ned in (10) under the following identi- cation 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:

(16) Cov

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 Mans eld 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 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 $1 A_{sgt}$ on $1 Q_{sgt}$. We therefore begin by analyzing changes in raw test scores $1 A_{sgt}$ across cohorts and then con rm that changes in control variables across cohorts are uncorrelated with $1 Q_{sgt}$. 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 rst 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 rst 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]

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 de ne 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 de ne 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 de ne 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]

[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 nd that teacher entry and exit yield broadly similar results, suggesting that spillovers across teachers are not a rst-order source of bias for our technique.

[32] learol-gradthe eveyears before and after the e ent.

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 variation comes from differential changes in teacher quality across subjects and grades within a school in a given year. The coefficient on ΔQ_{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 u77 -1.Ses7(and)-s and8 teachers' math and English VA are highly correlated ($\rho \simeq 0.6$) 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 ε_{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 ΔQ_{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 ΔQ_{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 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 5rof the id (16 d becaums

1 Q_{sgt} with this imputation procedure in the full sample is 0.877. The estimate falls because the imputation

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 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 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.

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 speci c 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

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

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

$$\Sigma_{A_{jt}\,mn} = \hat{\sigma}_{A,|s_m - s_n|}$$

and the mth element of γ_{jt} is

$$\gamma_{jt\,m} = \hat{\sigma}_{A,|t - s_m|}.$$

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 γ_{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 $\hat{\mu}_{jt}$.

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

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

as shown in (9). Let $\tilde{\mu}_j = \lim_{t\to\infty} \hat{\mu}_{jt}$ 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

$$\lambda_j = \tilde{\mu}_j - \mu_j. \tag{18}$$

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

VA estimates are biased at the teacher-level if they are inconsistent, i.e. if we systematically mispredict a forecast bias,

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

where the second step follows because $Cov(A_{it} - \mu_j, \hat{\mu}_j) = 0$ under random assignment in year t. Hence,

$$B(\hat{\mu}_j) = 0 \iff Var(\omega_j) + Cov(\mu_j, \omega_j) = 0.$$

This identity has two implications. First, if VA estimates are unbiased at the teacher-level, they must also be forecast-unbiased: $Var(\omega_j) = 0 \Rightarrow B(\hat{\mu}_j) = 0$. Second, and more importantly for our application, forecast-unbiased VA estimates can be biased at the teacher level only if $Cov(\mu_j, \omega_j) = -Var(\omega_j)$. Intuitively, if the teacher-level bias ω_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, $B(\hat{\mu}_j)$ could be 0 even if $Var(\omega_j) > 0$. In this sense, if a pre-specified value-added model produces VA estimates $\hat{\mu}_{jt}$ 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 $\hat{\mu}_{jt}$ without any additional inputs. In contrast, estimating teacher-level bias requires accounting for the impacts of estimation error in $\hat{\mu}_{jt}$ to construct the limit $\hat{\mu}_j$, 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

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 coef cients on the control vector 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 X_{it}, one would ideally regress teacher VA_{jt} on X_{it}:

(19) $$_{jt} \; D \quad C \quad X_{it} \; C \quad _{it}.$$

Since true VA is unobserved, we substitute VA estimates b_{jt} for $_{jt}$ on the left hand side of (19). This yields an attenuated estimate of because b_{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 by using b_{jt} as the dependent variable in (19) and multiplying the estimate of by $SD(_{jt})/SD(b_{jt})$. In the sample for which we observe lagged test scores, the standard deviation of teacher VA estimates is $SD(_{jt})/SD(b_{jt}) \; D \; 1{:}56$. We therefore multiply the estimate of obtained from estimating (19) with b_{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 $\hat{\mu}_{jt}$ 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 $\hat{\mu}_{jt}$ 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 $\hat{\mu}_{jt}$ 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.

	Elem. School English	Elem. School Math	Middle School English	Middle School Math

TABLE 3

Estimates of Forecast Bias Using Parent Characteristics and Lagged Scores

Dep. Var.:				
	(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 , , ,.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 innd hooli-8yeare 9.96 Tf 58.671 0 TTT0 9.96 Tf 1.36 Tw 18.995 0 Td16.32 test 4npr

Score	Score	Score	Predicted	Other Subj.

TABLE 5

Quasi-Experimental Estimates of Forecast Bias: Robustness Checks

	Teacher Exit Only	Full Sample	<25% Imputed VA	0% Imputed VA
Dependent Variable:	' Score	' Score	' Score	' 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

		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)		0.868	4.83
			(3.29)
(6)		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 aloe8tpe bac4s27etionsco4 .2-326sp[(co.8(ae;)- T* [(e52.2(4.8J T* [(Mdels)-5(aloe85-370.7(inclu06l)-361.7o6ris)-

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 (0.0006)			0.0123 (0.0006)			
Special education student		-0.003 (0.001)					
Parent Income ($10,000s)			0.00084 (0.00013)	0.00001 (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 , , 36 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

FIGURE 1

Drift in Teacher Value-Added Across Years

a) Autocorrelation Vector in Elementary School for English and Math Scores

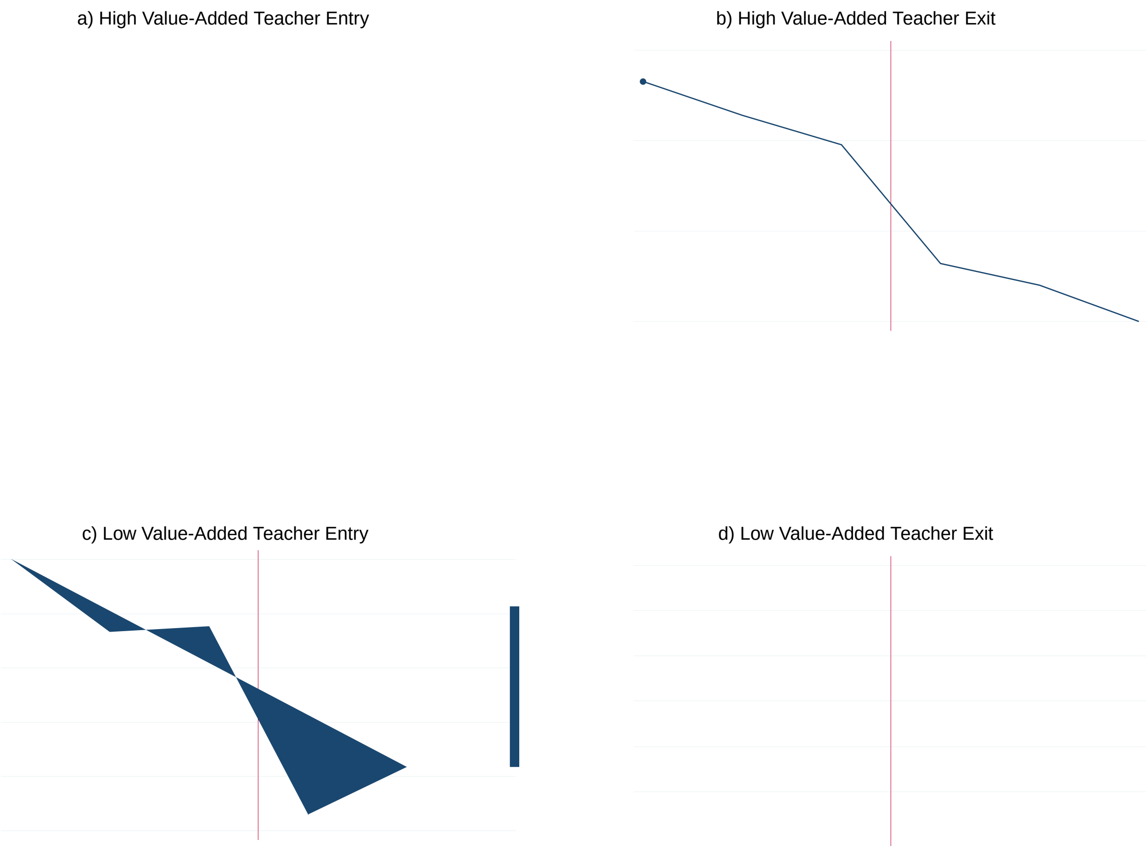
b) Autocorrelation Vector in Middle School for English and Math Scores

Notes: These gures show the correlation between mean test-score residuals across classes taught by the same teacher in di erent years. Panels A and B plot autocorrelation vectors for elementary and middle school. To calculate these vectors, we rst 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 coe cients 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

FIGURE 3

Impacts of Teacher Entry and Exit on Test Scores



Notes: These gures 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 yeart = 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 rst identify the set of teachers who entered or exited a school-grade-subject cell and de ne 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 2 [= 3; 2]. In Panel A, we identify the subset of teac2hers with VA estimates intry and [(2)]TJ/F19 8-388(e)-317(distribut617(in)-317(th (y)29(aw40n)-31 (y)2916-398(lea)2i(y)2916-ith)-316(V)

FIGURE 5

E ects of Changes in Teaching Sta on Scores in Other Subject

a) Changes in Other-Subject Scores: Middle Schools

b) Changes in Other-Subject Scores: Elementary Schools

Notes: This gure 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 de nitions 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 rst regress both the x- and y-axis variables on changes in mean teacher VA in the other subject as well as year xed e ects 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 t estimated on the underlying micro data using a weighted OLS regression as in Table 4. The coe cients show the estimated slope of the best- t line, with standard errors clustered at the school-cohort level reported in parentheses.