ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-354

Does Where Students Come From Affect Where Teachers Go?

Michael Gilraine
New York University

Odhrain McCarthy
New York University

We show that fade out biases value-added estimates at the teacher-level. To do so, we use administrative data from North Carolina and show that teachers' value-added depend on the quality of the teacher that preceded them. Value-added estimators that control for fade out feature no such teacher-level bias. Under a benchmark policy that releases teachers in the bottom five percent of the value-added distribution, fifteen percent of teachers released using traditional techniques are not released once fade out is accounted for. Our results highlight the importance of incorporating dynamic features of education production into the estimation of teacher quality.

VERSION: February 2021

Suggested citation: Gilraine, Michael, and Odhrain McCarthy. (2021). Does Where Students Come From Affect Where Teachers Go?. (EdWorkingPaper: 21-354). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/01kq-5k44

Does Where Students Come From Affect Where Teachers Go?*

Michael Gilraine and Odhrain McCarthy

February 12, 2021

We show that fade out biases value-added estimates at the teacher-level. To do so, we use administrative data from North Carolina and show that teachers' value-added depend on the quality of the teacher that preceded them. Value-added estimators that control for fade out feature no such teacher-level bias. Under a benchmark policy that releases teachers in the bottom five percent of the value-added distribution, fifteen percent of teachers released using traditional techniques are not released once fade out is accounted for. Our results highlight the importance of incorporating dynamic features of education production into the estimation of teacher quality. *JEL* Codes: H40, H75, I21, I22.

Keywords: Value-Added, Fade Out, Teacher Retention.

*Gilraine: Department of Economics, New York University, New York, NY 10003 (email: mike.gilraine@nyu.edu); McCarthy: Department of Economics, New York University, New York, NY 10003 (email: otm210@nyu.edu). We are grateful to Uros Petronijevic, Barbara Biasi, and Jesse Bruhn for insightful comments. We thank the North Carolina Education Research Data Center for providing the data. All remaining errors are our own.

I Introduction

It is widely accepted that teachers play a crucial role in education production. Research has long highlighted the importance of teachers but data limitations have curtailed progress (Hanushek, 1971, 1986). More recently, the availability of administrative matched student-teacher data has allowed researchers to precisely quantify the impacts of teachers – see pioneering studies by Rockoff (2004) and Rivkin et al. (2005) – and the effect of teachers on students' short- and long-run performance (e.g., as measured by test scores, college attendance, and earnings) are substantial: Chetty, Friedman, and Rockoff (2014b), for instance, estimate that replacing a teacher whose value-added ('VA') is in the bottom five percent with an average teacher would increase the present value of students' lifetime income by $400,000 per classroom.[1] In light of these findings, an increasing number of states are using VA measures to guide school personnel decisions.[2]

Crucial to the use of VA measures in personnel decisions is their ability to provide an unbiased estimate of a teacher's true quality. To ensure unbiasedness, practitioners condition on a large set of variables – most prominently lagged test scores – and a key focus of the literature has been determining whether this is sufficient to ensure VA estimates are "forecast-unbiased," defined as whether the VA of a randomly selected *group* of teachers equals the average test score growth of their students. The assessment of forecast-unbiasedness has been the subject of much debate (see Rothstein, 2010, 2017; Chetty, Friedman, and Rockoff, 2017), but in light of influential experimental (Kane and Staiger, 2008; Kane et al., 2013) and quasi-experimental evidence (Chetty, Friedman, and Rockoff, 2014a; Bacher-Hicks, Kane, and Staiger, 2014) a growing consensus is forming that teacher VA estimates are not meaningfully forecast-biased (Koedel, Mihaly, and Rockoff, 2015).

Nevertheless, the ability for lagged test scores to completely eliminate bias in VA estimates appears at odds with an empirical regularity in the education intervention literature that test score impacts universally "fade out" over time (e.g., Currie and Thomas, 1995; Krueger and Whitmore, 2001; McCaffrey et al., 2004; Jacob et al., 2010; Rothstein, 2010; Chetty et al., 2011, etc.). The strong fade out suggests that a substantial component of test score gains are transitory. If this is the case, the contemporaneous test score growth of students who experienced large test score gains in the prior year are likely to be muted as transitory knowledge disappears. Indeed, forecast unbiasedness does not imply the much stronger requirement of "teacher-level unbiasedness"– namely that each *indi-*

[1] Calculation: $522,000*1.34%*2.063*28.2. Refer to page 1672 in Chetty et al. (2014b).

[2] According to the National Council on Teacher Quality, thirty-four states use VA measures in their teacher evaluation systems as of 2019. By contrast, only fifteen states did so in 2009.

vidual teacher's VA estimate converges to their true quality. Chetty et al. (2014a) show that, under their specified education production function ('EPF'), the joint existence of forecast-unbiasedness and teacher-level biasedness only occurs in a knife-edge case.[3]

In fact, we show fade out implies that forecast-unbiasedness and teacher-level biasedness can jointly exist more generally. To do so, we adapt the Jacob, Lefgren, and Sims (2010) framework in specifying an EPF where teachers augment knowledge in ways that are both transient and permanent. Under this augmented EPF, knowledge gains imparted by teachers are positively correlated with fade out in the subsequent year, causing teacher-level bias as the estimated VA of a teacher depends upon the VA of the teacher that preceded them. Concretely, if students' test score gains in a given year are positively related to the amount of fade out they experience in the subsequent year, teachers persistently assigned students who had high- (low-) VA teachers in the prior year will have VA estimates that are downward (upward) biased because their students' test score growth will fall below (above) expectations.[4] This result holds even when teachers are initially assigned to classes at random, which would generate forecast unbiasedness.[5]

This fade out mechanism is distinct from the autocorrelated student-level error mechanism that could induce such results.[6] Specifically, Rothstein (2010) shows that fifth grade teachers' effects on fifth-grade scores are negatively correlated with their effects on fourth-grade scores. Chetty et al. (2014a) argue that the Rothstein (2010) result may be due to negatively autocorrelated student-level errors and advocate using jack-knife VA estimates to account for this. Throughout, we account for autocorrelated student-level errors by constructing jack-knife VA estimates and so any remaining relationship between a teacher's VA and the quality of the teacher that precedes them is driven by fade out.

Using detailed administrative data from North Carolina, we confirm the joint existence of forecast unbiasedness and teacher-level bias empirically. To show forecast unbiasedness, we follow Chetty et al. (2014a) and leverage quasi-experimental variation based on teacher staffing changes. The methodology involves regressing changes in school-grade test scores on changes in teacher VA; we find that this regression yields a point estimate of 1.005 (s.e. 0.035), indicating that, on average, there is a one-to-one relationship between the

[3]The knife-edge case occurs when the bias in each teacher's VA estimate and true VA offsets perfectly the variance in true VA. Intuitively, this occurs when bias co-varies negatively with true VA in such a way that the falling co-variance between true VA and estimated VA is exactly compensated by a proportionate fall in the variance of estimated VA.

[4]Where expectations are taken by conditioning on lagged test scores, as done in traditional value-added models, and so fade out is not accounted for.

[5]Refer to section II for a more formal exposition.

[6]The autocorrelated student-level error mechanism refers to the fact that when students do unobservably well in one year due to non-persistent factors (e.g., measurement error) their performance is likely to mean-revert the following year.

VA of teachers and the average test score growth of their students.

To reveal teacher-level bias, we adapt the methodology of Chetty et al. (2014a) and conduct event studies around the entry and exit of high- and low-VA teachers on the test score gains for these teachers' students in the *subsequent* grade (and year). We find that the entry of a teacher whose (jack-knife) VA is in the top five percent of all entrants causes a *decline* in mean test score gains for her students in the *subsequent* grade of 0.05σ. Similarly, the exit of such a teacher causes a 0.05σ *rise* in mean test score gains for her students in the *subsequent* grade.[7] To ensure our results are not driven by unobservable shocks, we also consider the impact a newly-entering or exiting teacher has on students' test score gains in the *grade below* and find no such test score changes for these students. As test score gains form the basis for constructing teacher VA, the VA of a teacher assigned students who receive high-quality teaching in the prior grade will be downward biased.

Next, we leverage all staffing changes to show more generally that the teacher VA in the prior grade last year impacts teacher VA in the subsequent year. Intuitively, we follow a teacher over her career and contrast her VA when she receives students who had low-VA teachers in the prior year to her VA when she receives students who had high-VA teachers in the prior year. Here, we find a one-standard deviation increase in mean teacher VA in the prior grade last year leads to a 0.052σ *decrease* in a teacher's VA in the current year.[8]

The dependence of VA measures on the teaching quality students experienced in the prior year raises important policy concerns. Not least, policymakers executing high-stakes personnel decisions may incorrectly release teachers who appear to be of low quality, but are just "unlucky" as his or her students were assigned high-quality teachers in the prior grade. These concerns are especially salient given that policymakers usually make personnel decisions on the basis of relatively few teacher-year observations, particularly so for teachers new to the profession.[9] To address these concerns, we account for fade out by controlling for test score growth in the prior year, which eliminates the relationship between the prior teacher's quality and the current teacher's quality. In addition, these VA estimates maintain forecast unbiasedness.

We find that accounting for fade out is policy-relevant. In particular, we hypothetically implement the widely-discussed policy proposed by Hanushek (2009, 2011) and evaluated in Chetty et al. (2014b) to release teachers in the bottom five percent of the teacher

[7]Low-VA teacher entry and exit event studies yield analogous results.

[8]Where VA is measured as a teacher-year fixed effect to allow VA to vary year-to-year.

[9]If teacher assignment to school-grades is (somewhat) random, teacher turnover will lead to the teacher-level bias in VA measures declining over time as teachers are evaluated using more years of data. As noted, however, teacher personnel decisions are made with relatively few years of data: The five most populous states award teacher tenure after only two to four years of experience.

VA distribution. Comparing the traditional VA measure to ours, we find that roughly fifteen percent of teachers who are released would not be if fade out was accounted for. Among these teachers, VA is one-quarter of a standard deviation higher. We then link these VA gains to student test scores along with longer-run outcomes such as high school dropout, suspensions, and SAT scores. Naturally, the higher levels of VA among affected classrooms leads to improvements in these outcomes. To gauge magnitudes, if we take benchmark estimates for earnings from Chetty et al. (2014b), the one-quarter standard deviation improvement in VA among affected classes would raise the net present value of classroom earnings by $45,000.

Our work contributes to three strands of literature. First, we advance the extensive literature on the estimation of teacher effects by helping eliminate teacher-level bias in VA estimates, aiding to reduce the many factors that can potentially plague VA estimates including measurement error, omitted variables, unobserved idiosyncratic shocks, and summer learning loss (e.g., Hanushek and Rivkin, 2010; McEachin and Atteberry, 2017; Gershenson and Hayes, 2018, etc.). Second, we contribute to the fade out literature (e.g., McCaffrey et al., 2004; Kane and Staiger, 2008; Jacob et al., 2010, etc.) by taking their insights and applying them to reduce teacher-level bias in VA estimates. Third, we add to our understanding of the education production technology, helping to reconcile or providing an alternative explanation for some prior findings.

In particular, our findings provide a causal explanation for the Rothstein (2010) result that fourth grade test score gains and the fixed effects of their fifth grade teachers are jointly significant.[10] Goldhaber and Chaplin (2015) demonstrate that the Rothstein (2010) result does not necessarily imply that VA estimates are biased. Specifically, they show that the sorting of students to teachers could induce this result, even if sorting is adequately controlled for via lagged test-scores.[11] In contrast, our findings indicate that the Rothstein (2010) result is partly driven by lagged test scores not fully accounting for periods of high test score growth due to fade out, potentially biasing VA estimates.

Nevertheless, since the sorting of students to teachers is limited (conditional on lagged test scores), forecast biasedness in VA measures is muted, re-inforcing the findings of

[10]Concretely, even if students are randomly assigned to teachers, fade out implies that teachers who randomly receive students with high-quality teaching in the prior period will have lower VA measures than those receiving students with low-quality teaching in the prior period. Refer to Section II for a more formal exposition.

[11]Formally, Goldhaber and Chaplin (2015) note that the Rothstein (2010) result is based on the correlation between fourth grade test score gains and fifth grade teacher assignment. Therefore, any sorting of students (e.g., by third grade test scores or fourth grade teaching quality) to fifth grade teachers would fail the Rothstein (2010) test even this sorting was adequately controlled for via lagged test scores.

Chetty et al. (2014a). Despite that, VA measures can still feature significant bias for some individual teachers. To that end, our work provides a theoretical justification as to how researchers can find forecast unbiasedness in VA estimates side-by-side with biased teacher-level VA estimates, echoing the results of Kane and Staiger (2008) and Kane et al. (2013).[12] Our results also provide an additional explanation for why teacher-year fixed effects are correlated within school-grades[13] and reinforce the findings of Koedel and Betts (2011) who highlight the value of using multiple years of data before making personnel decisions based on VA to mitigate non-persistent sorting.

More generally, our work speaks to the value of taking into account the dynamics of educational production when estimating the impacts of education inputs such as teachers (e.g., see Cunha and Heckman (2007); Cunha et al. (2010)). While dynamic features of the EPF are difficult to estimate, incorporating them into VA help us better target components of teaching that lead to knowledge gains that are unlikely to fade out, as in Macartney et al. (2018). Targeting knowledge components that do not fade out will then lead to large cumulative improvements in students' total knowledge, with these policies potentially raising the quality of teaching in this dimension through incentives (as in Biasi (forthcoming) for VA) or by selecting teachers that better impart long-term knowledge.

This paper is organized as follows. Section II evaluates forecast and teacher-level unbiasedness under an EPF that features short- and long-term knowledge components. The data are then introduced in Section III. Section IV presents our quasi-experimental design and communicates results, with Section V drawing out the policy implications. Section VI concludes.

II Conceptual Framework

This section augments the statistical model of Chetty et al. (2014a) to incorporate teacher-induced learning gains that have both short- and long-term components as in Jacob et al. (2010) and Cascio and Staiger (2012). We then assess the implications of this model for forecast unbiasedness and teacher-level biasedness. In particular, we show that teacher-level bias emerges in our model even if students are initially sorted randomly to teachers. We finish the section by highlighting the distinction between mean reversion

[12]In large-scale randomized control trials attempting to quantify the unbiasedness in teacher VA estimates, neither Kane and Staiger (2008) nor Kane et al. (2013) can statistically reject forecast unbiasedness. They also find, however, instances of large prediction errors at the individual-level.

[13]Without careful analysis, researchers may incorrectly interpret the correlation of teacher-year fixed effects within school-grades as a teacher peer effect. To avoid this concern, Jackson and Bruegmann (2009) carefully construct measures of a teacher's peer quality using (pre-determined) observable characteristics and VA estimated in an out-of-sample pre-period to show the importance of peers for teacher quality.

due to fade out and autocorrelated student-level errors.

II.1 Education Production Function

For simplicity, we assume that each teacher instructs one class per year (as in our empirical application). Teachers are indexed by j and differ in their quality in year t, μ_{jt}. To start, we follow Chetty et al. (2014a) and have that the test score of student i in year t assigned to teacher j, A^*_{ijt}, is given by:

$$A^*_{ijt} = \beta A^*_{ij,t-1} + \phi X_{ijt} + \mu_{jt} + \epsilon_{ijt}\,, \tag{1}$$

where $A^*_{ij,t-1}$ are lagged test scores, μ_{jt} is teacher j's contribution, X_{ijt} are observable determinants of student achievement (e.g., student characteristics), and ϵ_{ijt} represents idiosyncratic student variation.

Knowledge is this model is assumed to consist of transitory and permanent components as in Jacob et al. (2010). Intuitively, rote memorization might increase short-term knowledge, while learning a deep understanding of material will raise knowledge over a longer time horizon. We therefore model that the test score of student i in year t, A^*_{ijt}, can be divided into long-, A^{*L}_{ijt}, and short-term, A^{*S}_{ijt}, knowledge components:

$$A^*_{ijt} = A^{*L}_{ijt} + A^{*S}_{ijt}\,. \tag{2}$$

Short-term knowledge is assumed to be shallow and transient and so completely fades away between periods t and t+1. Therefore student achievement at time t is not a function of the lagged short-term knowledge component. Consequently, we can rewrite equation (1) as:

$$A^*_{ijt} = \beta A^{*L}_{ijt-1} + \phi X_{ijt} + \mu_{jt} + \epsilon_{ijt}\,. \tag{3}$$

Subbing equation (2) into (3) gives us:

$$A^*_{ijt} = \beta A^*_{ijt-1} - \beta A^{*S}_{ijt-1} + \phi X_{ijt} + \mu_{jt} + \epsilon_{ijt}\,. \tag{4}$$

The key difference in this model relative to Chetty et al. (2014a) is that test scores, A^*_{ijt}, are a function of the lagged long-term knowledge component, A^{*L}_{ijt}, and not lagged test scores, A^*_{ijt-1}. Since long-term knowledge is unobserved, however, the econometrician can only control for prior test scores, A^*_{ijt-1}, which include both the long- and short-term knowledge components, potentially biasing estimates of teacher VA, μ_{jt}.

For what follows we assume that teacher VA and student achievement follow a stationary process where, for simplicity, teacher quality is fixed over time and student-level errors are i.i.d.[14] (Section II.4 discusses the case where student-level errors are autocorrelated). Specifically,

Assumption 1 $\mathbb{E}[\mu_{jt}|t] = \mathbb{E}[\epsilon_{ijt}|t] = 0, \quad var(\mu_{jt}) = \sigma^2_{\mu}, \quad var(\epsilon_{ijt}) = \sigma^2_{\epsilon} \quad \forall \;\; t\,.$

Given this stationarity, we have that teacher VA is fixed over time and so $\mu_{jt} = \mu_j \;\; \forall \; t$.

Teachers and Knowledge Gain: Teachers augment knowledge in ways that are transitory and permanent. Given this, teachers with higher levels of value-added are likely to affect short-term knowledge differently than teachers with low levels of value-added. We parameterize this relationship with α as follows:

$$A^{*S}_{ijt} = \alpha\mu_j\,. \tag{5}$$

This parameterization implies that when $\alpha > 0$ teachers with higher test score contributions in period t also supply more to the short-term knowledge component. Intuitively, this make sense: teachers that increase test scores do so by augmenting both short- and long-term knowledge, since both components raise test scores. Therefore, we focus discussion on the case where $\alpha > 0$ in what follows (we also find this to be the case empirically).

Students in our context are assigned a new teacher every year. Given this, student i who is assigned to teacher j in period t was assigned to teacher k in the prior period, $t-1$. Equation (4) therefore becomes:

$$\begin{aligned} A^*_{ijt} &= \beta A^*_{ijt-1} - \beta A^{*S}_{ijt-1} + \phi X_{ijt} + \mu_j + \epsilon_{ijt} \\ &= \beta A^*_{ijt-1} - \alpha\beta\mu_k + \phi X_{ijt} + \mu_j + \epsilon_{ijt}\,, \end{aligned} \tag{6}$$

where the equality follows from equation (5).

Our model differs from that of Chetty et al. (2014a) in that equation (6) has the additional term $\alpha\beta\mu_k$ arising from the prior teacher imparting some short-term knowledge gain that fades out. To evaluate the impact of this unobserved short-term knowledge component on VA estimates, we next define the VA estimator for μ_j.

Estimation of Value-Added: We follow the prior literature and estimate VA under

[14] Teacher quality is assumed to be fixed for expositional simplicity and can be relaxed. Indeed, we follow Chetty et al. (2014a) and allow for drift in our empirical implementation. Accounting for drift generally makes teacher-level biasedness more pronounced under random sorting since it places greater weight on fewer observations, exacerbating the impact of fade out in the previous period.

the following model:

$$A^*_{ijt} = \beta A^*_{ijt-1} + \phi X_{ijt} + \mu_{jt} + \epsilon_{ijt}\,, \tag{7}$$

where the fixed effect μ_{jt} absorbs both the current and prior teachers' contributions. We then construct residual test scores, A_{ijt}, by removing the effect of observable characteristics, including lagged test scores:[15]

$$A_{ijt} = A^*_{ijt} - \hat{\beta} A^*_{ij,t-1} - \hat{\phi} X_{ijt}\,. \tag{8}$$

VA estimates are then constructed using empirical Bayes. Specifically, let $\bar{A}_{jt} \equiv \frac{1}{n}\sum_{i\in j} A_{ijt}$ denote the mean residual test score in the class taught by teacher j in year t and let $\bar{\mathbf{A}}_j^{-t} \equiv (\bar{A}_{j1}, ..., \bar{A}_{jt-1})'$ denote the vector of mean residual scores prior to year t in classes taught by teacher j. The following OLS regression is then run to obtain the best linear predictor of $\bar{A}_{jt}$:

$$\bar{A}_{jt} = \psi \bar{\mathbf{A}}_j^{-t}\,, \tag{9}$$

where $\mathbb{E}[\psi] = \frac{Cov(\bar{A}_{jt}, \bar{\mathbf{A}}_j^{-t})}{Var(\bar{\mathbf{A}}_j^{-t})}$. Using the empirical analogue for ψ, teacher j's VA in period t is then given by:

$$\hat{\mu}_{jt} = \hat{\psi} \bar{\mathbf{A}}_j^{-t}\,. \tag{10}$$

Armed with these equations the proceeding sections trace out the implications for forecast-unbiasedness and teacher-level biasedness under various assumptions about the sorting process of students to teachers. Before doing so, we formally define forecast-unbiasedness and teacher-level unbiasedness:

Definition 1 *Teacher VA estimates are forecast-unbiased if there is an unbiased relationship between mean test scores of students assigned to teacher j in year t and the corresponding VA estimates $\hat{\mu}_{jt}$. This is equivalent to testing if $\lambda = 1$, where λ is defined as:*

$$\lambda \equiv \frac{cov(A_{ijt}, \hat{\mu}_{jt})}{var(\hat{\mu}_{jt})}$$

Definition 2 *Teacher VA estimates are teacher-level unbiased if, as the number of classrooms each teacher instructs approaches infinity, each individual teacher's VA estimate approaches their true VA. Or, more formally, VA estimates are teacher-level unbiased if $lim_{t\to\infty}\hat{\mu}_{jt} = \mu_j \;\; \forall j$.*

[15] We follow Chetty et al. (2014a) and estimate β and ϕ using variation across students taught by the same teacher using an OLS regression of the form $A^*_{ijt} = \alpha_j + \beta A^*_{ij,t-1} + \phi X_{ijt}$.

II.2 Implications under Random Sorting

Consider a thought experiment where schools feature one class per grade and students do not switch schools. A total of J teachers are then randomly assigned to schools in two adjacent grades (e.g., grades 4 and 5) and remain at their posts for T periods.[16] With this setup some teachers acquire students who had a high-quality teacher in prior year, while other teachers are assigned students who received low-quality teaching in the prior year.[17]

Assume that student i is assigned teacher j in fifth grade and teacher k in fourth grade. Using equations (6) and (8), we have that in this case, $cov(\bar{A}_{jt}, \bar{A}_{j,t-s}) = cov(\mu_{jt} - \alpha\beta\mu_{k,t-1} + \bar{\epsilon}_t,\ \mu_{j,t-s} - \alpha\beta\mu_{k,t-s-1} + \bar{\epsilon}_{t-s}) = \sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu$ for all s $\neq$ t. Additionally, $var(\bar{A}_{jt}) = \sigma^2_A = \sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu + \frac{\sigma^2_\epsilon}{n}$ for all t. Here, we draw on the fact that $Cov(\mu_j, \mu_k) = 0$ due to random sorting. This means that equation (10) simplifies to:

$$\hat{\mu}_{jt} = \hat{\psi}\bar{\mathbf{A}}_j^{-t} = \bar{\mathbf{A}}_j^{-t} \frac{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu}{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu + \frac{\sigma^2_\epsilon}{n(t-1)}}\,, \tag{11}$$

and so $\psi = \frac{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu}{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu + \frac{\sigma^2_\epsilon}{n(t-1)}}$.

Claim 1 *VA estimates are forecast-unbiased.*

Proof. Here, we have:

$$\lambda = \frac{Cov(A_{it}, \hat{\mu}_{jt})}{Var(\hat{\mu}_{jt})} = \frac{Cov(\mu_{jt} - \alpha\beta\mu_{k,t-1} + \epsilon_{it}, \psi(\mu_{js} - \alpha\beta\mu_{k,s-1} + \bar{\epsilon}_s))}{Var(\psi(\mu_{js} - \alpha\beta\mu_{k,s-1} + \bar{\epsilon}_s))}\,.$$

Due to random assignment we have that $\mathbb{E}[\epsilon_{ijt}|\mu_{jt}] = 0$ and $Cov(\mu_j, \mu_k) = 0$ and so:

$$\lambda = \frac{1}{\psi}\frac{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu}{\sigma^2_\mu + \alpha^2\beta^2\sigma^2_\mu + \frac{\sigma^2_\epsilon}{n(t-1)}} = \frac{1}{\psi} * \psi = 1\,.$$

■

[16] The condition in which teachers remain at their posts and therefore repeatedly receive students from the same teachers in the prior year (roughly) corresponds to the manner in which education systems operate where teachers remain at the same school-grade for many years.

[17] Our thought experiment is set up this way to allow for easy comparison with Chetty et al. (2014a)'s definition of teacher-level unbiasedness where the VA estimates of individual teachers converge to their true quality as T $\rightarrow \infty$. Nevertheless, bias in teacher-level VA estimates over a short time horizon is of greater concern given that teacher tenure decisions are made using relatively few years of data. It is clear from our EPF that a given teacher's VA estimate would be biased over short time horizons under fairly weak conditions – namely, that at the time of assessment the teacher's students had received, on average, non-mean quality teaching in the prior year.

Claim 2 *Assuming $\alpha \neq 0$, VA estimates are biased at the teacher-level.*

Proof.

$$\hat{\mu}_{jt} = \mu_j - \alpha\beta\mu_k \neq \mu_j \quad \text{if} \;\; \mu_k \neq 0\,,$$

since $\bar{\mathbf{A}}_j^{-t} \rightarrow \mu_j - \alpha\beta\mu_k$ and $\psi \rightarrow 1$. ■

Intuitively, these results come from the fact that if we randomly select a teacher who receives students that previously had a high- (low-) VA teacher then these students will experience relatively large (small) short-term knowledge decay and that teacher's VA estimate will be downward (upward) biased. In contrast, if we randomly select a *group* of teachers, some would come after a high-VA teacher and others after a low-VA teacher, such that overall their average VA estimate would equal their students average residual test score growth and forecast-unbiasedness holds.

II.3 Implications under Non-Random Sorting

Now consider the case where teachers are non-randomly sorted to students such that $cov(\mu_j, \mu_k) = \sigma_{jk} \neq 0$. For simplicity, we maintain the orthogonality of the unobservable idionsyncratic component to VA (i.e., $\mathbb{E}[\epsilon_{ijt}|\mu_{jt}] = 0$).

Claim 3 *Assuming $\alpha \neq 0$, VA estimates are forecast-biased.*

Proof. In this case:

$$\lambda = \frac{1}{\psi} * \frac{\sigma_\mu^2 + \alpha^2\beta^2\sigma_\mu^2 - 2\alpha\beta\sigma_{jk}}{\sigma_\mu^2 + \alpha^2\beta^2\sigma_\mu^2 + \frac{\sigma_\epsilon^2}{n(t-1)} - 2\alpha\beta\sigma_{jk}} \neq \frac{1}{\psi} * \gamma\psi = \gamma \neq 1\,,$$

where γ is a scaling coefficient and the expression is not equal to one since the covariance term $2\alpha\beta\sigma_{jk} \neq 0$. ■

Claim 4 *VA estimates are biased at the teacher-level.*

Proof. This follows automatically from the fact that VA estimates are forecast-biased. ■

Claim 5 *If $cov(\mu_j, \mu_k) = \sigma_{jk} > (<)\ 0$ then VA estimates are downward (upward) biased.*

Proof. Since $cov(\mu_j, \mu_k) = \sigma_{jk} > (<)\ 0$ the scaling coefficient γ is less than (greater than) 1 and we have downward (upward) biasedness in VA estimates. ■

Intuitively, these results arise since if $cov(\mu_j, \mu_k) = \sigma_{jk} > 0$ then high-VA teachers are more likely to be followed by other high-VA teachers. Since the VA estimator does not account for the relatively large fade out of short-term knowledge imparted by the prior high-VA teacher, it is more difficult for the current teacher to raise test scores. Therefore, students assigned to high-VA teachers have residual test score growth below the actual knowledge imparted by those teachers. Consequently, we have downward biasedness in VA estimates. Analogously, if $cov(\mu_j, \mu_k) = \sigma_{jk} < 0$ we have upward biasedness in VA estimates.

II.4 Mean Reversion from Autocorrelated Student-Level Errors

This subsection highlights the theoretical distinction between mean reversion in student test scores due to fade out and mean reversion due to autocorrelated student-level errors. In particular, we consider the finding that fifth grade teachers impact their students' fourth grade test scores, as found in Rothstein (2010). In our model, this finding is consistent with the fifth grade teachers' VA being negatively correlated with the fourth grade teachers' VA due to fade out. Nevertheless, this finding would also be consistent with negatively autocorrelated student-level errors. We now proceed to show that while both mechanisms are able to generate this relationship, jack-knife estimators eliminate the autocorrelated student-level errors mechanism. Therefore, any relationship between a teacher's jack-knifed VA and the quality of the teacher that precedes them is driven by fade out.

Concretely, suppose that student-level errors are negatively autocorrelated such that $\epsilon_{ijt} = \nu_{ijt} - \gamma\nu_{ijt-1}$, where ν_{ijt} is i.i.d. with variance σ^2_ν. In this case, equation (6) becomes:

$$A^*_{ijt} = \beta A^*_{ijt-1} - \alpha\beta\mu_k + \phi X_{ijt} + \mu_j + \nu_{ijt} - \gamma\nu_{ijt-1}\,, \tag{12}$$

This more general error structure does not alter any of previous our claims: teacher-level biasedness and forecast unbiasedness still exist jointly under random sorting and forecast-biasedness and teacher-level biasedness occur under non-random sorting.[18]

As before, students assigned to teacher j in fifth grade (year t) were assigned to teacher k in fourth grade (year $t-1$). (We do not incorporate teacher assignments prior to fourth grade for expositional clarity). We gauge the effect a fifth grade teacher has on fifth grade test scores in year t by regressing:

$$A^*_{ijt,5} = \beta A^*_{ijt-1,4} + \phi X_{ijt,5} + \mu_{jt,5} + \psi_{ijt,5}\,, \tag{13}$$

[18]Formally, the assumption of random sorting on the idiosyncratic component (i.e., $\mathbb{E}[\nu_{ijt}, \nu_{ijt-1}|\mu_{jt}] = 0$) continues to ensure that $\mathbb{E}[\epsilon_{ijt}|\mu_{jt}] = \mathbb{E}[\nu_{ijt} - \gamma\nu_{ijt-1}|\mu_{jt}] = 0$. We can therefore follow the proofs in Sections II.2 and II.3 to obtain the same results with autocorrelated student-level errors.

where the subscripted number following the comma is a grade subscript, and $\mu_{jt,5}$ corresponds to teacher j's effect in year t on her students' fifth grade test scores. Note that the teacher effects have become year-specific as the autocorrelated error structure violates our stationarity assumption (Assumption 1) that made teacher effects time invariant. Under our EPF defined in equation (12), it must be that:

$$\mu_{jt,5} = \mu_j - \alpha\beta\mu_k + \bar{\nu}_{jt} - \gamma\bar{\nu}_{jt-1}\,. \tag{14}$$

Fourth grade test scores for the *same* students taught by teacher j in fifth grade are given by:

$$A^*_{ijt-1,4} = \beta A^*_{ijt-2,3} + \phi X_{ijt-1,4} + \mu_{kt-1,4} + \nu_{ijt-1} - \gamma\nu_{ijt-2}\,, \tag{15}$$

where $\mu_{kt-1,4}$ corresponds to the fourth grade teacher's effect on these students' fourth grade scores. Regressing fourth grade test scores, $A^*_{ijt-1,4}$, on the fifth grade teacher's effect on fifth grade test scores, $\mu_{jt,5}$, would therefore identify the following coefficient (using equations (14) and (15)):

$$\begin{aligned}\beta_{OLS} &= \frac{Cov(\mu_{jt,5}, A^*_{ijt-1,4})}{Var(\mu_{jt,5})}\\ &= \frac{Cov(\mu_j - \alpha\beta\mu_k + \bar{\nu}_{jt} - \gamma\bar{\nu}_{jt-1}, \beta A^*_{ijt-2,3} + \phi X_{ijt-1,4} + \mu_{kt-1,4} + \nu_{ijt-1} - \gamma\nu_{ijt-2})}{Var(\mu_{jt,5})}\end{aligned} \tag{16}$$

The coefficient identified by equation (16) is negative for two reasons. First, the negatively autocorrelated student-level error implies that $Cov(-\gamma\bar{\nu}_{jt-1}, \nu_{ijt-1}) < 0$, generating a negative coefficient, as noted in Chetty et al. (2014a). Second, students' test scores in period $t-1$ incorrectly contain the portion of knowledge induced by teacher k that has faded out, namely $Cov(-\alpha\beta\mu_k, \mu_{kt-1,4}) < 0$.

In our empirical implementation, we eliminate the portion of the negative relationship coming from the autocorrelation in student-level errors by using jack-knife estimates in the estimation of teacher j's VA, as in Chetty et al. (2014a). Specifically, our VA estimates carefully leave out any year that would contain the autocorrelated student-level error ν_{ijt-1}. Therefore, any relationship between a teacher's VA and the quality of the teacher that precedes them must be driven by fade out. We now proceed to show this empirically.

III Data and Construction of VA

This section details the rich administrative data that we use and the sample restrictions we make. We then detail how we operationalize the VA estimation procedure detailed above.

III.1 Data and Sample Restrictions

We use rich administrative data that cover all public school students from the state of North Carolina, available through the North Carolina Education Research Data Center.[19] Our data cover fourth and fifth grade from 1996-97 through 2010-11 and third grade from 1996-97 through 2004-05 and 2006-07 through 2008-09.[20] We provide a data overview here; further detail is available in Appendix A.

Our data contain yearly standardized test scores in mathematics and English along with encrypted identifiers for students and teachers. Students can therefore be tracked over time and linked to a teacher in any given year. We also have detailed demographic information including parental education, economically disadvantaged status, ethnicity, gender, limited English status, disability status, academically gifted status, and an indicator for grade repetition. These data cover around 4.5 million student-year observations, with 1.8 million students and 75,000 teachers.

We make several sample restrictions to end up at our final VA analysis sample (hereafter 'VA Sample'). Specifically, we restrict our data to: (i) students who can be matched to teachers, (ii) students with valid current and lagged mathematics scores, (iii) students classified as being in a single grade and school throughout the school year, (iv) classrooms with at least ten but fewer than forty students, and (v) non-special education classrooms. Our final VA sample consists of 2,675,254 student-year observations covering 1,385,237 students and 34,953 teachers.

Summary statistics for the full sample and the VA Sample are available in Columns (1) and (2) of Table 1, respectively. While the sample restrictions eliminate approximately forty percent of the observations, we see only minor differences in comparison to the full sample, with the VA sample showing slightly higher performance levels and being drawn

[19]Data citation: North Carolina Education Research Data Center (1996-2017).

[20]The third grade pre-test which we use as the lagged test score for third grade students is not available in 2005-06 and stopped being administered after 2008-09, necessitating the elimination of those years from our data for third grade. Third grade lagged test scores for English are also unavailable in 2005-06 and 2007-08.

from moderately higher socioeconomic backgrounds on average.[21]

In later sections, we account for fade out in our VA estimates using the 'Fade Out VA Sample,' with summary statistics for this sample being reported in column (3) of Table 1. To account for fade out, we require twice-lagged test scores to construct lagged test score *growth*. The requirement of twice-lagged test scores restricts the sample to fourth grade cohorts from 1997-98 through 2005-06 and 2007-08 through 2009-10 and fifth grade cohorts from 1997-98 through 2010-11. We also drop classes where prior year mathematics growth data is missing for fifty percent of the class to ensure that we can adequately control for fade out.[22] These restrictions lead the 'Fade Out VA Sample' to have forty percent fewer observations in comparison to the 'VA Sample', although there are only minor differences in student performance and demographics between the samples.

Long-Run Outcome Data: Our 'Fade Out VA Sample' is then linked to high school outcome data for policy analysis (see Section V). The high school outcome data cover a range of high school outcomes, including high school dropout, algebra scores, suspensions, PSAT scores, and SAT scores. These outcomes do not cover the entire 'Fade Out VA Sample' as some cohorts have yet to reach the required age to achieve that outcome (e.g., fourth grade students in 2010-11 will have not taken the SAT by 2016-17). Table B.1 describes each long-run high school outcome that we use and the cohorts in the Fade Out VA Sample that it covers. It also reports the match rate between our long-run outcome and eligible cohorts. The match rate is imperfect as any student leaving the North Carolina public school system between elementary school (grades 3-5) and high school will not be matched and some high-school outcomes are voluntary (e.g., the SAT). Nevertheless, we achieve match rates of over eighty percent for algebra scores (usually taken in grade 8), while match rates for the voluntary PSAT and SAT outcomes are around forty percent. For that reason, taking the SAT will also be used as a long-run outcome, both as it indicates an interest in college and because it highlights possible selection into taking the SAT.[23]

[21] Students in the VA sample being somewhat higher performing is similar to what Chetty, Friedman, and Rockoff (2014a) find (refer to their Table 1) and is likely related to the fact that we must drop students who are missing lagged test scores. A student may lack lagged test scores for several reasons: (i) entered the public North Carolina system in the past year, (ii) was not required to take the test last year (e.g., due to a learning disability), (iii) failed to take the test (e.g., due to sickness), or (iv) data was entered incorrectly. A missing lagged score due to these reasons are likelier to occur for lower performing students.

[22] The requirement for classes to have at least fifty percent of students with non-missing prior year mathematics growth scores represents a third of the 1,237,894 student-year observation drop from the VA Sample to the Fade Out VA Sample.

[23] We find no impact of being assigned higher VA teachers on PSAT taking rates. (Results available upon request).

Table 1: Summary Statistics

	Full Sample[1]	VA Sample[2]	Fade Out VA Sample[3]
	(1)	(2)	(3)
Cognitive Outcomes:			
Math Score (σ)	0.000	0.046	0.059
Reading Score (σ)	0.000	0.034	0.043
Student Demographics:[4]			
% Economically Disadvantaged[5]	46.3	44.6	43.7
% White	57.9	60.1	60.6
% Black	28.8	27.9	27.6
% Hispanic	7.4	6.5	6.5
% Asian	2.0	1.9	1.9
Parental Education:[6]			
% High School Dropout	11.5	10.5	10.1
% High School Graduate	47.3	46.9	46.6
% Vocational School Graduate	15.9	16.6	16.7
% College Graduate	25.3	25.9	26.6
Other Learning Measures:			
% Repeating Grade	1.5	1.5	1.0
% Limited English Proficiency	4.3	3.5	3.1
% Academically Gifted	11.9	12.8	15.8
% Disabled	5.3	5.0	5.1
# Students	1,843,217	1,385,237	995,948
# Teachers	74,986	34,953	21,656
# Student-Years	4,457,812	2,675,254	1,437,360

Notes:

[1] Our data cover fourth and fifth grade from 1996-97 through 2010-11 and third grade from 1996-97 through 2004-05 and 2006-07 through 2008-09.

[2] VA sample restrictions: (i) students who can be matched to teachers, (ii) students with valid current and lagged mathematics scores, (iii) students classified as being in a single grade and school throughout the school year, (iv) classrooms with at least ten but fewer than forty students, and (v) non-special education classrooms. See Appendix A for more details.

[3] Further reduces the VA Sample by requiring twice-lagged test scores which restricts the sample to fourth grade cohorts from 1997-98 through 2005-06 and 2007-08 through 2009-10 and fifth grade cohorts from 1997-98 through 2010-11. We also only keep classrooms which have valid math test score growth data for at least 50% of students.

[4] With the exception of parental education and economically disadvantaged (see notes 5 and 6), all variables have one percent or less of missing student-year observations. Summary statistics are reported for non-missing observations only.

[5] Economically disadvantaged status data are only available from 1998-99 onward. As such, approximately one-fifth of student-year observations are missing economically disadvantaged status.

[6] Parental education is only available for years 1996-97 through 2005-06. As such, approximately one-third of student-year observations are missing parental education. We also combine several education categories which are treated as separate in our control vector (as we include dummies for each educational category). 'College Graduate' includes parents with both college degrees and graduate school degrees, while 'Vocational School Graduate' includes parents with degrees from trade or business schools and those with degrees from community, technical or junior college.

III.2 Constructing Value-Added

We construct our VA estimates as described in Section II.1. Namely, we construct the best linear predictor of the teacher fixed effects, μ_{jt}. To do so, we first residualize test scores in equation (8), requiring us to explicitly define controls for the observable determinants of student achievement: lagged test scores, $A^*_{ij,t-1}$, and student demographics, X_{ijt}.

Control Vector: Our control vector for observable determinants of student achievement uses a similar set of controls to Chetty et al. (2014a,b). First, we control for lagged test scores with a cubic polynomial in prior-year scores in both mathematics and English, interacted with grade dummies.[24] Second, we control for student demographics including: parental education, economically disadvantaged status, ethnicity, gender, limited English status, disability status, academically gifted status, and grade repetition. Students with missing demographic data are assigned to a missing data group for that variable. We also include the following class and school-grade controls: (i) 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, (ii) class and school-grade means of all the student demographics, (iii) class size, and (iv) grade and year dummies.

When we account for fade out we add the following variables into our control vector: cubic polynomials of mathematics and English student test score growth in the prior grade last year, interacted with grade indicators. We also include cubic polynomials of mean (non-missing) mathematics and English prior-year student growth at the class and school-year level, interacted with grade indicators. For students missing prior year-growth data, we impute a value of zero and include missing indicators.

With our control vectors specified, we then calculate VA, allowing for drift.[25] The within-year variance components of our VA Sample are reported in column (1) of Appendix Table B.2. The standard deviation of our VA estimates is 0.178, which is very similar to the 0.180 estimate reported for the North Carolina data in Rothstein (2017).[26] We also report the autocorrelation vector that underlie these VA estimates, with the level of drift being very similar to that reported in Rothstein (2017). Column (2) then reports the variance and autocorrelation components when we control for fade out. To facilitate comparison among sub-samples, column (3) reports the same elements for the 'Fade Out VA Sample' when we do not include the additional fade out controls. Comparing columns

[24] When prior English test scores are missing, we set its value to zero and include an indicator for missing data interacted with the controls for prior mathematics test scores.

[25] To do so we use vam.ado, a STATA program published by Stepner (2013). We allow for a drift-limit of seven years.

[26] Refer to Column (4) of appendix Table A2 in Rothstein (2017).

(2) and (3), it is clear that the variance and drift of the VA estimates are very similar when fade out is and is not accounted for.

IV Empirical Methodology and Results

We proceed to show that fade out is a underlying component of the education production function and its correlation with VA leads to teacher-level bias. To do so, we start by looking at the impact that high- and low-VA teachers who switch schools have on the VA of teachers in the *subsequent* grade. We then show that teachers are no longer affected by the VA of the prior teacher once fade out is accounted for.

IV.1 School Switching Event Studies

To demonstrate that teacher VA estimates are affected by the quality of the teacher that precedes them, we start by conducting event studies around the entry and exit of high- and low-VA teachers. Our methodology mimics that of Chetty et al. (2014a), although we adapt it to investigate the impact of these events on test score gains in the *subsequent* grade. If fade out is a component of the underlying EPF and short-term knowledge gain is positively correlated to VA, as in our model, then we expect that entry of high- (low-) VA teachers will decrease (increase) test score gains for their students in the *subsequent* grade. As test score gains form the basis for constructing teacher VA, changes in test score gains caused by teacher entry will in turn affect the VA of teachers in the subsequent grade.

Methodology: Let event year '0' denote the school year a teacher enters or exits the school and define all other event years relative to that academic year (e.g., if a teacher enters a school in 2004-05, then event year '0' is 2004-05 and event year '1' is 2005-06 etc.). An entry event is defined as the the arrival of a teacher who did not teach in that school for the three preceding years and remains there for more than a year; an exit event is defined as the departure of a teacher who does not return to that school for at least four years.[27] A teacher is defined as high- (low-) VA if her estimated VA in her year of entry or exit is in the top (bottom) five percent of all entrants or leavers.[28]

[27]The asymmetric entry (exit) definition is due to the one year lag between the year of entry and its ensuing impact on the subsequent grade in the following year (i.e., event year '1'). We condition on the entering teacher remaining there for more than a year to alleviate teacher turnover which reduces treatment intensity. Results are robust to when we do not condition on the entering teacher remaining there for more than a year.

[28]In the case that multiple teachers enter or exit a school in the same grade at the same time, the mean VA of the switchers is used to decide whether the event falls in the top or bottom five percent of the VA distribution.

We estimate the VA of each entering teacher by excluding event years $t \in [0, 3]$ from their VA calculation. Analogously, we estimate the VA of each exiting teacher by excluding event years $t \in [-3, -1]$. The definitions of entry and exit combined with their corresponding jack-knife VA estimates ensure VA is calculated using data from students outside the six year school-grade event window and its preceding school-grade-year.[29]

We focus solely on the variation induced by school switchers because if we were to use variation generated by school-grade switchers (as in Chetty et al. (2014a)) this would induce spurious event study results in our context.[30] We limit our sample by only including events where we have test score data in the school-grade cell for at least three years before and after the event.[31]

Results: Panel A of Figure 1 plots the impact of high-VA teacher entry on mean test score gains across different cohorts in the *subsequent* grade. The solid line shows the event-year point coefficients throughout the six year event period, with whiskers denoting 95 percent confidence intervals. Test score gains are demeaned using year fixed effects to eliminate secular trends. We do not condition on any other covariates. We normalize test score gains to zero at event-year $t = 0$.

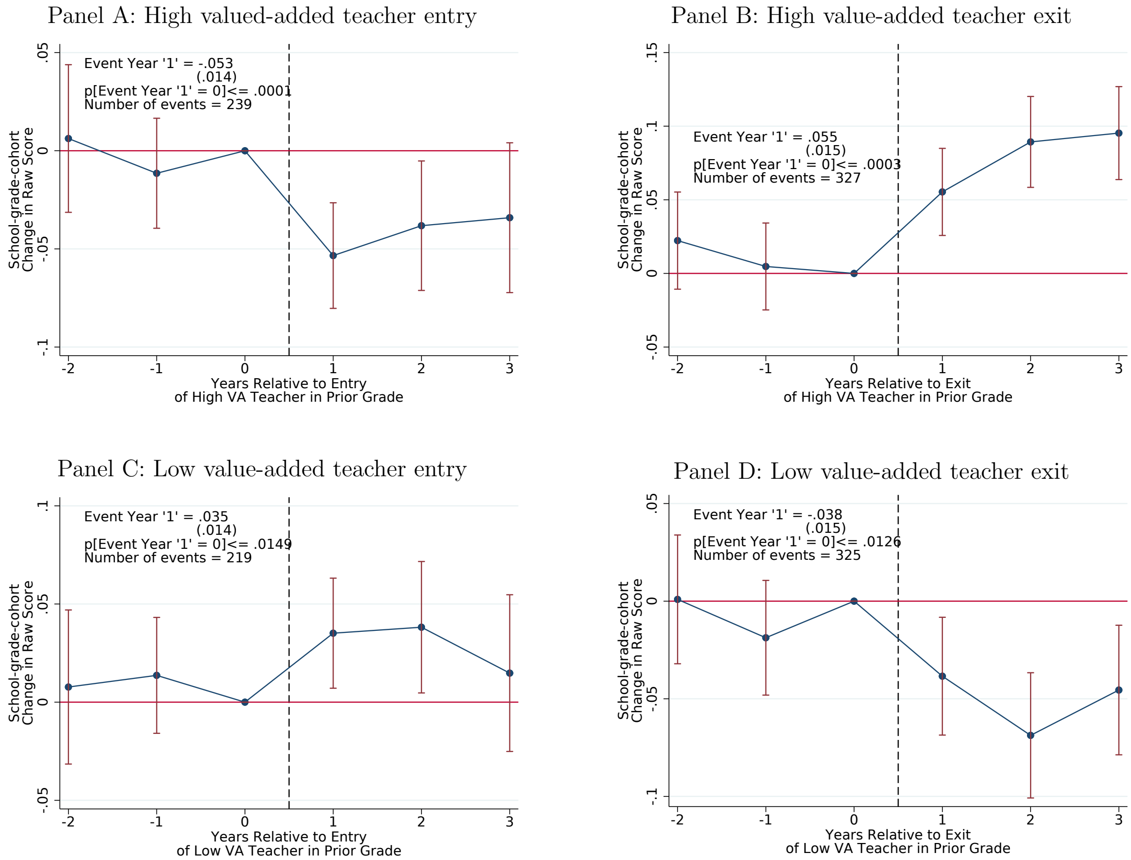
Panel A shows that when a high-VA teacher enters there is a steep drop in cohort test score gains in the subsequent grade in the following year. This decline in test score gains makes it significantly more difficult for the teacher of the subsequent grade to raise test scores above expectations, which in turn harms their VA estimates. This is exactly what one would expect due to fade out. For event time $t = 1$, the coefficient on mean test score gains is large and negative at -0.053 and is statistically significant at the one

[29]Since teacher VA is measured with error, calculating teachers' VA using test scores from the students within the event window creates a spurious correlation between VA estimates and test scores (Chetty et al., 2014a). Since the entering teacher was not in the school for event years $t \in [-3, -1]$, excluding event years $t \in [0, 3]$ from their jack-kinfe VA estimate ensures their VA calculation excludes student data within the event window years and its preceeding school-grade-year $t \in [-3, 3]$. Analogously, since the exiting teacher was not in the school for event years $t \in [0, 3]$, we exclude event years $t \in [-3, -1]$ when forming their jack-knife VA estimate.

[30]To see this, suppose a high-VA teacher switches from fifth grade to fourth grade within a school in 2004-05. Given that we investigate the impact of teacher switchers on teacher VA in the subsequent grade and year, event year '1' (i.e., 2005-06 in this instance) is the year of interest. The event study would then analyze the impact this entry has on the fifth grade cohort by comparing school-grade-cohort mean test scores three years before event year '1' (i.e., 2002-03 through 2004-2005) to school-grade-cohort mean test scores three years from event year '1' onwards (i.e., 2005-06 through 2007-08). Since fifth grade cohorts had a high-VA teacher before the event in 2003-04 (i.e., event year $t = -1$) and no longer have the high-VA teacher after the event in 2005-06 (i.e., event year $t = 1$) this would induce a spurious negative correlation in the event study.

[31]More specifically, if a teacher enters or exits school s in grade $g-1$ in year $t-1$, then we require that school-grade sg has test score data three years before and after year t. Following Chetty et al. (2014a), we stack the data and use the three years before and after each event for school-grades with multiple events occurring within six years (e.g., entry in both 2004-05 and 2006-07).

Figure 1: Impacts of Teacher Entry and Exit in the Preceding School-Grade-Year on Test Scores



Notes: These figures plot event studies of the entry and exit of high- and low-VA teachers in the preceding school-grade-year at event time $t = 0$ on current school-grade-cohort test score gains. Panels A and B plot the impact of the entry and exit of high-VA teachers. Analogously, Panels C and D plot the impact of the entry and exit of low-VA teachers. To construct panel A, we first identify the set of teachers who satisfy the definition of entry. Each year, we then compute the VA of every entrant leaving out event years $t \in [0, 3]$. We then identify the set of entrants who are in the top 5 percent of the distribution of all entrants, imputing a value of 0 for those entrants with missing VA estimates. We then plot the impact these entries have on the subsequent school-grade cell test score gains around the event window $t \in [-2, 3]$. Panel B is constructed by identifying the set of teachers who satisfy the definition of exit. Panels C-D are constructed analogously when the set of entrants/departers is defined as being in the bottom 5 percent of the distribution. Each year, the VA of every departer is computed leaving out event years $t \in [-3, -1]$ and the set of departers in top (bottom) 5 percent of the distribution are determined. Where necessary, we consider the student-weighted mean VA of entrants in cases where there are multiple entries into a single school-grade-year cell. We only keep events where we have school-grade data three years before and after the event to obtain a balanced panel. Test score gains are demeaned using year fixed-effects to eliminate secular trends. We do not condition on any other covariates. We normalize residual scores to zero at event year $t = 0$. Standard errors are clustered at the school-grade level. Each panel reports the event year $t = 1$ coefficient, standard error, and corresponding p-value.

percent level.

The remaining panels are constructed analogously. Panel B shows the impact of high-VA teacher exit on the subsequent grade. Here, there is a sharp increase in test score gains once the high-VA teacher exits in the previous year. This aligns with our fade out model as students now experience little fade out once a high-quality teacher exits, making it easier for teachers to increase cohorts' test scores in the subsequent grade. Panels C and D display results for the entry and exit of low-VA teachers. As expected, once a low-VA teacher enters (exits) the prior grade their is a significant increase (decrease) in cohort test score gains in the following year.

All panels show significant and sharp changes in test score gains upon the arrival or departure of a teacher whose quality is in the tail of the VA distribution. Moreover, they all have the appropriate sign that one would expect when the knowledge imparted by teachers fades out. These results indicate that the fade out phenomenon is an integral dynamic feature of the education production function that must be accounted for in the construction of VA as otherwise high- (low-) quality teachers will unduly punish (reward) teachers in the subsequent grade.

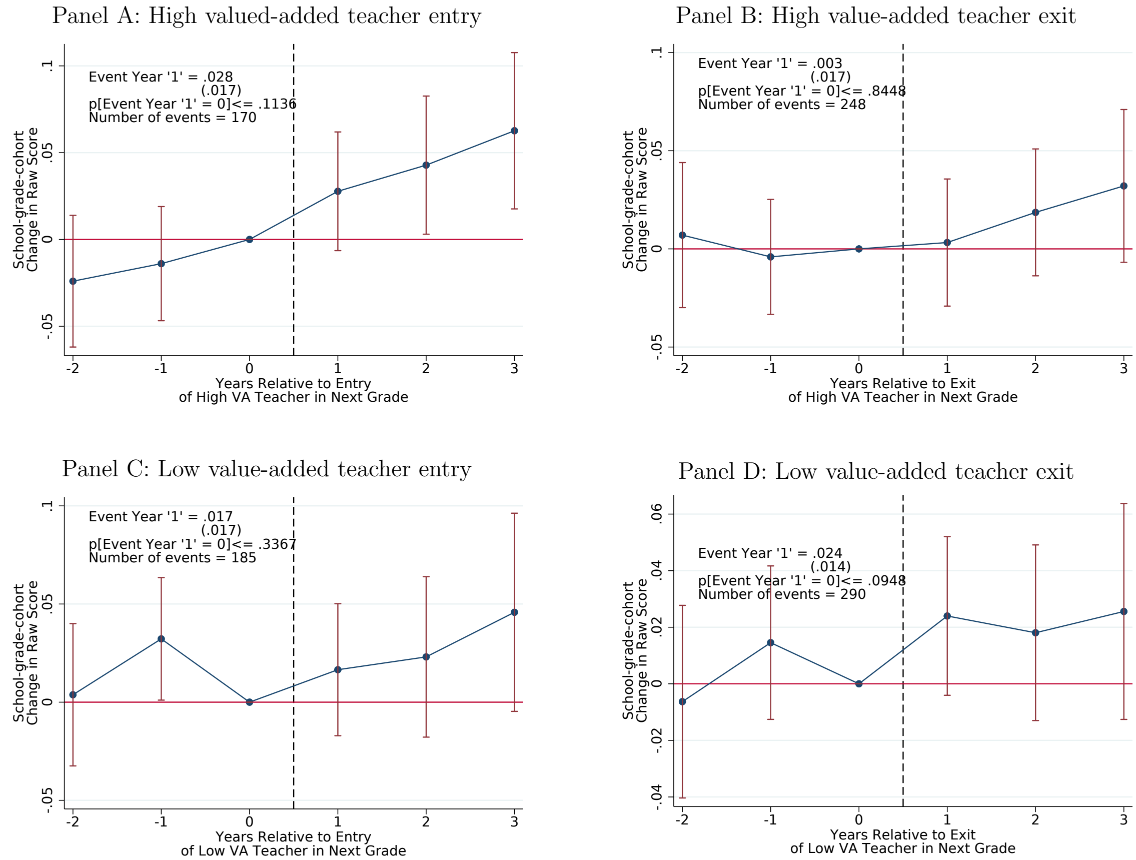
IV.2 Placebo Tests

The school-switching quasi-experimental design implemented above should eliminate concerns that our results are driven by spurious idiosyncratic, non-persistent factors (e.g., measurement error) since the VA estimates of the switchers exclude any student data used as the dependent variable in the event study. Nevertheless, there may be persistent concerns that the entry or exit of high- or low-VA teachers is correlated with unobservable determinants of student performance, compromising the event study design. We address these concerns by repeating the event study in Section IV.1, but looking at the impact of high- and low-VA school switchers on the estimated VA of teachers who teach students in the *grade below* the following year (as opposed to the *grade above* in the following year).

Figure 2 reports the results from these placebo tests. Panels B-D show that these teacher arrival or departure events have no impact on the grade below cohort test scores. Panel A does, however, show some impact of the arrival of a high-VA teacher on the test scores of students in the prior grade the following year. These effects appear partially driven by positive pre-trends in test scores and would arise if the entry of a high-VA teacher is correlated with an unobservable positive school shock (e.g., a school is looking to improve by continually attracting high-quality teachers over time). This type of correlation works *against* our results presented in Section IV.1, as there we found that the entry of a high-VA teacher causes a large and significant *decrease* in the VA of teachers in the proceeding grade. Our placebo results therefore serve to reinforce the validity of the quasi-experimental design in Section IV.1, with the correlation of any persistent school shocks to high-VA teacher entry working against our main results.

Furthermore, these results provide empirical validation for the Rothstein (2010) result that relates to the failure to account for fade out in the EPF (rather than autocorrelated student-level errors). In particular, Rothstein (2010) shows there is a negative correlation

Figure 2: Impacts of Teacher Entry and Exit in the Proceeding School-Grade-Year on Test Scores



Notes: These figures plot event studies of the entry and exit of high- and low-VA teachers in the *proceeding* (rather than *preceding* as in Figure 1) school-grade-year at event time $t = 0$ on current school-grade-cohort test score gains. Panels A and B plot the impact of entry and exit of high-VA teachers. Analogously, Panels C and D plot the impact of entry and exit of low-VA teachers. To construct panel A, we first identify the set of teachers who satisfy the definition of entry. Each year, we then compute the VA of every entrant leaving out event years $t \in [0, 3]$. We then identify the set of entrants who are in the top 5 percent of the distribution of all entrants imputing a value of 0 for those entrants with missing VA estimates. Where necessary, we consider the student-weighted mean VA of entrants in cases where there are multiple entries into a single school-grade-year cell. We then plot the impact these entries have on the preceding school-grade cell test scores around the event window $t \in [-2, 3]$. We only keep events where we have school-grade data three years before and after the event to obtain a balanced panel. We demean test score gains by year fixed-effects to eliminate secular trends and do not condition on any other covariates. We normalize residual scores to zero at event year t=0. We cluster standard errors at the school-grade level. Panels C-D are constructed analogously. Each panel reports the event year t=1 coefficient, standard error and corresponding p-value.

between a fifth grade teacher's effect on fifth grade test scores and their effect on fourth grade test scores. Chetty et al. (2014a) argue that the regression Rothstein (2010) runs may include estimation error on both the left-hand-side and right-hand-side potentially creating spurious correlation and advocate using jack-kinfe VA estimates to account for this. Nevertheless, per these results, if a fifth grade teacher was preceded by a newly entering high- (low-) VA teacher in the fourth grade then the fifth grade teacher's VA estimate will now include a portion of the fourth grade teacher's VA that corresponds to fade out and that is independent of autocorrelated student-level error, yielding the negative correlation reported by Rothstein (2010). If instead we were to condition on a newly entering high- (low-) VA teacher in fifth grade and look at their effects on fourth grade students we should find no effect, as verified by the placebo results in Figure 2.

IV.3 Quasi-Experimental Estimates of Bias

The preceding results focus exclusively on variation induced by the tails of the distribution of school switchers. We now turn to leveraging variation from the entire distribution to display the teacher-level biasedness that fade out induces more generally. Concretely, consider following a teacher over their career. Suppose in a given year that teacher receives students who had low-VA teachers in the prior year. Those students will have little knowledge fade out and so it should be relatively easy to raise those students' test scores in the current year. Conversely, suppose in the following year that same teacher receives students who had high-VA teachers from the prior grade last year (e.g., because the low-VA teachers in the prior grade last year were replaced with higher VA teachers). We would now expect it to be more difficult to raise those students' test scores. As such, we expect that the teacher's VA will fall across the two cohorts due to a rise in the prior-year teaching quality across the two cohorts.

Formally, let $\hat{\mu}_{jt}^{-\{t-\tau,...,t+\tau\}}$ be the estimated VA of teacher j in year t calculated excluding years $t \in \{t-\tau, ..., t+\tau\}$ and define n_{jt} as the class size of teacher j in year t. Further, define $Q_{s,g-1,t-1}^{jt} \equiv \frac{\sum_k n_{k,t-1}\hat{\mu}_{k,t-1}^{-\{t-2,\ t-1,\ t\}}}{\sum_k n_{k,t-1}}$ as the student-weighted average jack-knife VA (excluding years $t = \{t-2, t-1, t\}$ in VA calculation) of teachers who are assigned to grade $g-1$ in year $t-1$ at the same school s that teacher j teaches grade g in year t. $Q_{s,g-1,t-1}^{jt}$ is therefore the (jack-knife) estimate of the teaching quality students of teacher j received in the prior grade last year. Taking first differences, we define $\Delta Q_{s,g-1,t-1}^{jt} \equiv Q_{s,g-1,t-1}^{jt} - Q_{s,g-1,t-2}^{j,t-1}$,[32] and so $\Delta Q_{s,g-1,t-1}^{j}$ is the change in the teaching quality teacher j's current class received in the prior grade relative to the teaching quality teacher j's previous class received in the prior grade. By leaving out years $t-2$, $t-1$ and t in the calculation of VA for teachers in the prior grade last year, we ensure that changes in VA are being driven by changes in teaching staff rather than changes in VA estimates

[32]To be explicit, the latter term $Q_{s,g-1,t-2}^{j,t-1} \equiv \frac{\sum_k n_{k,t-2}\hat{\mu}_{k,g-1,t-2}^{-\{t-2,\ t-1,\ t\}}}{\sum_k n_{k,t-2}}$ and so is the the (jack-knife) estimate of the teaching quality students of that teacher j's last class receiving in the prior grade, excluding years $t = \{t-2, t-1, t\}$ in the VA calculation.

as done in Chetty et al. (2014a).[33]

Let A^{jt}_{sgt} be the mean residual test scores of teacher j's students in year t (i.e., the teacher-year fixed-effect for teacher j in year t). Define $\Delta A^{jt}_{sgt} \equiv A^{jt}_{sgt} - A^{j,t-1}_{sg,t-1}$ as the change in the mean residual test scores of teacher j between years t and $t-1$. We then regress:

$$\Delta A^{jt}_{sgt} = a + b\Delta Q^{jt}_{s,g-1,t-1} + \Delta\chi^{jt}_{sgt}\,. \tag{1}$$

The coefficient b identifies the degree of teacher-level biasedness due to changes in teaching quality in the prior grade last year for teacher j under the following identification assumption:

Assumption 2 *Across cohort changes in the prior year teaching quality for students assigned to teacher j are orthogonal to changes in other determinants of teacher j's students' residual test scores:*

$$cov(\Delta Q^{jt}_{s,g-1,t-1},\ \Delta\chi^{jt}_{sgt}) \ = \ 0 \tag{2}$$

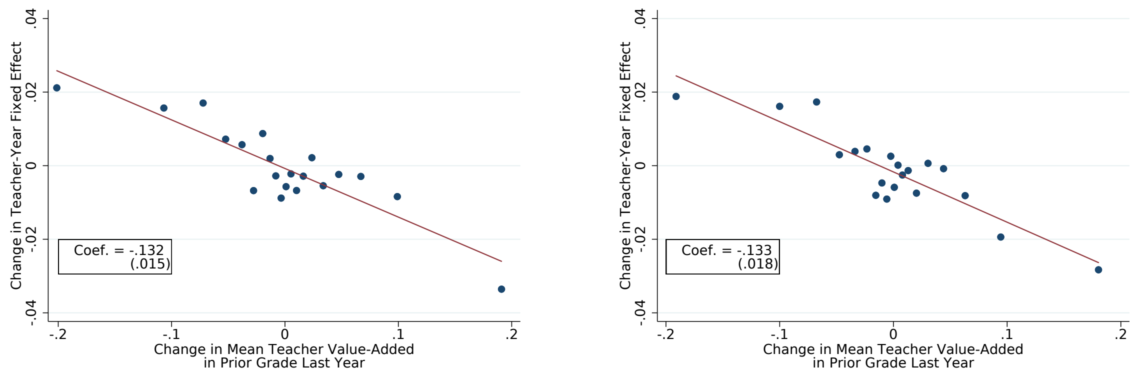
Under Assumption 2, we would expect $b = 0$ if VA estimates did not feature any teacher-level bias. In particular, we would expect the change in the quality of teaching teacher j's students received in the prior grade to have no impact on the change in residual test scores. As in Chetty et al. (2014a), residual test scores control for the large set of control variables that proxy for student, parent, and classroom quality. Nevertheless, these controls will be insufficient in the presence of fade out as they only control for the level of student performance in the prior year. To account for fade out, both the level and change in performance must be accounted for as two students could perform at the same level, but have experienced different knowledge gains in the prior year, which will subsequently fade out at different rates.

Results: Panel A of Figure 3 presents a binned scatter plot of the changes in teacher-year fixed-effects against changes in (jack-knifed) mean teacher VA in the prior grade last year. To construct the figure, we first collapse our sample to the teacher-year level and then compute the (jack-knifed) student-weighted teacher VA in each school-grade cell.
From this data, we can then calculate changes in the teacher-year fixed-effect, ΔA^{jt}_{sgt}, along with changes in (jack-knifed) mean teacher VA in prior grade last year, $\Delta Q^{jt}_{s,g-1,t-1}$. Changes in (jack-knifed) mean teacher VA in the prior grade last year is then divided into

[33] Concretely, when forming teacher-level jack-knife VA estimates $\hat{\mu}_{jt}$ we exclude the same years $t \in \{t-2, t-1, t\}$ for computing $Q^{j}_{s,g-1,t-1}$ and $Q^{j}_{s,g-1,t-2}$ and so the VA of any individual teacher present in both $Q^{j}_{s,g-1,t-1}$ and $Q^{j}_{s,g-1,t-2}$ will not have materially changed. Thus, the majority of variation in $\Delta Q^{j}_{s,g-1,t-1}$ must come from personnel changes.

Figure 3: Impact of Teacher VA in Prior Grade Last Year on Teacher-Year Fixed-Effect

Panel A: Three year leave-out VA Estimates | Panel B: Five year leave-out VA Estimates



Notes: These figures plot the changes in teacher-year fixed-effects against the changes in mean teacher VA in the prior grade last year. To construct panel A, we first exclude classrooms with missing VA estimates and then collapse the core-sample to the teacher-year level. We then compute the student-weighted (jack-knife) teacher VA in each school-grade cell. Changes in the teacher-year fixed-effect, ΔA^{jt}_{sgt}, along with changes in mean (jack-knifed) teacher VA in prior grade last year, $\Delta Q^{jt}_{s,g-1,t-1}$, are then calculated, dropping any observations with missing values. We then divide changes in (jack-knifed) mean teacher VA in the prior grade last year into twenty equal sized bins (vingtiles) and plot the mean changes in the teacher-year fixed-effect in each bin against mean changes in (jack-knifed) teacher VA in the prior grade last year. The solid line shows the OLS line of best fit which corresponds to running regression equation (1). Panel B is constructed analogously with the exception that VA estimates are five-year leave out estimates – namely, every year we exclude the two years prior to the event, the two event years, and the year after the event. Each panel reports the slope coefficient and corresponding standard error from the underlying OLS, with standard errors clustered at the teacher-level. Panels A and B contain 41,254 and 34,358 observations, respectively.

twenty equal sized bins (vingtiles) and we plot the mean changes in teacher-year fixed-effect in each bin against mean changes in (jack-knifed) teacher VA in the prior grade last year.

The figure indicates a clear negative relationship and so increases (decreases) in mean teacher VA in the prior grade last year cause a decline (increase) in the teacher-year fixed-effects. Given that the teacher-year fixed-effects are used to construct VA measures, this shows that the quality of teaching in the prior year impacts VA estimates. The slope coefficient that we estimate is statistically significant at the one percent level and has a magnitude of -0.13. Given that the standard deviation of school-grade-year mean teacher VA is 0.10 and the standard deviation of teacher-year fixed-effects is 0.25, this means that a one-standard deviation increase in mean teacher VA in the prior grade last year leads to a 0.052 SD fall in teacher-year fixed-effects in the current year.

Therefore we have shown that $b < 0$. Intuitively, this arises because we do not control for fade out when forming students' residual test scores and so a positive change in the teaching quality in the prior grade last year (i.e., $\Delta Q^{jt}_{s,g-1,t-1} > 0$) leads to an increase in the amount of fade out experienced across cohorts resulting in a negative change in residual test scores across cohorts (i.e., $\Delta X^{jt}_{sgt} < 0$). Therefore, $cov(\Delta Q^{jt}_{s,g-1,t-1},\ \Delta\chi^{jt}_{sgt}) < 0$, resulting in $b < 0$.

We note that our finding that $b < 0$ cannot be explained by unobservable persistent school shocks, as these would induce a positive slope in regression equation (1) (Chetty et al., 2017).[34] To further alleviate any remaining concerns, we repeat the analysis excluding a full five years from teachers' VA calculations: the two years prior to the event, the two event years, and the year after the event. Panel B of Figure 3 reports the results, which are virtually identical to those from Panel A. We now turn to accounting for fade out and showing that the teacher-level bias found in our quasi-experimental design is removed once fade out is accounted for.

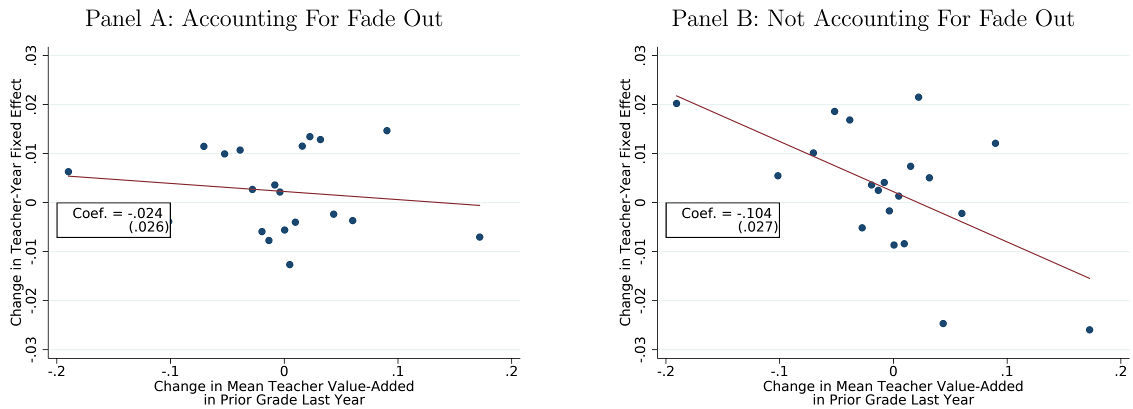
IV.4 Accounting for Fade Out in the Quasi-Experimental Design

We now re-estimate the above quasi-experiment except that we control for fade out in our calculation of VA measures. To do so, we add a parsimonious set of variables designed to control for fade out in our control vector X_{ijt} in addition to the original set of controls. (See Section III.2 for more detail). To control for fade out we include cubic polynomials of mathematics and English student test score growth in the prior grade last year, interacted with grade indicators. We also include cubic polynomials of mean (non-missing) mathematics and English prior-year student growth at the class and school-year level, interacted with grade indicators. For students missing prior year-growth data we impute a value of zero and include missing indicators. To ensure that we adequately control for fade out, we only keep classes which have prior year mathematics growth data for at least 50% of students.

Panel B of Figure 4 replicates Panel A of Figure 3 among the restricted 'Fade Out VA Sample' that requires twice-lagged test scores. As expected, the sample restriction has little effect on the negative relationship between changes in mean teacher VA in the prior grade last year and changes in teacher-year fixed-effects. In Panel A, we control for

[34]For example, suppose we are looking at teacher j who taught grade 5 in 2005-06 and 2006-07. Suppose, there is a persistent AR(1) positive school shock in 2003-04. Since Panel A of Figure 3 includes data from 2003-04 in the calculation of $\Delta Q^{jt}_{s,g-1,t-1}$ (it only excludes data from 2004-05, 2005-06, and 2006-07), the shock will raise the VA of existing teachers in the school in 2003-04. The positive shock implies that the third grade 2003-04 cohort will have, in expectation, higher growth in fifth grade in 2005-06 than the 2006-07 fifth grade cohort (who was in second grade in 2003-04) and so there will be a decline in fifth grade test score growth from 2005-06 to 2006-07. If an existing teacher (whose VA estimate includes this positive school shock) then leaves the school in fourth grade in 2005-06 there will also be, in expectation, a negative change in the (jack-knifed) VA estimate in the prior grade last year across the 2005-06 and 2006-07 fifth grade cohorts, yielding $b > 0$. An analogous story occurs in the case of a negative school shock. Indeed, Chetty et al. (2017) argue that these school shocks could be an explanatory factor for why Rothstein (2017) finds a positive correlation between changes in average prior year scores and changes in teacher VA (refer to his Figure 2).

Figure 4: Impact of Teacher VA in Prior Grade Last Year on Teacher-Year Fixed-Effect That Accounts for Fade Out



Notes: Panel A reproduces Figure 3, but accounts for fade out in the estimation of VA. We account for fade out in Panel A by including cubic polynomials of mathematics and English student test score growth in the prior grade last year, interacted with grade indicators. We also include cubic polynomials of mean (non-missing) mathematics and English prior-year student growth at the class and school-year level, interacted with grade indicators. For students missing prior year-growth data we impute a value of zero and include missing indicators. To ensure that we are adequately controlling for fade out, we restrict our sample in Panel A to only include classes where less than 50% of students have missing prior year mathematics growth data. Panel B then replicates Panel A of Figure 3 (that does not account for fade out) among this restricted sample. Each panel reports the slope coefficient and corresponding standard error from the underlying OLS, with standard errors clustered at the teacher-level. Each panel A contains 12,470 observations.

fade out in the estimation of VA and re-estimate the relationship. We find no significant relationship between changes in mean teacher VA in the prior grade last year and changes in teacher-year fixed-effects once fade out is accounted for. Together these results show that accounting for fade out removes the teacher-level biasedness present in the quasi-experimental design.

Forecast Unbiasedness: Recall that our model in Section II raised the possibility that failing to account for fade out leads to teacher-level bias, but does not create forecast bias when students are sorted randomly to teachers (conditional on controls). We now show that forecast unbiasedness holds regardless of whether fade out is accounted for, aligning with these theoretical predictions.

To show that forecast unbiasedness is maintained regardless of whether fade out is accounted for, we use our fade out sample and implement the quasi-experimental design of Chetty et al. (2014a) to assess forecast biasedness.[35] Let Q_{sgt} to be the student-weighted mean VA of $\hat{\mu}_{j,t}^{-\{t,\ t-1\}}$ across teachers in school s in grade g in year t. Define $\Delta Q_{sgt} \equiv Q_{sgt} - Q_{sg,t-1}$ as the change in mean teacher VA in school s in grade g between years $t-1$ and t. Let A^*_{sgt} be the mean of student test scores in school s in grade g in year t and define $\Delta A^*_{sgt} \equiv A^*_{sgt} - A^*_{s,g,t-1}$ as the change in those test scores in school s in grade g

[35] i.e., we replicate Table 4 in Chetty et al. (2014a).

Table 2: Quasi-Experimental Estimates of Forecast Bias

	Accounting for Fade Out		Without Accounting for Fade Out	
	Δ score (1)	Δ score (2)	Δ score (3)	Δ score (4)
Change in mean teacher VA across cohorts	1.010 (0.034)	0.985 (0.036)	1.005 (0.035)	0.979 (0.037)
R-squared	6.74%	9.54%	6.59%	9.39%
Year Fixed Effects	X	X	X	X
School Fixed Effects		X		X
Grades	4 to 5	4 to 5	4 to 5	4 to 5
Number of School-Grade-Year Cells	19,951	19,951	19,951	19,951

Notes: This table reports the coefficient b from equation (3) using the Fade Out VA Sample where we cluster standard errors at the school-cohort level. To construct column (1), we include the aforementioned set of variables that control for fade out and compute two-year leave-out VA estimates for each teacher. Classrooms with missing two-year leave-out VA estimates are excluded. We then compute mean (student-weighted) teacher VA and test scores at the school-grade-year level. Finally, we compute the change in test scores at the school-grade level across years t and $t-1$ to obtain ΔA^*_{sgt}. Analogously, we compute the difference in mean teacher VA at the school-grade level across years t and $t-1$ to obtain ΔQ_{sgt}. We regress ΔA^*_{sgt} on ΔQ_{sgt}, including year fixed-effects to eliminate secular trends. We cluster standard-errors at the school-grade level. Column (2) is constructed analogously except we now include both year fixed-effects and school fixed-effects in the regression. To construct columns (3) and (4), we follow the same steps used to construct columns (1) and (2), except we do not include controls for fade out when computing two-year leave-out VA estimates.

between years $t-1$ and t. We then run the following regression:

$$\Delta A^*_{sgt} = a + b\Delta Q_{sgt} + \Delta\chi_{sgt} \tag{3}$$

Chetty et al. (2014a) show that the coefficient b in (3) identifies the degree of forecast bias (defined in Definition 1) under the identification assumption that $cov(\Delta Q_{sgt}, \Delta\chi_{sgt}) = 0$ which they argue holds in their data.[36] In particular, a point estimate where $b = 1$ indicates forecast unbiasedness.[37]

Column (1) in Table 2 shows the results of equation (3) when we account for fade out in the estimation of VA, while column (3) displays the results when we do not control for fade out when computing VA estimates. Both cases cannot reject forecast unbiasedness in VA estimates as our estimates of b are statistically indistinguishable from one. We continue to be unable to reject forecast unbiasedness when we repeat the analysis controlling for both year and school fixed-effects in columns (2) and (4). We also note that accounting

[36] In particular, they show that the exclusion of salient observable characteristics, including parent characteristics and twice-lagged test scores, does not induce any meaningful forecast bias in their results.

[37] Recall that Definition 1 defines forecast unbiasedness as when the relationship between mean (residual) test scores of students assigned to teacher j in year t and the corresponding VA estimates $\hat{\mu}_{jt}$ are one-to-one.

for fade out increases the forecast precision of teacher quality, as shown by the higher R-squared in columns (1) and (2) relative to columns (3) and (4).

These results combined with our prior findings indicate that teacher VA estimates that do not account for fade out are forecast unbiased, but feature bias at the teacher-level. These results align with findings in the prior literature that the sorting of students to teachers is limited conditional on controls for lagged test scores and that VA estimates are forecast unbiased (Chetty et al., 2014a; Bacher-Hicks et al., 2014). Nevertheless, we find that accounting for fade out is necessary to eliminate bias for those teachers who were unlucky (lucky) to receive students who had teachers in the top (bottom) of the VA distribution in the prior year.

Furthermore, our results shed new light on how researchers can find forecast unbiased VA estimates side-by-side with biased teacher-level VA estimates. For example, our results help reconcile the concerns highlighted in Rothstein (2010) with the forecast unbiasedness results in Chetty et al. (2014a). Additionally, they help explain how Kane et al. (2013) – in the largest RCT to date on quantifying unbiasedness in teacher VA estimates – find that, as a group, students' test score gains were largely in line with teacher VA estimates, but there was a portion of low- and high-VA teachers whose students' test score gains were considerably better or worse than their VA estimates.

V Policy Implications

This section considers the policy implications of teacher-level bias arising from fade out. We start by considering the implications for *who* is fired, then consider gains in policy effectiveness from accounting for fade out. We pay particular attention to a benchmark policy of releasing teachers in the bottom five percent of the VA distribution, given its prominence in the literature.

Who is Released: We start by looking at how accounting for fade out affects who is released. To do so, we contrast our VA estimates that do and do not account for fade out. We find that many teachers are released under one VA measure, but not the other. Specifically, our VA sample of fourth and fifth grade teachers consists of 14,289 teachers and so a policy that deselects teachers in the bottom five percent of the VA distribution releases 715 teachers. We find that 106 of those 715 teachers are released when VA does not account for fade out, but are *not* released when fade out is accounted for (and vice versa). Therefore, the termination decision for fifteen percent of released teachers relies on whether fade out is properly accounted for.

Since accounting for fade out affects who is released, the quality of released teachers is lower when fade out is accounted for which, in turn, will drive higher policy gains. Given that eighty-five percent of teachers are released regardless of whether fade out is accounted for, we focus the discussion that follows on the improved policy gains among the fifteen percent of teachers whose termination decision depends upon whether fade out is properly accounted for.[38]

Impacts on Test Scores: Let μ^{fade} denote teacher VA that accounts for fade out and $\mu^{no\ fade}$ teacher VA that does not account for fade out. Under the assumption that underlying teacher quality is normally distributed,[39] classrooms with teachers that are released under the policy, on average, will see test scores increase by:

$$G = \mathbb{E}[\mu^{fade} | \mu^{fade} < \Phi^{-1}(0.05)] \,, \tag{1}$$

where $\Phi(\cdot)$ is the cdf of the normal distribution with variance σ_μ^2. When teacher VA does not account for fade out then the policymaker bases her decision on $\mu^{no\ fade}$ instead. Policy gains among affected classrooms in this case will be:

$$G = \mathbb{E}[\mu^{fade} | \mu^{no\ fade} < \Phi^{-1}(0.05)] \,. \tag{2}$$

Among the fifteen percent of classrooms whose teachers' release status are dependent on the VA measure used, the extra policy gain achieved by accounting for fade out in terms of test scores is:

$$\begin{aligned} MG = \; & \mathbb{E}[\mu^{fade} | \mu^{fade} < \Phi^{-1}(0.05), \mu^{no\ fade} \geq \Phi^{-1}(0.05)] \\ & - \mathbb{E}[\mu^{fade} | \mu^{fade} \geq \Phi^{-1}(0.05), \mu^{no\ fade} < \Phi^{-1}(0.05)] \,. \end{aligned} \tag{3}$$

Given the variance of the VA distribution that we estimate ($\sigma_\mu = 0.173$),[40] this implies that test scores will rise by 0.042 standard deviations in classrooms whose teachers' retention status depends on the VA measure being used.

[38] Since the other eighty-five percent of teachers do not affect policy gains as they are released regardless of whether fade out is accounted for, the gains in terms of total policy can be determined by multiplying our gains by fifteen percent.

[39] The assumption that underlying teacher quality is normally distributed can be relaxed using the nonparametric empirical Bayes estimator introduced by Gilraine et al. (2020). Those authors find, however, that teacher quality is close to normally distributed in North Carolina and so maintaining normality causes little bias in this context.

[40] Accounting for fade out also affects the estimated variance of teacher effects. For brevity, we treat the variances when fade out is and is not accounted for as the same since the estimated variances are near-identical (see Table B.2).

These policy calculations assume that the policymaker knows true teacher VA and thus makes retention decisions based on teachers' true quality. In reality, teacher VA is estimated. To account for this, we replace the test score gains in equation (3) with their sample analogs.[41] We calculate these sample analogs via Monte Carlo simulation assuming that we have three years of data for each teacher and all teachers have class sizes of twenty students. When retention decisions are formulated based on estimated VA, the test score gains of the policy among affected classrooms become 0.041 standard deviations.

Impacts on Long-Run Outcomes: We next evaluate the additional policy gains in terms of long-run outcomes that can be achieved by releasing teachers based on VA that accounts for fade out. Given our data, we consider several important long-run outcomes: algebra scores, drop-out rates, suspensions, PSAT scores, SAT taking , and SAT scores. Once again we focus on the benchmark policy whereby teachers in the bottom five percent of the VA distribution are released and are replaced by mean quality teachers.

We link long-run outcomes with teacher VA using the method proposed by Chetty et al. (2014b). First, the benefits of a one-unit increase in VA is calculated. We start by constructing long-term outcome residuals using variation across students taught by the same teacher j, based on the regression equation

$$Y_{ij}^{*} = \alpha_j + \beta^{Y} X_{ijt} + u_{ijt}\,, \tag{4}$$

where Y_{ij}^{*} is the long-run outcome of interest, α_j is a teacher fixed effect, and X_{ijt} are observed determinants of student achievement, including lagged test scores. We then define the long-run residuals as

$$Y_{ijt} = Y_{ij}^{*} - \hat{\beta}^{Y} X_{ijt}\,, \tag{5}$$

where $\hat{\beta}^{Y}$ is estimated in equation (4).

These long-run residuals, Y_{ijt}, are then regressed on each teacher's (normalized) VA, pooling across grades:

$$Y_{ijt} = \delta + \kappa \hat{m}_{jt} + \eta_{ijt}\,, \tag{6}$$

where $\hat{m}_{jt} \equiv \frac{\hat{\mu}_{jt}^{fade}}{\hat{\sigma}_{\mu}}$ is normalized teacher VA, which is just teacher VA (that accounts for

[41] So equation (3) becomes $\mathbb{E}[\mu^{fade}|\hat{\mu}^{fade} < \hat{\Phi}^{-1}(0.05), \hat{\mu}^{no\ fade} \geq \hat{\Phi}^{-1}(0.05)] - \mathbb{E}[\mu^{fade}|\hat{\mu}^{fade} \geq \hat{\Phi}^{-1}(0.05), \hat{\mu}^{no\ fade} < \hat{\Phi}^{-1}(0.05)]$.

fade out) scaled by the estimated standard deviation ($\hat{\sigma}_\mu$) of the teacher VA distribution.[42]

Figure B.1 shows the impact of be being assigned a teacher with higher VA for one year on long-run outcomes. The panels plot residual long-run outcomes for students in school year t versus the estimated (normalized) VA that accounts for fade out of the teacher j who taught them in that year, $\hat{m}_{jt}$. We build the binned scatter plots in three steps: (i) residualize the long-run outcome as described in equation (5), (ii) divide our VA estimates, $\hat{m}_{jt}$, into twenty equal-sized bins and plot the mean of the long-run outcome residuals in each bin against the corresponding bin mean of $\hat{m}_{jt}$, and (iii) add back in the mean long-run outcome in the estimation sample to facilitate the interpretation of the scale.

The slope coefficients suggest that being assigned to a teacher whose test score VA is one standard deviation higher in a single grade increases algebra scores by 0.02 standard deviations, reduces dropout by 0.23 percentage points, lowers the number of days suspended in middle and high school by 0.03, raises PSAT scores by 2 points, increases the likelihood of taking the SAT by 0.8 percentage points, and boosts SAT scores by 3 points. All these impacts are statistically significant.

Long-Run Effects of Releasing Teachers in the Bottom 5 Percent: The relationship between teacher VA and long-run outcomes feeds into our calculations of the long-run gains of our benchmark policy. Specifically, among teachers whose release status depends on the VA measure used, accounting for fade out in VA estimates raises student outcomes by:

$$G = (\Delta m_\sigma^{fade} - \Delta m_\sigma^{no\ fade}) \times \kappa\,, \tag{7}$$

where Δm_σ^{fade} represents the average increase in VA among those teachers released when fade out is accounted and $\Delta m_\sigma^{no\ fade}$ when it is not.

Table 3 reports the additional policy gains coming from accounting for fade out in terms of the long-run outcomes. We find significant improvements in the long-run when teachers are released based on VA that accounts for fade out: We find that VA is one-quarter of a standard deviation higher among affected classrooms when VA is accounted for, leading to roughly a fifteen percent gain in long-term outcomes. For instance, we find that accounting for fade out leads to a roughly one point increase in SAT scores among

[42]Due to Bayes shrinkage the standard deviation of the normalized teacher VA measure is less than one as in Chetty et al. (2014b).

Table 3: Long-Run Policy Benefits from Accounting for Fade Out
when Releasing Bottom 5% of Teachers

Long-Run Outcome:	Algebra Score (σ) (1)	Percent Drop Out (2)	Days Suspended (3)	PSAT Score (4)	Percent Took SAT (5)	SAT Score (6)
Sample Mean	0.054	8.76	3.03	1297.3	35.5	999.7
Benefit ($\hat{\kappa}$)	0.019	-0.23	-0.028	1.85	0.81	3.34
Change in VA when True VA is Observed:						
VA Accounts for Fade Out (Δm_{σ}^{fade})	1.76	1.76	1.76	1.76	1.76	1.76
VA Does not Account for Fade Out ($\Delta m_{\sigma}^{no\ fade}$)	1.51	1.51	1.51	1.51	1.51	1.51
Long-Run Gain of Accounting for Fade Out (G)	0.005	-0.058	-0.007	0.46	0.20	0.84

Notes: This table shows the estimated gains in terms of various long-run outcomes of accounting for fade out among teachers whose retention decision depends on whether VA accounts for fade out under a policy that releases the bottom 5% of teachers and replaces them with mean quality teachers when true teacher quality. The results assume that the policymaker can observe the actual VA of the teacher; Table B.3 reports the additional policy gains when true teacher quality is unobserved to the policymaker and so teacher releases are based on estimated (rather than true) value-added. Long-run outcomes are: algebra scores (standardized by year taken), high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores. For PSAT and SAT scores, we combine the math and English components and take the values from the student's first attempt. See Table B.1 for more details on the coverage of long-run outcomes. 'Benefit' represents the increase in the long-run outcome associated with having a teacher with VA one standard deviation higher for one grade as described by equation (6); this benefit is shown graphically in Figure B.1. The third and fourth rows then calculate the increase in VA when affected teachers are replaced by mean quality teachers when VA is accounted for (third row) and not accounted for (fourth row). The difference between these two rows is then calculated and then multiplied by the benefit to give the additional policy gains of accounting for fade out among teachers who retention decision depends on whether VA accounts for fade out, as described by equation (7).

affected classrooms.[43]

To gauge the magnitude of these additional policy gains, we place them in terms of the additional earnings findings in Chetty et al. (2014b): They find that replacing a teacher whose VA is in the bottom 5% with an average teacher increases the present value of students' lifetime income by approximately $400,000 per classroom. Given that accounting for fade-out increases VA among affected classrooms by 0.25 standard deviations, this implies that accounting for fade out could increase the net present value of earnings in affected classrooms by $45,000.[44] It is therefore clear that accounting for fade out in VA estimates allows policymakers to both release teachers more *fairly* by correcting teacher-level bias *and* increase the total gains achieved by a given policy in terms of both test

[43] The policy gains are computed in the case where the policymaker observes true VA. Table B.3 reports the additional policy gains when teacher VA is estimated (calculated via Monte Carlo). Results are similar when when teacher VA is estimated, although policy gains are somewhat lower as the policymaker releases some teachers whose true VA is not in the bottom five percent.

[44] Calculation: $522,000*1.34%*0.24, where the first two numbers are from Chetty et al. (2014b) and 0.24 is the difference in VA among affected classrooms when VA is estimated (see Table B.3).

scores and long-run outcomes.

VI Conclusion

We have shown that fade out biases value-added estimates at the teacher-level. Using quasi-experimental variation that exploits school-level teacher turnover for identification, we show that a one-standard deviation increase in mean teacher VA in the prior grade last year causes a 0.05σ decrease in a teacher's fixed-effect in the current year. We find that the dependence of teacher VA on the quality of the prior teacher is eliminated once fade out is controlled for in the VA estimation.

Accounting for fade out both increases the fairness and efficiency of VA policies. In particular, under the widely-discussed policy to release teachers in the bottom 5% of the VA distribution, roughly 15% of teachers are wrongly fired when fade out is not accounted for in VA estimation. Furthermore, accounting for fade out raises the benefits of the policy both in terms of test scores and long-run outcomes, potentially raising the net present value of earnings among affected classrooms by $45,000.

Our results shed new light on the importance of incorporating dynamic features of the education production function into empirical specifications. In particular, much of the education literature controls for the full history of prior inputs using lagged scores. In the context of teacher VA, we find that lagged scores are an insufficient control due to the fade out of test scores that has been documented in the literature. To account for fade out, both the level and change in performance must be accounted for as two students could perform at the same level, but have experienced different knowledge gains in the prior year, thereby experiencing different fade out levels in the subsequent year.

More generally, the presence of fade out naturally raises a distinction between short- and long-term knowledge in the education production function. Given that the learning process is inherently cumulative, better-targeting pieces of knowledge that persist can lead to large gains in students' total knowledge over time. Along these lines, recent research has highlighted the importance of non-cognitive skills which have been found to be more enduring than test score gains (e.g., Cunha et al., 2010), with several recent papers estimating teachers' non-cognitive VA (e.g., Jackson, 2018; Petek and Pope, 2018). An alternative approach is to decompose teachers' impacts on test scores into short- and long-term components, allowing policymakers to target teachers' contribution to long-term knowledge directly – something we are exploring in related work.

References

Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger (2014), "Validating teacher effect estimates using changes in teacher assignments in Los Angeles." Working Paper 20657, National Bureau of Economic Research, URL http://www.nber.org/papers/w20657.

Biasi, Barbara (forthcoming), "The labor market for teachers under different pay schemes." *American Economic Journal: Economic Policy*.

Cascio, Elizabeth U. and Douglas O. Staiger (2012), "Knowledge, tests, and fadeout in educational interventions." Working Paper 18038, National Bureau of Economic Research, URL http://www.nber.org/papers/w18038.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2011), "How does your kindergarten classroom affect your earnings? Evidence from Project STAR." *Quarterly Journal of Economics*, 126, 1593–1660.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014a), "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review*, 104, 2593–2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2014b), "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood." *American Economic Review*, 104, 2633–79.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff (2017), "Measuring the impacts of teachers: Reply." *American Economic Review*, 107, 1685–1717.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006), "Teacher-student matching and the assessment of teacher effectiveness." *Journal of Human Resources*, 41, 778–820.

Cunha, Flavio and James Heckman (2007), "The technology of skill formation." *American Economic Review*, 97, 31–47.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach (2010), "Estimating the technology of cognitive and noncognitive skill formation." *Econometrica*, 78, 883–931.

Currie, Janet and Duncan Thomas (1995), "Does Head Start make a difference?" *American Economic Review*, 85, 341–364.

Gershenson, Seth and Michael S. Hayes (2018), "The implications of summer learning loss for value-added estimates of teacher effectiveness." *Educational Policy*, 32, 55–85.

Gilraine, Michael, Jiaying Gu, and Robert McMillan (2020), "A new method for estimating teacher value-added." Working Paper 27094, National Bureau of Economic Research, URL http://www.nber.org/papers/w27094.

Goldhaber, Dan and Duncan Dunbar Chaplin (2015), "Assessing the"Rothstein falsification test": Does it really show teacher value-added models are biased?" *Journal of Research on Educational Effectiveness*, 8, 8–34.

Hanushek, Eric A. (1971), "Teacher characteristics and gains in student achievement: Estimation using micro data." *American Economic Review*, 61, 280–288.

Hanushek, Eric A. (1986), "The economics of schooling: Production and efficiency in public schools." *Journal of Economic Literature*, 24, 1141–1177.

Hanushek, Eric A. (2009), "Teacher deselection." In *Creating a New Teaching Profession* (Dan Goldhaber and Jane Hannaway, eds.), 165–180, Urban Institute Press, Washington, DC.

Hanushek, Eric A. (2011), "The economic value of higher teacher quality." *Economics of Education Review*, 30, 466–479.

Hanushek, Eric A. and Steven G. Rivkin (2010), "Generalizations about using value-added measures of teacher quality." *American Economic Review, Papers & Proceedings*, 100, 267–71.

Jackson, C. Kirabo (2018), "What do test scores miss? The importance of teacher effects on non-test score outcomes." *Journal of Political Economy*, 126, 2072–2107.

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, 85–108.

Jacob, Brian A., Lars Lefgren, and David P. Sims (2010), "The persistence of teacher-induced learning." *Journal of Human Resources*, 45, 915–943.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas Staiger (2013), "Have we identified effective teachers? Validating measures of effective teaching using random assignment." *MET Project Research Paper, Bill & Melinda Gates Foundation.*

Kane, Thomas J. and Douglas O. Staiger (2008), "Estimating teacher impacts on student achievement: An experimental evaluation." Working Paper 14607, National Bureau of Economic Research, URL http://www.nber.org/papers/w14607.

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, 18–42.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff (2015), "Value-added modeling: A review." *Economics of Education Review*, 47, 180–195.

Krueger, Alan B. and Diane M. Whitmore (2001), "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR." *Economic Journal*, 111, 1–28.

Macartney, Hugh, Robert McMillan, and Uros Petronijevic (2018), "Teacher value-added and economic agency." Working Paper 24747, National Bureau of Economic Research, URL http://www.nber.org/papers/w24747.

McCaffrey, Daniel F., J.R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton (2004), "Models for value-added modeling of teacher effects." *Journal of Educational and Behavioral Statistics*, 29, 67–101.

McEachin, Andrew and Allison Atteberry (2017), "The impact of summer learning loss on measures of school performance." *Education Finance and Policy*, 12, 468–491.

North Carolina Education Research Data Center (1996-2017), "Student, class and personnel files." URL http://childandfamilypolicy.duke.edu/research/hc-education-data-center/.

Petek, Nathan and Nolan Pope (2018), "The multidimensional impact of teachers on students." URL http://www.econweb.umd.edu/~pope/Nolan_Pope_JMP.pdf. Unpublished.

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*, 94, 247–252.

Rothstein, Jesse (2010), "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics*, 125, 175–214.

Rothstein, Jesse (2017), "Measuring the impacts of teachers: Comment." *American Economic Review*, 107, 1656–84.

Stepner, Michael (2013), "Vam: Stata module to compute teacher value-added measures." URL http://fmwww.bc.edu/RePEc/bocode/v/vam.ado.

For Online Publication

A Construction of the Teacher Value-Added Sample

Our data cover fourth and fifth grade from 1996-97 through 2010-11 and third grade from 1996-97 through 2004-05 and 2006-07 through 2008-09. Our VA sample is restricted to those grades and years to ensure that we have both students' lagged test scores[45] and that we can accurately match students to their teachers.[46] These data cover 4,457,812 student-year observations with 1,843,217 students and 74,986 teachers.

To be included in our value-added analysis data set we must be able to match students to their teachers. We start by requiring that students are classified as being in a single grade and school throughout the school year, dropping about 100,000 observations. We also drop 16,000 observations where we lack data on teacher experience. Our matching procedure then follows Clotfelter et al. (2006) and subsequent research using North Carolina data and links student to teachers using the end-of-grade (EOG) files. These files record the test proctor during the end-of-grade tests, which for third through fifth grade is typically the teacher who taught the students throughout the year. We follow the sample restrictions used by prior researchers to ensure high-quality student-teacher matches and only count a student-teacher match as valid if the test proctor in the EOG files taught a self-contained class for the relevant grade in that year and that teacher only administered tests for students in that grade. We drop roughly 1.35 million student-year observations from our sample that we cannot accurately match to a teacher.

We then make several additional data restrictions. First, we drop 200,000 students who are missing lagged mathematics scores. Second, we only include classes with at least ten but fewer than forty students, dropping 125,000 observations. Third, we drop classes where at least fifty percent of students are classified as disabled, eliminating 3,000 observations. Our final VA sample consists of 2,675,254 student-year observations covering 1,385,237 students and 34,953 teachers.

For demographics, we have information about parental education (six education groups, 1996-97 through 2005-06 only), economically disadvantaged status (1998-99 through 2010-11 only), ethnicity (six ethnic groups), gender, limited English status, disability status,

[45] The third grade pre-test which we use as the lagged test score for third grade students is not available in 2005-06 and stopped being administered after 2008-09, necessitating the elimination of those years from our data for third grade. Third grade lagged test scores for English are also unavailable in 2005-06 and 2007-08.

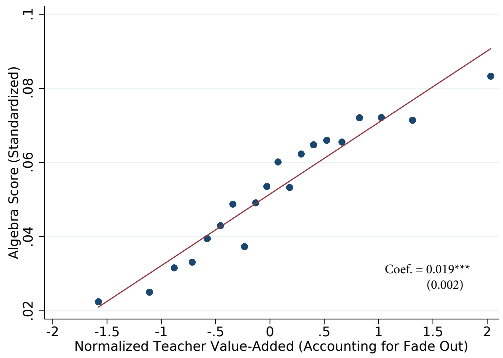
[46] We lose the ability to match students to teachers after 2010-11 as the teacher identifier is no longer recorded in the end-of-grade (EOG) files.

academically gifted status, and grade repetition. Besides the missing data in some years for parental education and economically disadvantaged status our demographic data cover over 99 percent of all student-year observations. Whenever demographic information is missing, we create a missing indicator for that variable.

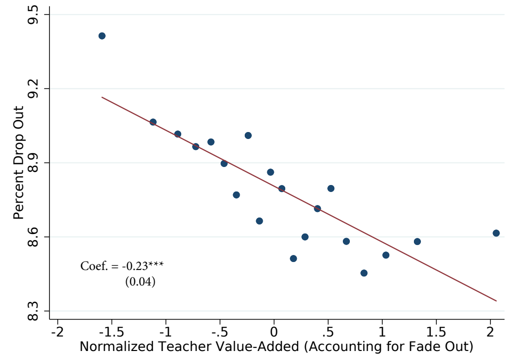
B Appendix Figures and Tables

Figure B.1: Effects of Teacher Value-Added on Long-Run Outcomes

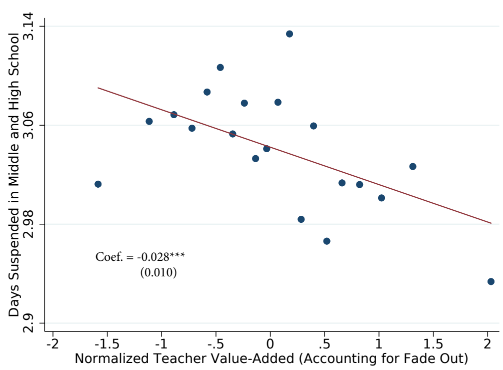
(a) Algebra Score



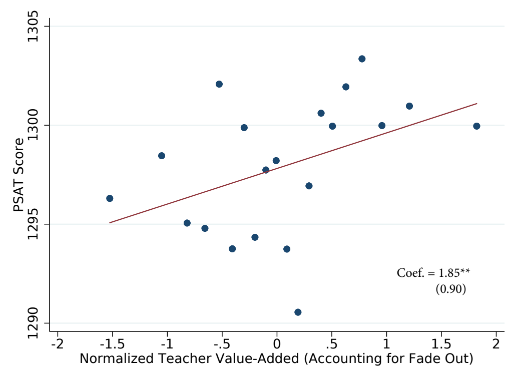
(b) High School Drop Out



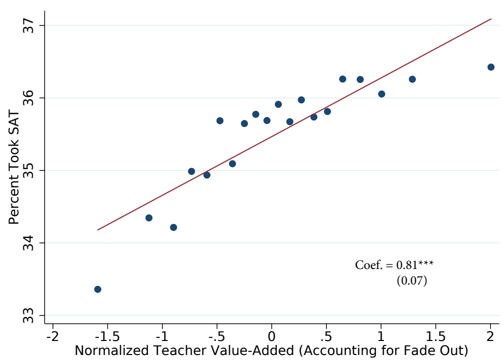
(c) Days Suspended



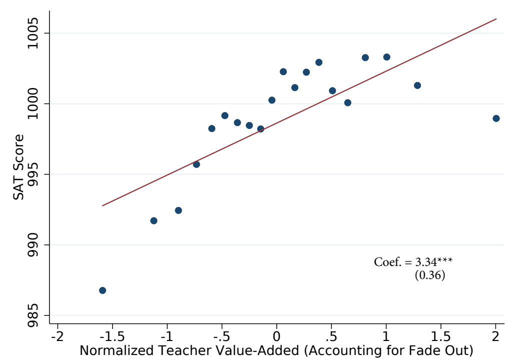
(d) PSAT Score



(e) Took SAT



(f) SAT Score



Notes: These figures link long-run outcomes to teacher value-added that accounts for fade out by comparing the long-run outcomes of students who were assigned to teachers with different value-added, controlling for a rich set of student characteristics. To do so, we follow the steps outlined in Chetty et al. (2014b): (i) residualize the long-run outcome as described in equation (5), (ii) divide our value-added estimates $\hat{m}_{jt}$ into twenty equal sized vingtiles and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}_{jt}$ in each bin, and (iii) add back in the mean long-run outcome in the estimation sample to facilitate interpretation of the scale. Coefficient estimates are reported in the figures with standard errors clustered at the school by cohort level in parenthesis below. Coefficient estimates are the same as those reported for κ in the second row of Table 3. Long-run outcomes are: algebra scores (standardized by year taken), high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores. For PSAT and SAT scores we combine the math and English components and take the values from the student's first attempt. See Table B.1 for more details on the coverage of long-run outcomes.

Table B.1: Coverage of Long-Run Data Linkage

	Data Coverage (1)	Grades Usually Taken (2)	Fade Out VA Sample Cohorts Covered[a] (3)	Match Rate (% of covered sample) (4)
Algebra I	1997-98 to 2016-17	Grades 7-9	All cohorts	81.4% (1,170,123 of 1,437,360)
High School Dropout	2003-04 to 2016-17	Grade 10-12	Entering 4^{th} grade after 1996-97 or before 2009-10	N/A[b]
PSAT	2012-13 to 2016-17	Grade 10	Entering 4^{th} grade after 2006-07	44.4% (148,383 of 334,191)
SAT	2008-09 to 2016-17	Grades 11-12	Entering 4^{th} grade after 1999-00 or before 2010-11	35.6% (330,751 of 930,474)
Days Suspended	2000-01 to 2016-17	Grades 6-12	All cohorts	N/A[b]

[a] Our fade out VA sample cover fourth grade cohorts from 1997-98 through 2005-06 and 2007-08 through 2009-10 and fifth grade cohorts from 1997-98 through 2010-11. The column then refers to which of these cohorts are covered by the long-run data. For example, for high school dropout "entering 4^{th} grade after 1996-97 or before 2009-10" mean that the following cohorts are linked to the dropout data: fourth grade students 1997-98 through 2008-09 (excluding 2006-07), and fifth grade students 1998-99 through 2009-10.

[b] If we do not observe the student to have dropped out or suspended according to the data students are assumed to not have dropped out or been suspended. Note that for high school dropout we only code a student as dropping out if their withdrawal code indicates that they are dropping out of high school (out-of-state or private school transfers are thus not coded as a drop out). We use dropout rather than graduation as the high school graduate data does not start until the 2008-09 school year.

Table B.2: VA Autocorrelation and Variance Estimates

	(1)	(2)	(3)
Autocorrelation Vector			
Lag 1	0.55	0.58	0.56
Lag 2	0.49	0.50	0.50
Lag 3	0.45	0.45	0.45
Lag 4	0.42	0.42	0.41
Lag 5	0.40	0.41	0.40
Lag 6	0.38	0.38	0.38
Lag ≥ 7	0.36	0.37	0.36
Within-year variance components			
Total SD	0.543	0.515	0.517
Individual-level SD	0.497	0.472	0.472
Class + teacher-level SD	0.217	0.207	0.211
Estimates of teacher SD			
Lower bound based on lag 1	0.178	0.173	0.173
Quadratic estimate	0.188	0.186	0.187
Data Used			
VA Sample	X		
Fade Out VA Sample		X	X
Variables Used			
VA Controls	X	X	X
Fade Out Controls		X	
Student-Year Observations	2,675,254	1,437,360	1,437,360

Notes: This table gives the drift autocorrelation estimates across years for the same teacher used to compute VA estimates. It also reports the raw standard deviation of test score residuals and decompose this variation into components driven by idiosyncratic student-level and class+teacher variation. The sum of the student-level and class+teacher variances equals the total variance. These estimates are outputs of the vam.ado file constructed by Stepner (2013). To obtain estimates of teacher SD we replicate the procedure used by Chetty et al. (2014a) – refer to their Table 2 notes.

Table B.3: Long-Run Policy Benefits from Accounting for Fade Out when Releasing Bottom 5% of Teachers

Long-Run Outcome:	Algebra Score (σ) (1)	Percent Drop Out (2)	Days Suspended (3)	PSAT Score (4)	Percent Took SAT (5)	SAT Score (6)
Sample Mean	0.054	8.76	3.03	1297.3	35.5	999.7
Benefit ($\hat{\kappa}$)	0.019	-0.23	-0.028	1.85	0.81	3.34
Change in VA when True VA is Unobserved:						
VA Accounts for Fade Out ($\Delta\hat{m}_{\sigma}^{fade}$)	1.53	1.53	1.53	1.53	1.53	1.53
VA Does not Account for Fade Out ($\Delta\hat{m}_{\sigma}^{no\ fade}$)	1.29	1.29	1.29	1.29	1.29	1.29
Long-Run Gain of Accounting for Fade Out ($\hat{G}$)	0.005	-0.056	-0.007	0.45	0.20	0.82

Notes: This table is the analog to Table 3 when the policymaker cannot observe true teacher quality and so teacher releases are based on estimated (rather than true) VA. The table therefore shows the estimated gains in terms of various long-run outcomes of accounting for fade out among teachers whose retention decision depends on whether VA accounts for fade out under a policy that releases the bottom 5% of teachers and replaces them with mean quality teachers when true teacher quality is unobserved by the policymaker. Policy calculations are done via Monte Carlo simulation under the assumption that we have three years of data for each teacher and all teachers have class sizes of twenty. Long-run outcomes are: algebra scores (standardized by year taken), high school drop out, total days suspended in middle and high school, PSAT scores, whether student takes the SAT, and SAT scores. For PSAT and SAT scores, we combine the math and English components and take the values from the student's first attempt. See Table B.1 for more details on the coverage of long-run outcomes. 'Benefit' represents the increase in the long-run outcome associated with having a teacher with VA one standard deviation higher for one grade as described by equation (6); this benefit is shown graphically in Figure B.1. The third and fourth rows then calculate the increase in VA when affected teachers are replaced by mean quality teachers when VA is accounted for (third row) and not accounted for (fourth row). The difference between these two rows is then calculated and then multiplied by the benefit to give the additional policy gains of accounting for fade out among teachers who retention decision depends on whether VA accounts for fade out, as described by equation (7).