Making Teaching Last: Long- and Short-Run Value-Added*

Michael Gilraine, New York University
Nolan Pope, University of Maryland

December 3, 2020

ABSTRACT

Recent research indicates that teachers with high value-added (VA) improve their students' later life outcomes, leading to VA measures featuring ever more widely in high-stakes decision-making. Yet concerns remain that VA measures might capture undesirable dimensions of learning such as rote memorization that are unlikely to contribute to a student's lifelong success. This paper separates test score value-added measure into two components: long-run VA, which persists to the following year, and short-run VA, which does not. We find that students assigned to high long-run VA teachers fare substantially better in terms of long-run outcomes. In contrast, students assigned to high short-run VA teachers do not perform any better in terms of long-run outcomes. We also find that long-run VA is correlated with non-cognitive VA measures, whereas short-run VA is not. Policy simulations indicate that the use of long-run VA improves policy effectiveness by a factor of two over traditional VA measures.

Keywords: Value-Added

JEL Classifications: H40, I21, I22

*We would like to thank Jesse Bruhn along with seminar participants at University College Dublin and the Northeast Economics of Education Workshop for helpful comments. Gilraine: Department of Economics, New York University, New York, NY 10003 (email: mike.gilraine@nyu.edu); Pope: Department of Economics, University of Maryland, College Park, MD 20742 (email: npope@umd.edu). All remaining errors are our own.

1 Introduction

A main objective of educators and policymakers is to improve the long-run outcomes of their students, such as high school and college graduation, income, and health. A growing body of work has shown that high quality teachers are a key component to improving these later life outcomes (Chetty et al., 2014b). Over the last several decades, the predominant measure of teacher quality – teacher value-added (VA) – has focused on how teachers directly improve the contemporaneous test scores of students.

Many teachers, however, argue that the contemporaneous impact teachers have on students' test scores is a poor proxy for the permanent influence teachers have on their students' future lives. This is particularly true if teaching foundational principles of learning or developing behavioral improvements are not initially measured by tests, but are important for future learning. In addition, the incentivization of teachers' contributions to contemporaneous test scores has led to concerns that VA measures largely capture temporary improvements in students' academic achievement due to teaching practices such as rote memorization or 'teaching to the test' in common parlance (Koretz, 2002).

This paper separates a teachers' contribution to student performance into two components: 'long-run' VA, which persists to the following year, and 'short-run' VA which does not. To do so, we use a model of education production with short- and long-term knowledge as in Jacob et al. (2010) and Cascio and Staiger (2012) to show that teachers' contribution to next year's knowledge can be estimated by replacing contemporaneous test scores with subsequent test scores as the dependent variable in VA estimation. Short-run VA can then be calculated by using the difference between contemporaneous and subsequent year test scores as the dependent variable. Incorporating fixed effects for students' teachers in the next year then accounts for the influence of their teacher in the following year. Both components of VA can be estimated using standard VA approaches that have been developed in the literature (Rockoff, 2004; Kane and Staiger, 2008; Kane et al., 2008; Jacob and Lefgren, 2008; Chetty et al., 2014a).

We estimate long- and short-run VA using over a decade of data from third to fifth grade students in a large urban school district. We show that long- and short-run VA contribute approximately 40 and 60 percent, respectively, to standard teacher VA. While long-run and

short-run VA are positively correlated with the standard teacher VA measure (by construction), they are negatively correlated with each other. In addition, long-run VA is positively correlated with non-cognitive VA measures, whereas short-run VA is not. Given that non-cognitive VA measures are better predictors than traditional VA measures of long-term behavioral outcomes such as high-school completion, college-going, and crime (Jackson, 2018; Petek and Pope, 2018; Rose et al., 2019), this suggests that our long-run VA measure partially captures teachers' influence on student behavioral improvements that are crucial for future outcomes.

We next investigate whether our long-run VA measure better captures a teacher's contribution to students' lifelong success. To do so, we calculate the impact of being assigned to high long- and short-run VA teachers in elementary school on high school outcomes. We find no evidence that being assigned a higher short-run VA teacher improves a student's high school outcomes. In contrast, being assigned a higher long-run VA teacher significantly improves a host of high school outcomes, such as graduation, high school test scores, GPA, absences, and taking and performing well on the SAT. In addition, long-run VA is a far better predictor of these outcomes than either traditional or non-cognitive VA measures. While these results require the assumption that students are not sorted to teachers based on unobservable determinants of student achievement, we find no evidence of forecast bias or sorting on omitted observables.[1]

Following Chetty et al. (2014a) and Bacher-Hicks et al. (2014), we use quasi-experiments that exploit teacher turnover to show that our long-run VA measure affects contemporaneous and future student performance, while short-run VA only affects contemporaneous performance. First, we conduct event studies around the arrival or departure of a teacher in the bottom or top five percent of the VA distribution, finding that contemporaneous test scores change sharply across cohorts when high or low VA teachers enter or exit a school-grade. In contrast, test scores for those cohorts in the subsequent year change sharply with the entry or exit of a high or low long-run VA teacher, but not a high or low short-run VA teacher.

Second, we leverage variation from the entire distribution of teacher turnover and show that the change in contemporaneous school-grade mean test scores are statistically indistinguishable

[1]Whether teacher VA estimates are biased has been the subject of much debate in the literature (see Rothstein, 2010, 2017; Chetty et al., 2017). While we conduct the standard tests to test for bias in VA estimates, including our long- and short-run VA measures, we view our contribution as showing that teachers' contributions to later life outcomes come solely from the portion of VA that persists.

from the change in teacher long- or short-run VA caused by teacher staffing changes. As before, long-run VA affects the test scores of the students in the subsequent year, whereas short-run VA does not. Using similar quasi-experimental variation, we estimate the impact of long- and short-run VA teachers on high school outcomes and find that long-run VA positively affects these outcomes, while short-run VA does not. These findings provide direct evidence that while both the long- and short-run components of VA impact contemporaneous test scores, only the long-run component of VA affects future performance and lifelong success.

Our results demonstrate that more than half of the variation in the standard teacher VA measure has little to no predictive power on how a teacher impacts students' later academic outcomes. As such, the efficacy of many education policies could be improved by solely focusing on teachers' long-run VA. To demonstrate this point, we assess a benchmark policy in which teachers in the bottom 5 percent of the VA distribution are replaced with mean VA teachers. Because the correlation between standard VA and long-run VA is relatively low, over 70 percent of teachers released under a policy that employs standard VA are not released when long-run VA is adopted.

Using long-run VA leads to large improvements in policy efficiency. Releasing teachers based on true long-run VA instead of true standard VA increases the average beneficial effects of our benchmark policy on high school graduation, SAT-taking, SAT scores, PSAT scores, high school GPA, absences, and AP class attendance two- to three-fold. Unfortunately, policymakers can only use estimated VA measures for policy decisions which reduces the relative efficacy of using long-run relative to standard VA as long-run VA is noisier and features more drift. Even so, the policy gains from using long-run VA remain substantial: Releasing teachers based on long-run VA estimates after three years of data leads to an additional 0.2 students graduating and 1.4 students taking the SAT over the subsequent 10 years in comparison to releasing teachers based on standard VA. These are substantial improvements over the additional 1.0 student graduating and 0 students taking the SAT under the benchmark policy from using the standard VA measure. These benefits are achieved with no additional cost to schools, students, or teachers, and simply arise from using an alternative measure of teacher quality.

Our paper contributes to the literature in the followings ways. First, we demonstrate that more than half of the variation in standard teacher VA does not meaningfully predict students'

future academic outcomes. Second, we demonstrate that the component of the standard VA that predicts students' future outcomes can be easily estimated and used in lieu of standard VA measures for education policy. Third, because our standard measure of teacher quality is diluted by the noise of short-run VA, teachers are approximately twice as important at improving later-life outcomes than previously estimated. As such, previous estimates of the value of having a higher quality teacher are roughly half of the actual value of a higher quality teacher. Fourth, many education policies involving teacher quality could increase their efficacy two-fold if they replaced standard VA measures with long-run VA. Fifth, due to the negative correlation between long- and short-run VA, teachers may substitute time and energy between teaching methods that lead to long-term or short-term learning. Since we find no impact of short-run VA on future outcomes, teachers may be able to move away from teaching practices that lead to higher short-run VA to those that lead to higher long-run VA and thereby improve students' later life outcomes. Lastly, since long-run VA is positively correlated with non-cognitive VA measures and short-run VA is not, this suggests the possibility that long-run VA is partially capturing a teacher's non-cognitive VA and their ability to improve students' behavioral skills.

More generally, our work speaks to the importance of taking knowledge fade out into consideration in the assessment of teachers. In particular, our research shows that some teachers are good at producing a temporary recognition that disappears the following year, while other teachers impart deep knowledge that persists. To the extent that teachers can substitute between teaching these two forms of knowledge, policies targeting VA improvements can improve contemporaneous test scores (as shown by and Biasi (Forthcoming) and Macartney et al. (2018)), but worsen long-term educational outcomes. Targeting teachers based on their ability to impart long-term knowledge either directly or through incentives has the potential to lead to large improvements in longer term educational outcomes over policies targeting more myopic performance measures.

The rest of the paper is organized as follows: The next section introduces the theoretical model which highlights how teachers' contributions to long- and short-term knowledge can be separately identified. Section 3 introduces the administrative data that we use and our VA estimation procedure. In Section 4, we estimate long- and short-run VA, describing our model estimates and then linking these VA measures to long-run outcomes. We check for bias in VA

estimates in Section 5. Section 6 then conducts policy analysis, highlighting the benefit of using our long-run VA measure. Section 7 concludes.

2 Conceptual Framework

This section introduces a dynamic model of education production whereby teacher-induced learning gains have both short- and long-term components as in Jacob et al. (2010) and Cascio and Staiger (2012). The model highlights how teachers' contribution to long- and short-term knowledge can be separately estimated using standard VA methods.

Standard VA Model: Teachers are indexed by j and teach one class per year (as in our empirical application). For simplicity, we ignore classroom shocks in our model, although we do account for them in our empirical implementation. Teacher j increases the test score of student i who is assigned to them in period t by their value-added, μ_{jt}. Formally, the test score of student i in year t assigned to teacher j, A^*_{ijt}, is given by:

$$A^*_{ijt} = \beta X_{ijt} + \mu_{jt} + \epsilon_{ijt}\,, \tag{2.1}$$

where X_{ijt} are observable determinants of student achievement including lagged test scores and student demographics, μ_{jt} is teacher j's contribution, and ϵ_{ijt} represents idiosyncratic student variation. As is standard in the VA literature, we work with residualized test scores, A_{ijt}, which we construct by removing the effect of observable characteristics:

$$A_{ijt} \equiv A^*_{ijt} - \beta X_{ijt} = \mu_{jt} + \epsilon_{ijt}\,, \tag{2.2}$$

where β is estimated using variation across students taught by the same teacher using the following OLS regression:

$$A^*_{ijt} = \alpha_j + \beta X_{ijt}\,, \tag{2.3}$$

where α_j is a teacher fixed effect.

The estimation of μ_{jt} in equation (2.2) has been the focus of the prior VA literature. For

now, it is sufficient to note that μ_{jt} can be estimated using shrinkage estimators as in Chetty et al. (2014a). We discuss our estimation procedure for μ_{jt} formally in Section 3.2. Hereafter, we call estimates of μ_{jt} 'normal' VA.

Regardless of the exact estimation procedure, for $\hat{\mu}_{jt}$ to provide unbiased estimates of μ_{jt} it must (almost) surely be the case that the observable characteristics X_{ijt} used to construct the test score residuals A_{ijt} are sufficiently rich. Specifically, the unobservable determinants of student achievement must be balanced across teachers so that remaining unobserved heterogeneity in ϵ_{ijt} is balanced across teachers with different VA estimates. To ensure this is the case, we impose the commonly-invoked (in the VA literature) no sorting on unobservables assumption:

Assumption 1 *Students are not sorted to teachers based on unobservable determinants of student achievement and so* $\mathbb{E}[\epsilon_{ijt}|j] = \mathbb{E}[\epsilon_{ijt}]$.

Long- and Short-Run VA Model: We depart from the standard VA model and consider knowledge 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 model that teachers augment both types of knowledge and therefore a teacher's total impact on test scores, μ_{jt}, can be separate into a short-, μ_{jt}^S, and long-term, μ_{jt}^L, knowledge component:

$$\mu_{jt} = \mu_{jt}^L + \mu_{jt}^S \,. \tag{2.4}$$

Henceforth, we call μ_{jt}^L *long-run VA* and μ_{jt}^S *short-run VA*.

We therefore have that (residualized) test scores at time t are given by:

$$A_{ijt} = \mu_{jt}^L + \mu_{jt}^S + \epsilon_{ijt} \,. \tag{2.5}$$

Unfortunately, the econometrician can only estimate the sum of the teacher components, μ_{jt}, in equation (2.5) and not the teacher's contribution to the individual knowledge components. To do so, we consider achievement the following year when student i is assigned to teacher k. In

this period, (residualized) achievement will be given by:

$$A_{ijk,t+1} = \delta^L \mu^L_{jt} + \delta^S \mu^S_{jt} + \mu_{k,t+1} + \epsilon_{ijk,t+1}\,, \tag{2.6}$$

where δ^L and δ^S parametrize the fade out of the short and long-term components of knowledge between periods t and $t+1$.

We now make clear our conceptualization of short- and long-term knowledge. On one hand, we view short-term knowledge as being shallow and transitory (e.g., rote memorization) and so will not persist into future periods. Long-term knowledge, on the other hand, is formed either through a deep understanding of material or through behavioural improvements (e.g., time spent studying) that will persist into future periods. In our context, we specifically coin 'long-term knowledge' as any learning gained while being taught by teacher j that fully persists into the next period.[2] Formally, our conceptualization assumes:

Assumption 2 *Short-term knowledge completely fades out and so* $\delta^S = 0$.

Assumption 3 *Long-term knowledge persists in its entirety between periods* t *and* $t+1$ *and so* $\delta^L = 1$.

While these assumptions are imposed here, we can *test* whether the assumptions hold in the data by checking whether $\delta^S = 0$ and $\delta^L = 1$ when we regress our short- and long-run VA measures on (residualized) test scores in period $t+1$. Under these assumptions, (residualized) achievement in $t+1$ is:

$$A_{ijk,t+1} = \mu^L_{jt} + \mu_{k,t+1} + \epsilon_{ijk,t+1}\,. \tag{2.7}$$

To estimate μ^L_{jt}, we require that Assumption 1 – that students are not sorted to teachers based on unobservable components of student achievement – holds for two years instead of just one. In addition, students cannot be sorted to teachers in period $t+1$ based on teacher assignment in period t. This assumption is identical to that of Petek and Pope (2018). Formally:

[2] In light of this definition, long-run VA will not capture learning gained while being taught by teacher j that is not captured by period $t+1$ test scores (e.g., a deep understanding of material that is useful two grades into the future, but not next grade). In principle, we could replace period $t+1$ test scores with period $t+2$ test scores to capture these components, but the gains of doing so are limited relative to costs such as reduced sample sizes.

Assumption 4 *Students are not sorted to teachers in period $t+1$ based on unobservable determinants of student achievement or teacher assignment in t and so $\mathbb{E}[\epsilon_{ijk,t+1}|j,k] = \mathbb{E}[\epsilon_{ijk,t+1}]$.*

We now rewrite equation (2.7) as:

$$A_{ijk,t+1} = \mu^L_{jt} + \tilde{\epsilon}_{ijk,t+1}\,, \tag{2.8}$$

where $\tilde{\epsilon}_{ijk,t+1} \equiv \mu_{k,t+1} + \epsilon_{ijk,t+1}$. Under Assumption 4, we have that the expected VA of teacher k is zero and so $\mathbb{E}[\tilde{\epsilon}_{ijk,t+1}] = 0$. This equation looks very similar to the standard VA estimating equation (see equation (2.3)) and so can be estimated using standard VA techniques (formally discussed in Section 3.2). With an estimate of the long-run VA component, μ^L_{jt}, in hand, we could estimate the short-run VA component using equation (2.4) (i.e., $\mu^S_{jt} = \mu_{jt} - \mu^L_{jt}$). Equivalently, since $\delta^L = 1$, we could sub equation (2.6) into (2.5) and rearrange:

$$A_{ijt} - A_{ijk,t+1} = \mu^S_{jt} + \breve{\epsilon}_{ijkt}\,. \tag{2.9}$$

where $\breve{\epsilon}_{ijt} \equiv \epsilon_{ijt} - \tilde{\epsilon}_{ijk,t+1}$.

To recap, we have constructed three different VA measures: 'normal' VA, μ_{jt}, long-run VA, μ^L_{jt}, and short-run VA, μ^S_{jt}. The estimating equations for each are as follows:

1. **Normal Value-Added:** $A_{ijt} = \mu_{jt} + \epsilon_{ijt}$
2. **Long-Run Value-Added:** $A_{ijk,t+1} = \mu^L_{jt} + \tilde{\epsilon}_{ijk,t+1}$.
3. **Short-Run Value-Added:** $A_{ijt} - A_{ijk,t+1} = \mu^S_{jt} + \breve{\epsilon}_{ijkt}$.

3 Data and VA Estimation

This section describes the administrative data that we use and provides descriptive statistics. We then detail the control vector we use to residualize test scores followed by our VA estimation procedure.

3.1 Data

We use administrative data from a large urban school district which links students to teachers over time. Our data span school years 2002-03 through 2016-17 and cover all students enrolled in the district from grades K-12. We lack student-teacher links in middle schools, so we restrict attention to elementary grades. State standardized tests in math and English are run for grades 2-8 from 2002-03 through 2012-13[3] and so our main VA analysis sample cover third through fifth grades from 2003-04 through 2011-12 to ensure that we have lagged and subsequent test scores for our entire sample. Our main VA sample thus cover roughly 650,000 students with 1.45 million student-year observations.

In addition to test scores, our data also contain behavioral variables which we use to construct non-cognitive VA measure as in Jackson (2018) and Petek and Pope (2018). In particular, for all grades K-12 we observe the number of days a student was suspended, the number of days a student was absent,[4] whether a student progresses on time to the next grade (i.e., held back), and achievement grades in 10 subjects (e.g., reading, mathematics, art, etc.) for each trimester, which we use to construct a student's GPA. For absences, we take the log of a student's days absent (plus one), while for suspensions we employ an indicator for being suspended. We limit our non-cognitive VA sample to third through fifth grades from 2003-04 through 2011-12 to align with our test score VA sample.

Our data also include detailed student demographics, including information about parental education (five education groups), economically disadvantaged status, ethnicity (seven ethnic groups), gender, limited English status, and age. Demographic coverage is nearly one hundred percent for all demographic variables with the exception of parental education, which is missing for twenty-nine percent of the sample.

We make several sample restrictions to end up at our final VA analysis sample. First, we drop 148,644 student-year observations that cannot be matched to a teacher or the teacher is

[3] In 2013-14, the state switched to a new testing system to align their content to Common Core. Due to this switch, no test scores are available for 2013-14. Starting in 2014-15 test scores become available again for grades 3-8. Given the requirement for lagged and subsequent test scores for our methodology, we could only incorporate 2015-16 test scores for grades 4-5 into the VA analysis sample. For consistency, however, we focus our analysis on pre-2014 cohorts.

[4] Unfortunately, data on absences are unavailable for 2002-03.

assigned to multiple grades or schools within the same year.[5] Second, we only include classes with more than seven but fewer than forty students with valid current, lagged, and subsequent test scores in that subject, losing 14,368 student-year observations. Third, we exclude 99,336 observations that lack a valid current or lagged test score in that subject.[6] Our final sample is roughly 1.2 million student-year observations, covering roughly 550,000 students and 13,000 teachers.

Table 1 provides summary statistics for our data. Column (1) reports these for the full sample, while column (2) does so for our VA analysis sample. Our district is highly-disadvantaged with seventy percent of students being eligible for free or reduced price lunch and over a third coming from families whose parents are high school dropouts. The student body is also three-quarters Hispanic, with black and white students each making up a further ten percent. The VA analysis sample is similar to the full sample, although is somewhat positively selected. This positive selection is common in VA papers and is driven by the requirement for lagged test scores, which drops newly-arrived students who tend to be lower-performing. There appears to be little selection into our VA sample that is linked to high school outcome data (see column (3)).

Long-Run Outcomes Data: Our VA analysis data are then linked to high school outcome data. The high school outcome data cover a range of high school outcomes, including: algebra scores, exit exam scores, PSAT scores, SAT-taking and scores, AP classes taken, graduation, and high school GPA, 'effort GPA,' absences, suspensions, and grade repetition. For outcomes which students can retake (e.g., exit exam, SAT, etc.) we take the score from their first attempt. High school outcomes (e.g., GPA, absences) incorporate their entire grade 9-12 high school career.

Our long-run outcomes do not cover all students in our VA sample as some cohorts have yet to reach the required age to achieve that outcome (e.g., third grade students in 2012-13 have not taken the SAT by 2016-17). Table A.1 describes each long-run high school outcome that we use and the cohorts in the VA analysis data set that it covers. It also reports the match rate between the two data sets among eligible cohorts. The match rate is not perfect as any student

[5] We define a teacher as teaching multiple grades or schools if more than three of their students come from a different grade or school than the modal grade or school.

[6] As the last two restrictions are subject-specific, our VA samples are also subject-specific. Our math VA sample has 1,185,181 observations while our English VA sample has 1,181,362 observations.

leaving the school district between elementary school (grades 3-5) and high school (grades 9-12) will not be matched. For most high school outcomes, the match rate is around seventy percent, although outcomes where participation is voluntary (e.g., the SAT) have lower match rates. We do not find any evidence that any of our VA measures influence the probability a student is present in our long-run outcome data (see Figure A.2).

3.2 Estimation of Value-Added

Constructing Test Score Residuals: We construct test score residuals for each subject (math and English), A_{ijt}, by regressing raw standardized test scores, A^*_{ijt}, on a vector of covariates, X_{ijt}, and teacher fixed effects, as described by equation (2.3). Our baseline control vector is similar to that of Chetty et al. (2014a), although we also include lagged non-cognitive measures as in Jackson (2018) and Petek and Pope (2018) to account for potential sorting based on student behavior (in addition to test scores).

We control for the following student-level controls: (i) lagged test scores using a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English, interacted with grade dummies,[7] (ii) four lagged non-cognitive measures (suspensions, absences, held back, and GPA) using a cubic polynomial in each of these prior-year measures interacted with grade dummies,[8] (iii) demographics, including: parental education (five education groups plus a missing data category), economically disadvantaged status, ethnicity (seven ethnic groups), gender, limited English status, and age interacted with grade dummies. We also include the following class- and school-grade level 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) interacted with grade dummies, (ii) cubics in class and school-grade means of prior-year non-cognitive measures (suspensions, absences, held back, and GPA) interacted with grade dummies, (iii) class and school-grade means of all the demographic covariates, (iv) class size, and (v) grade and year dummies.

[7] If data on lagged scores is missing, we exclude observations if prior scores in the subject for which we are estimating VA are missing; otherwise we set the other subject prior score to zero and include an indicator for missing data in the other subject interacted with the controls for prior own-subject test scores.

[8] If data is missing for a student's non-cognitive measure we set the non-cognitive measure equal to the mean of that measure for the student's grade level and include an indicator for missing data.

VA Estimator: We follow Chetty et al. (2014a) and estimate μ_{jt} using a jack-knife empirical Bayes estimator that allows for drift. Specifically, let $\bar{A}_{jt} \equiv \frac{1}{n}\sum_{i \in j} A_{ijt}$ denote the mean (residualized) test score in the class taught by teacher j in year t. Let $\bar{\mathbf{A}}_j^{-t} \equiv (\bar{A}_{j1}, ..., \bar{A}_{jt-1}, \bar{A}_{jt+1}, ..., \bar{A}_{jT})'$ denote the vector of mean (residualized) scores in all classes taught by teacher j *except* in period t. The following OLS regression is then run to obtain the best linear predictor of $\bar{A}_{jt}$:

$$\bar{A}_{jt} = \psi \bar{\mathbf{A}}_j^{-t}, \tag{3.1}$$

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:

$$\hat{\mu}_{jt} = \hat{\psi} \bar{\mathbf{A}}_j^{-t}. \tag{3.2}$$

Estimation of the long-run components of VA follow similar methods, but redefines $\bar{A}_{jt}$ based on (residualized) test scores in period $t+1$. Specifically, $\bar{A}_j^{t+1} \equiv \frac{1}{n}\sum_{i \in j} A_{ij,t+1}$ and $\bar{\mathbf{A}}_j^{t+1,-t} \equiv (\bar{A}_{j1}^{t+1}, ..., \bar{A}_{jt-1}^{t+1}, \bar{A}_{jt+1}^{t+1}, ..., \bar{A}_{jT}^{t+1})'$. The empirical analog for ϕ^{t+1} is then estimated regressing $\bar{A}_{jt}^{t+1} = \psi \bar{\mathbf{A}}_j^{t+1,-t}$. The estimate for the long-run component of teacher j's VA in period t is then given by:

$$\hat{\mu}_{jt}^{L} = \hat{\psi}^{t+1} \bar{\mathbf{A}}_j^{t+1,-t}. \tag{3.3}$$

Analogously, the short-run component of VA is estimated where $\bar{A}_{jt}$ is redefined as $\bar{A}_{jt}^{\Delta} \equiv \frac{1}{n}\sum_{i \in j} A_{ijt} - A_{ij,t+1}$. We estimate non-cognitive VA as in Petek and Pope (2018) by redefining $\bar{A}_{jt}$ based on (residualized) non-cognitive outcomes in period $t+1$.

A few comments are due about the VA estimator we use and how it relates to tests of biasedness (implicitly testing Assumption 1) in the literature. Specifically, the literature tests whether VA estimates $\hat{\mu}_{jt}$ accurately predict differences in the mean test scores of students assigned to teachers in year t. To do so, it runs a regression of the following form:

$$A_{ijt} = \phi_t + \lambda \hat{\mu}_{jt} + \xi_{ijt}. \tag{3.4}$$

Under Assumption 1, $\lambda = 1$, and so the literature uses this as a test for so-called 'forecast unbiasedness.'

Two features of the VA estimator help $\lambda = 1$. First, VA estimates are leave-year-out (jackknife) measures of teacher quality. This avoids bias arising from including student outcomes in year t on teacher VA estimates without leaving out the data from year t, which introduces the same estimation errors on both the left and right hand side of the regression, biasing VA estimates. Second, the VA estimator accounts for drift which takes into account that more recent classes taught by a teacher are better predictors of current teacher quality. Omitting such drift attenuates λ away from one.

The VA estimator defined by equation (3.2), however, is not the best predictor of VA (with available data) for two reasons. First, the best predictor should utilize all available data to form predictions, including year t, even if this introduces some mechanical bias. Second, the VA estimator defined by equation (3.2) is only the best *linear* predictor; the best non-linear predictor of $\hat{\mu}_{jt}$ can be constructed using nonparametric empirical Bayes (NPEB) (Gilraine et al., 2020). Unfortunately, however, NPEB cannot currently account for drift[9] and so leave-year-out NPEB estimates are attenuated by drift causing estimates of λ to be less than one. For these reasons, we estimate $\hat{\mu}_{jt}$ using equation (3.2) in what follows to check for forecast unbiasedness in our estimates. When considering policy simulations, however, the goal is to form the most accurate predictions of $\hat{\mu}_{jt}$ (rather than unbiasedness) and therefore we use all available years of data for teacher j and NPEB to construct estimates of $\hat{\mu}_{jt}$. Gilraine et al. (2020) show that NPEB can substantially improve the accuracy of policy predictions.

Combining VA Measures: We reduce the dimensionality of the VA measures by constructing two VA indices: test score VA and non-cognitive VA. Test score VA merely combines our math and English VA estimates giving each subject equal weight (i.e., $VA_{test} = \frac{1}{2}VA_{math} + \frac{1}{2}VA_{English}$). This is done for all three of our VA measures: 'normal' VA, long-run VA, and short-run VA.

The non-cognitive VA index is computed using VA for suspensions, log days absent, GPA, and not progressing to the next grade on time (i.e., held back). We compute the index by summing the standardized value-added variables, recoded so each has the same expected sign, and then standardizing the resulting index to be mean zero, standard deviation one. Alternatively, one could use exploratory factor analysis to choose the factor load on each VA variable as done in

[9] Although some exciting advances are being made in this direction – see Gourieroux and Jasiak (2020), for instance.

Petek and Pope (2018); our results are robust to these factor loadings.[10]

4 Results

We estimate non-cognitive, 'normal,' long-, and short-run VA using the methodology described above. Here, we describe our VA model estimates, paying particular attention into how 'normal' VA decomposes into its long- and short-run components and how these measures correlated to each other and to non-cognitive VA. We then link these VA measures to long-run outcomes using the methodology from Chetty et al. (2014b).

4.1 VA Model Estimates

As discussed above in Section 3.2, we estimate our VA models separately for each subject and then combine them into a single test score VA index. We focus our discussion on Table 2 which reports parameter estimates for normal, long- and short-run VA in mathematics. (Analogously, Table A.2 reports parameter estimates for English).

The first six rows of Table 2 report the autocorrelations of mean test score residuals across classes taught in different years by a given teacher. These autocorrelations represent the reliability of mean class test scores for predicting teacher quality s years later. Unsurprisingly, reliability decays over time and so more recent test scores are better predictors of current teacher performance. The reliability of long- and short-run VA are lower than that of normal VA. Intuitively, this arises since normal VA is the sum of the long- and short-run components and so it should be more reliable over time than either of its constituent components. Long-run VA features lower reliability than short-run VA. Speculatively, this may indicate that teachers' contributions to long-term knowledge are more difficult to maintain or that teachers substitute long-term knowledge contributions with their short-term counterpart.[11] Such differences in 'drift' will be incorporated into the policy analysis later.[12]

[10] We find that our factor loadings are near-identical to those in Petek and Pope (2018) and so we refer interested readers to their paper, although our results are available upon request.

[11] For example, if long-term knowledge contributions are costlier than short-term contributions teachers may substitute toward short-term contributions if test score gains are incentivized or their motivation for teaching fades.

[12] In particular, less reliable VA estimates decrease the expected policy gains since low-VA teachers are more likely to drift back toward the mean.

Table 2 also reports the estimated standard deviation of teacher effects. Since teachers teach only one class per year, we cannot point identify the standard deviation of teacher effects as (unforecasted) innovations in teacher effects cannot be separated from idiosyncratic class shocks. Regardless, we can obtain a lower bound or use a quadratic approximation to the standard deviation of teacher effects; estimates from both methods are similar since rate of drift across one year is small.

We find that the standard deviation of teacher effects is 0.277 for normal VA. Note that since long- and short-run VA are negatively correlated, the sum of the variance for long- and short-run VA is greater than the variance of normal VA. For long- and short-run VA, we estimate that the standard deviation of teacher effects is 0.193 for long-run VA and 0.235 for short-run VA, suggesting that the distribution of long-run teacher effects is less dispersed than its short-run counterpart. A simple variance decomposition indicates that sixty percent of the variation in normal VA measures come from contributions to short-term knowledge, while forty percent of the variation come from long-term knowledge contributions.

Relationships Between VA Measures: Table 3 reports the relationship between our various VA measures (Table A.3 further reports the correlations of all VA components). Here, some interesting patterns emerge. First, unsurprisingly both short- and long-run VA are highly correlated to the 'normal' VA measure given that these two measures are the underlying foundation of 'normal' VA. The short- and long-run VA components themselves, however, are negatively correlated. Interestingly, the long-run component of VA is strongly correlated with our non-cognitive VA measure, while the short-run VA component is not. This suggest that long-run VA is picking up components of teaching that not just raise test scores in the following period but *also* improve student behaviour.

4.2 Impact on Future Outcomes

Methodology: We link future outcomes to teacher VA using the method proposed by Chetty et al. (2014b). This method compares the future outcomes of students who were assigned to teachers with different VA, controlling for a rich set of student characteristics. To start, we construct future outcome residuals using variation across students taught by the same teacher,

based on the regression equation

$$Y_i^* = \alpha_j + \beta^Y X_{it}\,, \tag{4.1}$$

where Y_i^* is the long-run outcome of interest, α_j is a teacher fixed effect, and X_{it} are observed characteristics of the student and the teacher. Using the estimates from equation (4.1), the long-run residuals, Y_{it}, are defined as:[13]

$$Y_{it} = Y_i^* - \hat{\beta}^Y X_{it}\,. \tag{4.2}$$

When long-run outcomes are defined as a future test score (e.g., test scores in the subsequent year), we simply regress these long-run residuals on our jack-knife teacher VA measure, pooling across all grades. This allows us to test model assumptions, such that long-run VA will raise test scores in the subsequent year by one unit (and short-run VA by zero units). For non-test-score based long-run outcomes, we pool across all grades and regress them on teachers' jack-knifed *normalized* VA to account for differences in the dispersion across the various VA measures. Formally,

$$Y_{it} = \delta + \rho^{\kappa}\hat{m}_{jt}^{\kappa} + \eta_{ij}\,, \quad \kappa \in \{N, L, S, NC\}, \tag{4.3}$$

where $\hat{m}_{jt}^{\kappa} \equiv \hat{\mu}_{jt}^{\kappa}/\hat{\sigma}_{\mu}$ denotes 'normalized' teacher VA, which is our estimate of a teacher's VA scaled by the estimated standard deviation ($\hat{\sigma}_{\mu}^{T}$) of the teacher VA distribution for VA type κ.[14]

Future Test Scores: Figure 1 plots the impacts of our three test score based VA measures on test scores in subsequent years. The coefficients at zero are statistically indistinguishable from one for all three of our test score VA measures – in line with the modeling assumptions. In the following period, long-run VA continues to raise test scores by one, while the impact of short-run VA on test scores falls to zero – in line with assumptions 2 and 3. For future periods, short-run does not impact test scores significantly while long-run VA continues to significantly raise test scores, giving confidence that long-run VA is capturing teachers' contributions to a

[13] This is the same method we used to construct residualized test scores. Therefore, if the long-run outcome is test scores in the subsequent period, the future outcome residuals will be identical to our residualized $t+1$ test scores used to calculate our long-run VA measure.

[14] As in Chetty et al. (2014b), the standard deviation of the normalized VA measures are less than one since Bayes shrinkage is applied to the VA estimates.

deeper understanding of the material or improved behaviour that continues to persists well into the future. The impact of 'normal' VA on future test scores is between the impacts of short- and long-run VA, which is sensible given that 'normal' VA consists of these two underlying components.

Impacts on High School Outcomes: Figure 2 plots (residualized) long-run outcomes for students in school year t versus (jack-knifed) *normalized* long- and short-run VA. For visual clarity, the impact of 'normal' VA is omitted from the figures, its effects lies between the long- and short-run VA impacts. We construct the binned scatter plots in three steps: (i) residualize the long-run outcome with respect to our control vector using within-teacher variation as described by equation (4.1), (ii) divide the normalized long- and short-run VA measures, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the long-run outcome in the estimation sample to facilitate interpretation of the scale.

Figure 2 reports the point estimates of equation (4.3) and visualizes the impact of long- and short-run VA on twelve high school outcomes. In terms of the slope coefficients underlying each panel, we find that being assigned to a teacher whose *long-run* VA is one standard deviation higher in a single grade increases algebra scores by 0.05 standard deviations, raises exit exam scores by 3 points, improves PSAT scores by 10 points (on the 2400 scale), boosts SAT scores by 9 points (on the 1600 scale) *in addition to* increase SAT-taking rates by 0.6 percentage points, raises number of AP classes taken by 0.04, improves high school graduation rates by 0.7 percentage points, boosts high school GPA by 0.02, raises high school 'effort GPA' by 0.01, *decreases* days absent in high school by 3 percent, reduces days suspended in high school by 0.003, and lowers grade repetition by 0.22 percentage points. All of these improvements in long-run outcomes are statistically significant at the one percent level. In contrast, being assigned to a teacher whose *short-run* VA is one standard deviation higher in a single grade either has little impact or causes a deterioration in the long-run outcome.

Prior research has found that non-cognitive VA can independently affect students' performance in high school and is particularly effective at improving behavioral-based high school outcomes such as GPA and absences (Jackson, 2018; Petek and Pope, 2018). We contrast the

impact of non-cognitive VA to our long-run VA measure in Figure 3 using the same method as in Figure 2. In line with prior findings, we find that being assigned to a teacher whose *non-cognitive* VA is one standard deviation higher in a single grade somewhat improves test score based outcomes (i.e., algebra scores) and meaningfully enhances behavioral-based high school outcomes (i.e., suspensions, absences, GPA). Despite that, being assigned to a teacher whose *long-run* VA is one standard deviation higher in a single grade leads to far more substantial improvements in *every* long-run outcome that we consider.

Multivariate VA Effects: The prior analysis looked at how each dimension of teacher quality affect high school outcomes. Given that these VA measures are correlated, however, we want to determine whether each VA measure independently affects long-term outcomes. For example, Figure 2 indicates that being assigned a high short-run VA teacher lead to a deterioration in long-run outcomes, which could either be caused by short-run VA itself *or* because short-run VA is negatively correlated with the long-run VA that drives vast improvements in long-run outcomes. We investigate the extent to which different dimensions of teacher quality matter for long-term student outcomes by regressing:

$$Y_{it} = \delta + \rho^{L}\hat{m}^{L}_{jt} + \rho^{S}\hat{m}^{S}_{jt} + \rho^{N}C\hat{m}^{N}C_{jt} + \eta_{ij}\,, \tag{4.4}$$

where the superscripts L, S, and NC represent long-run, short-run, and non-cognitive VA, respectively. 'Normal' VA is omitted from this analysis given that it would be collinear with long- and short-run VA.

Table 4 reports the results of the multivariate regression. Conditioning on long-run VA causes short-run VA to have little material effect on long-run outcomes, likely as some of the negative impact of short-run VA in the univariate analysis captured the negative correlation between short- and long-run VA (as foreshadowed). Importantly, accounting for long-run VA subsumes a large proportion of the improvements caused by being assigned to high non-cognitive VA teachers, with the impact of being assigned high non-cognitive VA teachers becoming negative for several test score-based outcomes such as PSAT and SAT scores. Non-cognitive VA continues to improve behavioral-based long-run outcomes such as graduation and GPA independently. The

inclusion of non-cognitive VA, on the other hand, only marginally affects the point estimates for long-run VA: Being assigned to a teacher whose long-run VA is one standard deviation higher continues to substantially improve both test score and non-test-score-based long-run outcomes. In fact, given the superiority of long-run VA teachers in improving students' outcomes, there appears to be little advantage in incorporating non-cognitive VA in high-stakes decision-making.[15]

5 Testing for Bias

The estimates of the impact of long- and short-run VA on long-run outcomes in the previous section rely on the assumption that the unobserved determinants of students' long-term outcomes are uncorrelated with teacher quality conditional on observables. In this section we verify the impact of teachers' long-term impacts four ways: (i) ensure that long- and short-run VA are forecast unbiased, (ii) test for sorting based on twice-lagged outcomes, (iii) conduct event studies on the entry or exit of high or low VA teachers, and (iv) use quasi-experimental variation that leverages staffing changes among the full distribution of teachers.

5.1 Forecast Unbiasedness

'Normal' VA estimates are "forecast unbiased" if teachers whose estimated VA is one-unit higher cause students' test scores to increase by one-unit on average (Chetty et al., 2014a). In our context, forecast unbiasedness similarly requires that teachers whose estimated long- and short-run VA is one-unit higher cause students' test scores to increase by one-unit on average. Forecast unbiasedness also requires that teachers whose estimated long-run VA is one-unit higher cause students' test scores to increase by one-unit on average in the *subsequent* year; correspondingly, forecast unbiasedness requires that teachers whose estimated short-run VA is one-unit higher cause students' test scores to increase by *zero* units on average in the *subsequent* year.

Table 5 tests for forecast unbiasedness in all three VA measures. We see that 'normal,' long-, and short-run VA increase test scores in the current period by an amount that is not

[15] With the possible exception of SAT-taking. Obviously, since non-cognitive VA continues to improve long-run behavioral outcomes independently, there is information contained in it that is not capture by long-run VA. If researchers could more precisely identify teachers' contributions to these behaviors, then it will become optimal for policymakers to incorporate this type of VA into their decision-making.

statistically distinguishable from one and so forecast unbiasedness cannot be rejected. In the subsequent period, we similarly cannot statistically reject that long-run VA raises test scores by one unit or that short-run VA increases test scores by zero units. Forecast unbiasedness can therefore not be rejected for all three of our test score-based VA measures.

5.2 Sorting on Twice-Lagged Outcomes

Although we cannot observe whether students sort on unobservable determinants of student achievement, we can assess whether students sort on variables that predict test score residuals but are omitted from the VA model. While such observable determinants are limited given the expansive control vector that we use, one notable observable remains: twice-lagged outcomes such as test scores and non-cognitive outcomes.

We estimate forecast bias using twice-lagged outcomes following the methodology outlined in Chetty et al. (2014a). First, we construct residual outcomes $\mathbf{Y}_{it}^{-2}$ by regressing each element of $\mathbf{Y}_{it}^{*-2}$ on our control vector X_{ijt} and teacher fixed effects, as in equation (2.3). Second, we regress residualized test scores A_{ijt} on $\mathbf{Y}_{it}^{-2}$, again including teacher fixed effects, and calculate predicted values $A_{ijt}^{Y} = \hat{\boldsymbol{\rho}}\mathbf{Y}_{it}^{-2}$. The twice-lagged outcomes we use are: math scores, English scores, suspensions, absences, held back, and GPA. The need for twice-lagged outcomes eliminates third grade students from our sample for this test.

Figure 4 visualizes the sorting based on twice-lagged outcomes by dividing the long- or short-run VA estimates into twenty equal-sized groups (vingtiles) and plotting the means of the residuals, A_{ijt}^{Y}, within each bin against the mean value of the VA estimate within each bin. Residual test scores are also shown, which nonparametrically plots test score residuals against the VA estimates and has a slope near one indicating the forecast unbiasedness found in the prior subsection. Similar to Chetty et al. (2014a), we find a small positive relationship between predicted scores based on twice-lagged outcomes and 'normal' VA: the coefficient is 0.013 (s.e. 0.002).[16] Interestingly, we find that this is driven entirely by the relationship between predicted scores based on twice-lagged outcomes and long-run VA. Regardless, the relationship between long-run VA and predicted scores, A_{ijt}^{Y}, is small relative to the relationship between VA and test score residuals.

[16] Chetty et al. (2014a) find a coefficient of 0.022 (s.e. 0.002) in their analysis.

5.3 Event Studies

While long- and short-VA estimates are forecast unbiased and sorting based on twice-lagged outcomes to teachers is limited, this does not rule out the possibility that students are sorted to teachers based on unobservable characteristics that are orthogonal to twice-lagged outcomes. Here, we use quasi-experiments that exploit naturally occurring teacher turnover to test for bias. We start by leveraging movements of teachers in the tail of the long- and short-run VA distributions, finding that the entry or exit of short- and long-run VA teachers influence contemporaneous test scores *but* only the entry or exit of long-run VA teachers affects test scores in the following year.

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 2009-10, then event year '0' is 2009-10 and event year '-1' is 2008-09). An entry event is defined as the arrival of a teacher who did not teach in that school for the three preceding years; an exit event is defined as the departure of a teacher who does not return to that school for at least three years. A teacher is defined as high- (low-) VA if her estimated VA in her year of entry or exit is in the top (bottom) 5 percent of all entrants or leavers.[17] We estimate the VA of each entering teacher by excluding event years $t \in [-3, 2]$ from their VA calculation,[18] ensuring that VA is calculated using data from students outside the six year school-grade event window.[19] Our analysis focuses solely on the variation induced by school switchers because school-grade switchers may impact students in the subsequent period which could induce spurious event study results.[20]

[17] Following Chetty et al. (2014a), we use mean VA to decide whether the event falls in the top or bottom 5 percent of the VA distribution if multiple teachers enter or exit at the same time. We also 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 2008-09 and 2010-11).

[18] Precisely, we follow Chetty et al. (2014a) and estimate VA for each teacher excluding a five-year window (two years prior, the current year, and the two subsequent years). Couple with the entry and exit definitions this ensures that no data from the relevant school-grade between event years $t \in [-3, 2]$ are used to compute teacher VA.

[19] 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 $t \in [0, 2]$ in their VA calculations is sufficient to address this concern. Analogously, since the exiting teacher was not in the school for event years $t \in [0, 2]$ excluding event years $t \in [-3, -1]$ from their VA calculations addresses this concern.

[20] For example, suppose a high short-run VA teacher switches from fourth to fifth grade within the same school. Then we will see that the exit of the high short-run VA teacher is associated with a decrease in fourth grade test scores but an increase in those fourth grade students subsequent test scores. The increase in subsequent test

Results: Panel A of Figure 5 plots the impact of the entry of a high-VA teacher on mean residualized test scores in the current and subsequent year. Specifically, the solid series plots school-grade-subject-year test scores in the current (left-side figure) and subsequent year for those students (right-side figure) before and after a high long-run VA teacher enters the school-grade. Similarly, the dashed line does so for the entry of a high short-run VA teacher. Effects are normalized to zero in the period before the teacher enters (i.e., period '-1').

When a high long- or short-VA teacher arrives, residualized contemporaneous test scores in the grade taught by the teacher rise immediately. Test scores before the teacher arrives are stable, indicating that there are no trends in school quality or unobserved student characteristics before the teacher's arrival.

The arrival of a high long-run VA teacher raises mean long-run VA in that school-grade by 0.03, while the arrival of a high short-run VA teacher raises mean short-run VA in that school-grade by 0.08. The higher increase in mean VA caused by the arrival of high short-run VA teachers relative to long-run VA teachers is due to the fact that the standard deviation in short-run VA is twice that of long-run VA. The arrival of a high long and short-run VA teacher increases contemporaneous test scores by 0.04σ and 0.08σ, respectively. Both of these test score increases are very similar to the change in mean teacher VA. In fact, the hypothesis that the observed impact on contemporaneous test scores equals the increase in mean VA is not rejected for either long- or short-run VA, consistent with long- and short-run VA estimates being forecast unbiased.

In contrast, the arrival of high long- and short-run VA teachers have very different impacts on the test scores of their students in the *following* year. On one hand, the arrival of a high long-run VA teacher raises their students' test scores in the subsequent year by 0.04σ. On the other hand, the arrival of a high short-run VA teacher does not affect their students' test scores in the subsequent year (point estimate of 0.01σ). Indeed, the hypothesis that the observed impact on subsequent test scores equals the increase in mean short-run VA is rejected. These results therefore provide direct evidence for our hypothesis that while both the long- and short-run components of VA impact contemporaneous test scores, only the long-run component of VA

scores, however, is being driven by those students being taught by the high short-run VA teacher in fifth grade rather than high short-run VA teachers influencing subsequent test scores.

affects future performance.

Panel B of Figure 5 repeats the event study for low-VA teacher entry. Here, the entry of a low long- or short-run VA teacher lowers contemporaneous test scores. Once again, only the entry of a low long-run VA teacher negatively affects her students' test scores in the subsequent year. Figure A.3 then conducts the teacher exit event studies. These event studies are somewhat noisier, but yield similar conclusions in that the hypotheses that the observed impact on subsequent test scores equals the increase in mean VA is always rejected for short-run VA, but never rejected for long-run VA.

5.4 Quasi-Experimental Estimates

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 show that an increase in the long-run VA of teachers increases both current *and* future test scores, while an increase in short-run VA only raises current test scores.

Methodology: Let $\hat{\mu}_{jt}^{L,-\{t,t-1\}}$ and $\hat{\mu}_{jt}^{S,-\{t,t-1\}}$ respectively denote the long- and short-run VA estimates for teacher j in year t constructed as described in Section 3.2 using data from all years except t and $t-1$. Let Q_{sgt}^{L} and Q_{sgt}^{S} denote the (student-weighted) mean of $\hat{\mu}_{jt}^{L,-\{t,t-1\}}$ and $\hat{\mu}_{jt}^{S,-\{t,t-1\}}$ across teachers in school s in grade g, respectively. We define the change in mean long- and short-run teacher VA from year $t-1$ to year t in grade g in school s as $\Delta Q_{sgt}^{k} = Q_{sgt}^{k} - Q_{sg,t-1}^{k}$ for $k \in \{L, S\}$. By leaving out both years t and $t-1$ when estimating VA, we ensure that the variation in ΔQ_{sgt}^{k} is driven by changes in the teaching staff rather than changes in VA estimates.[21] Leaving out two years eliminates the correlation between changes in mean test scores across cohorts t and $t-1$ and estimation error in ΔQ_{sgt}^{k}.

Let A_{sgt}^{*} denote the mean test scores, A_{ijt}^{*}, for students in school s in grade g in year t and define the change in test scores as $\Delta A_{sgt}^{*} = A_{sgt}^{*} - A_{sg,t-1}^{*}$. Similarly, let $A_{s,g+1,t+1}^{*}$ denote the mean test scores for students in school s in grade g in year t in the *following* year (and grade) and define the change in these subsequent test scores as $\Delta A_{s,g+1,t+1}^{*} = A_{s,g+1,t+1}^{*} - A_{s,g+1,t}^{*}$. We

[21] Part of the variation in ΔQ_{sgt}^{k} comes from drift because for a given teacher predicted VA will change over time because our forecast of VA varies across years. Because the degree of drift is small across a single year, however, drift has little impact on the results.

check whether our short- and long-run VA measures feature forecast bias by regressing changes in mean *current* and *subsequent* test scores across cohorts on changes in mean long- and short-run teacher VA:

$$\Delta A^*_{sgt} = a + b\Delta Q^k_{sgt} + \Delta\xi_{sgt}\,, \quad k \in \{L, S\}$$
$$\Delta A^*_{s,g+1,t+1} = a + b\Delta Q^k_{sgt} + \Delta\xi_{s,g+1,t+1}\,, \quad k \in \{L, S\}\,. \tag{5.1}$$

The change in test scores in equation (5.1) can be replaced by changes in high school outcome residuals to check for potential bias in our estimated impacts of long- and short-run VA on long-run outcomes.

Results for Test Scores: Figure 6 reports the changes in school-grade mean raw test scores across cohorts for the current period and the next four periods against changes in mean teacher long- and short-run VA (alongside changes in 'normal' VA), weighting by the number of students in each cell. Whiskers in the figure indicate 95% confidence intervals. Changes in the quality of the teaching staff strongly predict changes in test scores across cohorts in a school-grade cell. Indeed, the estimated coefficient on ΔQ^k_{sgt} for school-grade test scores in the current period (i.e., $t = 0$) is statistically indistinguishable from one for 'normal,' long-run, and short-run VA, although standard errors are higher than their non-experimental counterparts. Similarly, a point estimate of one cannot be rejected for long-run VA for school-grade test scores in the *subsequent* year (i.e., $t = 1$), in line with our non-experimental evidence. The point estimate on short-run VA on test scores in the *subsequent* year is statistically greater than zero, indicating some potential bias in the estimated effects of short-run VA teachers on subsequent scores.[22] Point estimates for test scores in future periods (i.e., $t > 1$) demonstrate that being assigned to high long-run VA teachers lead to large improvements, whereas assignment to high short-run VA teachers do not.

Results for High School Outcomes: Figure 7 visualizes how changes in mean school-grade-year short- and long-run teacher VA affects changes in the long-run outcomes of that school-grade-year. We create these figures by dividing normalized long- and short-run VA estimates into twenty equal-sized groups (vingtiles) and plotting the school-grade-year means of the long-run

[22] One reason this may occur is that schools adding new high short-run VA teachers may request these teachers focus their effort more on long-term knowledge acquisition, causing short-run VA changes from teacher turnover to lead to some improvements in future test scores.

outcome residuals defined by equation (4.2), Y_{it}, within each bin against the school-grade-year mean value of the VA estimate within each bin. Points estimates of equation (5.1) where school-grade-year long-run outcomes are used as the dependent variable are also reported in the figures. Figure 7 reinforces the non-experimental results in Figure 2: Long-run VA leads to dramatic improvements in long-run outcomes whereas short-run VA does not.

6 Policy Analysis

This section evaluates the benefits of using teacher long-run VA in terms of policy. We start by considering the implications for *who* is fired, then consider gains in policy effectiveness. We pay particular attention to a benchmark policy proposed by Hanushek (2009, 2011) and evaluated by Chetty et al. (2014b) that releases teachers in the bottom five percent of the VA distribution, given its prominence in the literature.

Who is Released: We start by looking at who is released when long-run VA is used in place of 'normal' VA. Figure 8 displays a scatter plot of teacher quality as measured by 'normal' and long-run VA in a given year. The dashed lines delineate teachers in the bottom five percent of the VA distribution for a given VA measure. We find that many teachers are released under 'normal' VA, but not long-run VA. Visually, these teachers are represented by the dots that fall in the first or fourth quadrant of the figure (as delineated by the dashed lines); for instance, those in the first quadrant are released if the 'normal' VA measure is used, but not if long-run VA were used instead. We find that about 70% of teachers released under the benchmark policy using normal VA are not released under a policy based on long-run VA.

Policy Efficiency: Since using long-run VA in place of 'normal' VA affects who is released, the quality of released teachers according to long-run VA will be lower when policymakers use long-run VA for high-stakes decision-making. Given that high long-run VA teachers increase long-run outcomes more than high 'normal' VA teachers, this will in turn drive higher policy gains.

We calculate the policy gains under policies that target the bottom five percent of the VA distribution according to 'normal' and long-run VA following the methodology in Chetty et al.

(2014b). In particular, we ignore general equilibrium effects and assume that both 'normal' and long-run teacher VA are normally distributed.[23] In addition, we assume that replacement teachers are of mean quality. We quantify the value of improving teacher quality according to each VA measure by calculating the gains in terms of long-run outcomes from selecting teachers based on their true normalized 'normal' and long-run VA. Given that teacher VA is not observed in practice, we then calculate the gains from selecting teachers based on VA estimates.

Our estimates from Section 4.2 of the impact of being assigned to a teacher whose 'normal' and long-run VA is one standard deviation higher in a single grade feed into our calculations of the long-run gains of replacing bottom five percent teachers with a teacher of mean quality. Doing so would raise a student's long-run outcomes by:

$$G^{\kappa} = \Delta m_{\sigma}^{\kappa} \times \rho^{\kappa}, \quad \kappa \in \{L, N\}, \tag{6.1}$$

where ρ^{κ} is our estimated impact of a one standard deviation higher teacher, with the superscripts N and L representing 'normal' and long-run VA, respectively. $\Delta m_{\sigma}^{\kappa}$ represents the average improvement in VA (measured in terms of test scores) of the policy. Under normality, the expected value of VA conditional on being a teacher in the bottom five percent is given by $\mathbb{E}[\alpha|\alpha < \Phi^{-1}(0.05)]$, where $\Phi(\cdot)$ is the cdf of the normal distribution. We therefore have that $\Delta m_{\sigma}^{PEB} = 2.06$, following from the normality assumption.

Table 6 reports the policy gains in terms of twelve high school outcomes. The policy gains from using long-run VA in place of 'normal' VA is substantial: point estimates indicate two- to three-fold increases in high school outcomes. If we were to use estimates from Chetty et al. (2014b) on earnings, this suggests that the total net present value earnings impact from replacing a bottom five percent teacher according to long-run VA is $800,000. In contrast, replacing a bottom five percent teacher according to 'normal' VA only raises the total net present value earnings by $400,000.

Selection on Estimated VA: In practice, teachers can only be selected on the basis of estimated VA $\hat{m}_{jt}$. This reduces the gains from selection for two reasons: (i) estimation error in

[23] We will relax the normality assumption in a later draft using NPEB as in Gilraine et al. (2020). We expect the *differences* in policy gains to be similar under NPEB as we are calculating differences rather than levels and so while each VA distribution may depart from normality we expect the departures to be highly-correlated.

VA, and (ii) drift in teacher quality over time. Given that long-run VA features more drift and estimation error (see Table 2 and A.2), we expect that policy gains when using estimated long-run VA will underperform their true VA benchmarks substantially more than those of normal VA.

Replacing bottom five percent teachers in year $n+1$ according to their estimated VA using the preceding $t = 1, ..., n$ years of data raises a student's long-run outcomes by:

$$G^{\kappa}(n) = \mathbb{E}\left[m^{\kappa}_{j,n+1} | \hat{m}^{\kappa}_{j,n+1} < F^{-1}_{\hat{m}^{\kappa}_{j,n+1}}(0.05)\right] \times \rho^{\kappa}\,, \quad \kappa \in \{L, N\}\,, \tag{6.2}$$

where $\mathbb{E}\left[m^{\kappa}_{j,n+1} | \hat{m}^{\kappa}_{j,n+1} < F^{-1}_{\hat{m}^{\kappa}_{j,n+1}}(0.05)\right]$ represents the expected value of teacher VA conditional on the teacher's estimated VA being in the bottom five percent. We follow Chetty et al. (2014b) and calculate the expected value using Monte Carlo simulations.[24] Note that we can only use $n-1$ of the n preceding years of data to estimate a teacher's long-run VA given that outcomes are only observed in year $t+1$.

Panel A of Figure 9 plots the mean gain per classroom of releasing bottom five percent teachers according to 'normal' and long-run VA in terms of two outcomes that are particularly salient to policymakers: high school graduation and SAT-taking rates (a measure of college interest). The gain from releasing teachers based on their true normal and long-run VA is shown by the horizontal lines in the figures. The gains from releasing teachers based on estimated VA are significantly smaller than the true VA benchmarks, particularly for long-run VA (as expected). Even so, releasing teachers based on estimated long-run VA substantially improves long-run outcomes relative to using normal VA, even when the policymaker can observe true normal VA. One drawback of long-run VA is that one can only use $n-1$ of the n years of data when estimating a teacher's long-run VA. In the figure we can see this additional year wait as the long-run VA series only begins after we have two years of data to estimate VA.

Once we have at least two years of data, the gains of using long-run VA over normal VA are

[24] To calculate the conditional expectation in equation (6.2) we start by using the variances and autocovariances reported in Tables 2 and A.2 to construct the variance-covariance matrix Σ_A. We set a drift limit of six periods and then simulate draws of average class scores from a normal distribution $N(0, \Sigma_A)$ for one million teachers and calculate $\hat{m}_{j,n+1}$ based on scores from the first n periods. The conditional expectation in equation (6.2) is then computed as the mean test score in year $n+1$ for teachers with $\hat{m}_{j,n+1}$ in the bottom five percent of the distribution. Our procedure is done separately for math and English and we report the average.

substantial. After three years of data, an additional 0.082 students per classroom graduates high school when a bottom five percent teacher is released according to long-run VA in comparison to normal VA, a 66% increase relative to the policy gain using normal VA of 0.125. For SAT-taking the gains from using long-run over normal VA are even greater.

The values in Panel A of Figure 9 represent the gains in the first year after the release of teachers. The drift in teacher quality, however, will causes the gains for students in subsequent years to decline. Given the higher levels of drift in long-run relative to normal VA, this could potentially drive down the gains from using long-run VA.

Panel B of Figure 9 reports the gains per class in future school years from releasing teachers using three year of data.[25] As expected, the higher levels of drift in long-run relative to normal VA erode some of the gains in subsequent school years, as teachers released according to long-run VA are more likely to have reverted back towards the mean than those released according to normal VA. The gains from releasing teachers based off long-run VA remain substantial: Releasing teachers based on long-run VA estimates after three years of data leads to an additional 0.2 students graduating and 1.4 students taking the SAT over the subsequent 10 years in comparison to releasing teachers based on normal VA.

7 Conclusion

This paper decomposes VA into its two components: the portion that persists to the next period, long-run VA, and the portion that completely fades out, short-run VA. While we find that both portions of VA affect contemporaneous test scores, only the long-run portion influences future test scores and long-run outcomes. Since more than half of the variation in standard teacher VA comes from the short-run component, the use of long-run VA is able to better measure teachers' true contributions to students' later life success.

Given that the goal of educators and policymakers is to improve the lifelong success of their students, long-run VA has the potential to greatly improve policy efficiency. We show that targeting long-run VA raises the efficacy of policies considerably: policy gains in terms of our

[25] We estimate the impacts of releasing a bottom five percent teacher in year n on the outcomes of students in a subsequent school year $n+m$ using the following equation: $G^{\kappa}(m,n) = \mathbb{E}\left[m^{\kappa}_{j,n+m} | \hat{m}^{\kappa}_{j,n+1} < F^{-1}_{\hat{m}^{\kappa}_{j,n+1}}(0.05)\right] \times \rho^{\kappa}$, $\kappa \in \{L, N\}$.

high school outcomes, on average, increase by twofold when using long-run VA instead of the standard VA measure under our benchmark policy where the bottom five percent of teachers according to VA are released. The higher noise and drift in long-run VA estimates attenuate these gains somewhat, but the benefits of using long-run VA remain substantial.

Our results shed new light on the importance of incorporating non-contemporaneous test score measures in assessing teachers. While contemporaneous test score measures have the advantage of being readily available, they are unable to capture teaching of foundational learning principles and behavioural improvements that are not initially measured by tests, but are important for future learning.

Incentivizing teachers also needs to take into account that teachers may be able to substitute between methods that encourage long- or short-term success Macartney et al. (2018). The use of future test score performance to evaluate teachers should help in this dimension as it incentivizes teachers to teach to *next year's* test rather than the current test. While teacher's can likely raise current test scores through unproductive teaching methods such as rote memorization, it is harder to use such techniques to increase student test scores in the following year. Instead, teachers should find it easiest to raise next year's test score by teaching foundational concepts and behaviours that help students succeed in the future – exactly the behavior we want to encourage to ensure students' lifelong success. Investigating teachers' response to incentive schemes in terms of substitution between long- and short-run VA is something that we are exploring in related work in designing optimal dynamic incentive schemes (Macartney, 2016; Gilraine, 2018).

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, 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.

Gilraine, Michael (2018), "School accountability and the dynamics of human capital formation.", URL https://docs.google.com/a/nyu.edu/viewer?a=v&pid=sites&srcid=bnl1LmVkdXxnaWxyYWluZXxneDo0ZWFhNjYzMDhlZDRlMWY2. Unpublished.

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.

Gourieroux, Christian and Joann Jasiak (2020), "Dynamic deconvolution of (sub)independent autoregressive sources." URL http://www.jjstats.com/papers/dynamdec.pdf. Unpublished.

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.

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.

Jacob, Brian A. and Lars Lefgren (2008), "Can principals identify effective teachers? Evidence on subjective performance evaluation in education." *Journal of Labor Economics*, 26, 101–136.

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., Jonah E. Rockoff, and Douglas O. Staiger (2008), "What does certification tell us about teacher effectiveness? Evidence from New York City." *Economics of Education Review*, 27, 615–631.

Kane, Thomas J. 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.

Koretz, Daniel M. (2002), "Limitations in the use of achievement tests as measures of educators' productivity." *Journal of Human Resources*, 752–777.

Macartney, Hugh (2016), "The dynamic effects of educational accountability." *Journal of Labor Economics*, 34, 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.

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.

Rockoff, Jonah E. (2004), "The impact of individual teachers on student achievement: Evidence from panel data." *American Economic Review*, 94, 247–252.

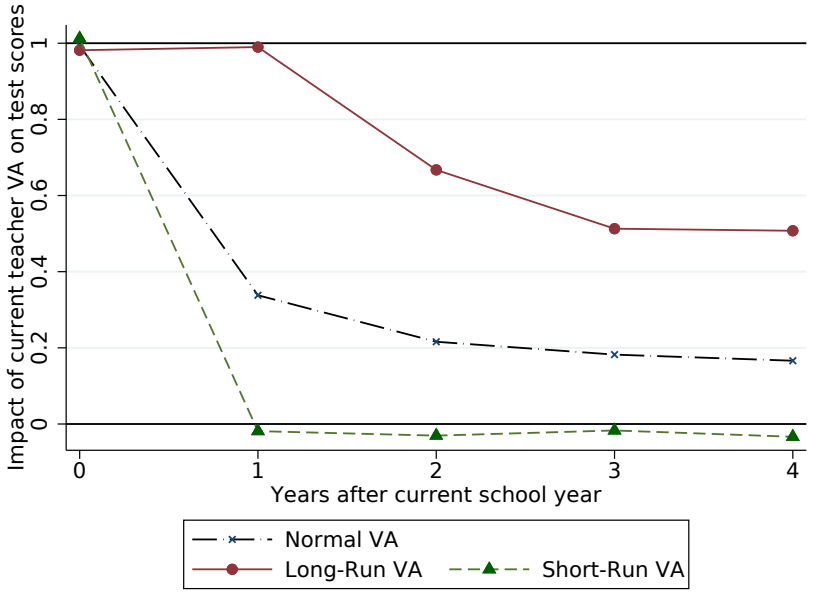
Rose, Evan K., Jonathan Schellenberg, and Yotam Shem-Tov (2019), "The effects of teacher quality on criminal behavior." URL https://drive.google.com/file/d/1agkUuMjtPIPoQlgQEel3tVVofs2WFVsA/view. Unpublished.

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.

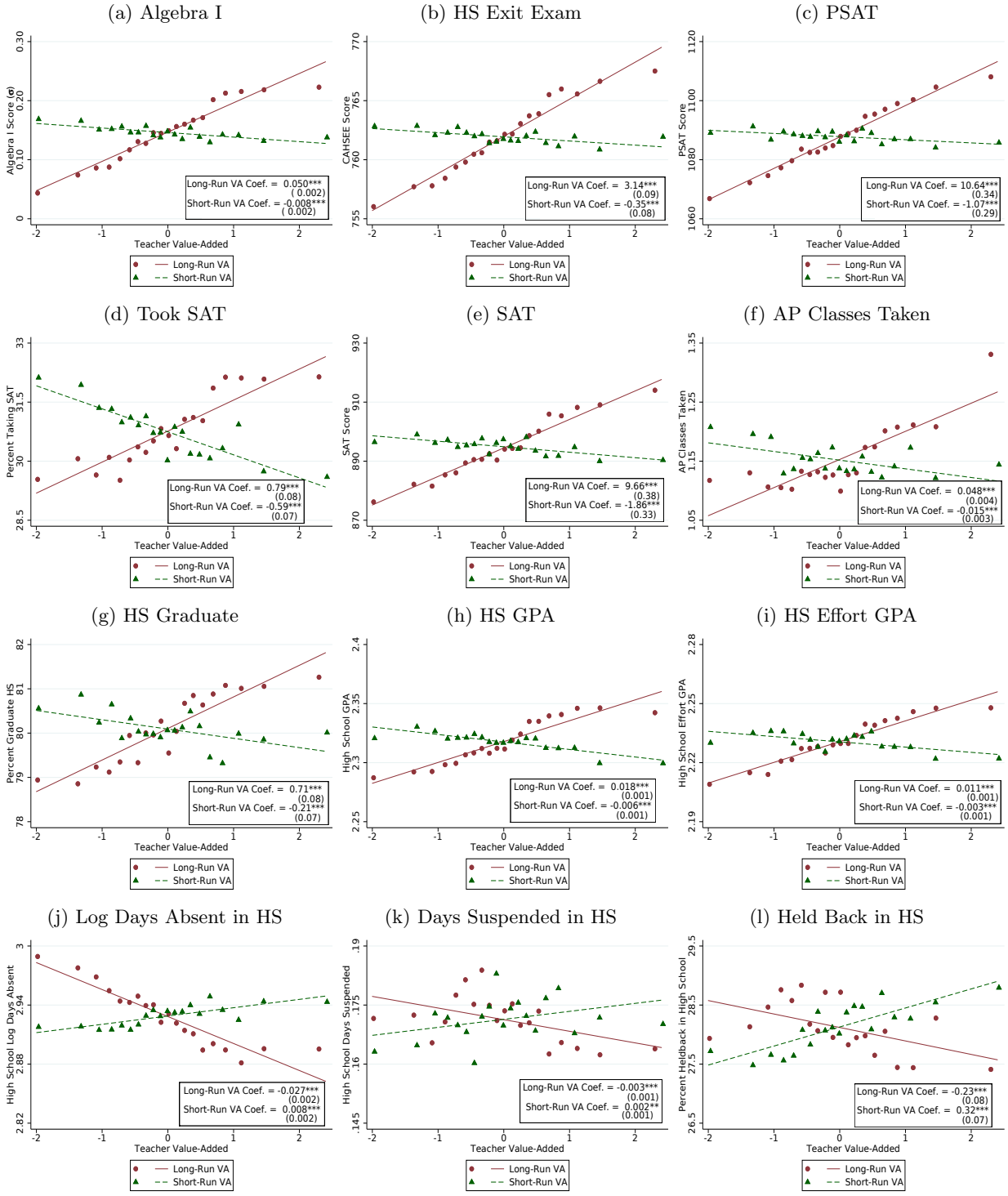
Stepner, Michael (2013), "Vam: Stata module to compute teacher value-added measures." URL http://fmwww.bc.edu/RePEc/bocode/v/vam.ado.

Figure 1: Effects of Short- and Long-Run Value-Added on Future Test Scores



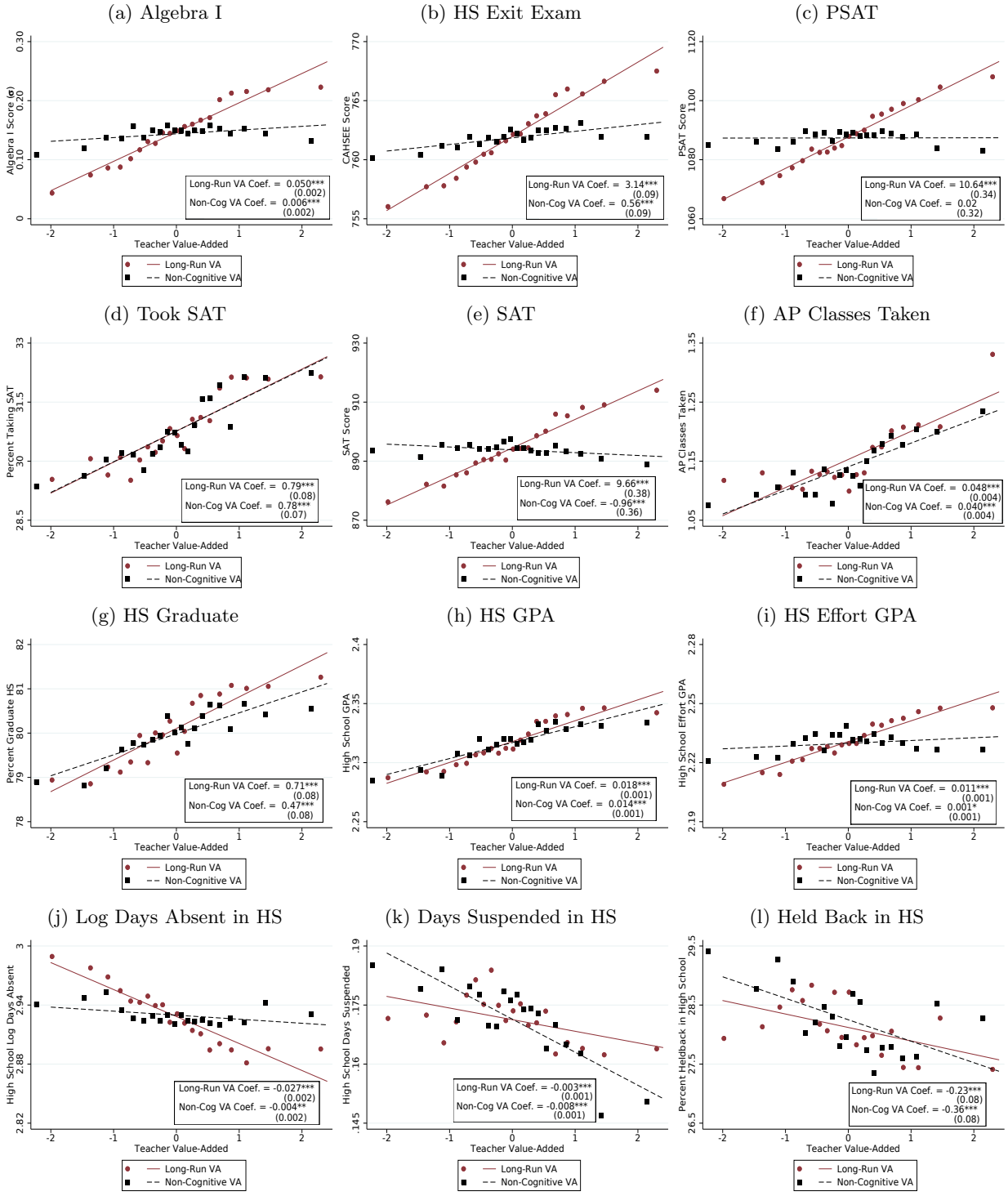
Notes: This figure shows the effect of teacher VA on test scores at the end of the current and subsequent school years for short- and long-run VA alongside 'normal' VA. The figure is constructed by regressing residualized end-of-grade math and English test scores in year $t + s$ on the teacher VA measure in year t in that subject. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (2.3). The coefficients and standard errors of the point estimates underlying the figure are reported in Table 5.

Figure 2: Effect of Short- and Long-Run Value-Added on Long-Run Outcomes



Notes: This figure shows the effect of teacher VA on long-run outcomes for short- and long-run value-added. Both VA measures are normalized by the standard deviation of their respective teacher effects. Each figure is constructed in three steps: (i) residualize the long-run outcome with respect to our control vector using within-teacher variation as described by equation (4.1), (ii) divide the normalized long- and short-run VA measures, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the long-run outcome in the estimation sample to facilitate interpretation of the scale. A line of best fit is then superimposed. Figures also report estimates of ρ from equation (4.3), which represent the effect of being assigned to a teacher whose long- or short-run VA is one standard deviation higher in a single grade on the long-run outcomes, along with its standard errors. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

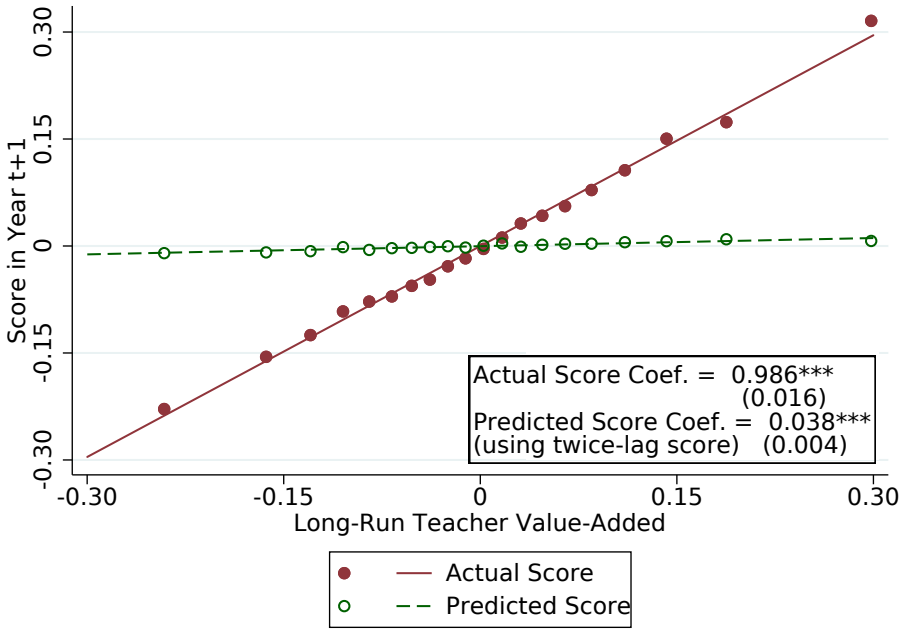
Figure 3: Effect of Non-Cognitive and Long-Run Value-Added on Long-Run Outcomes



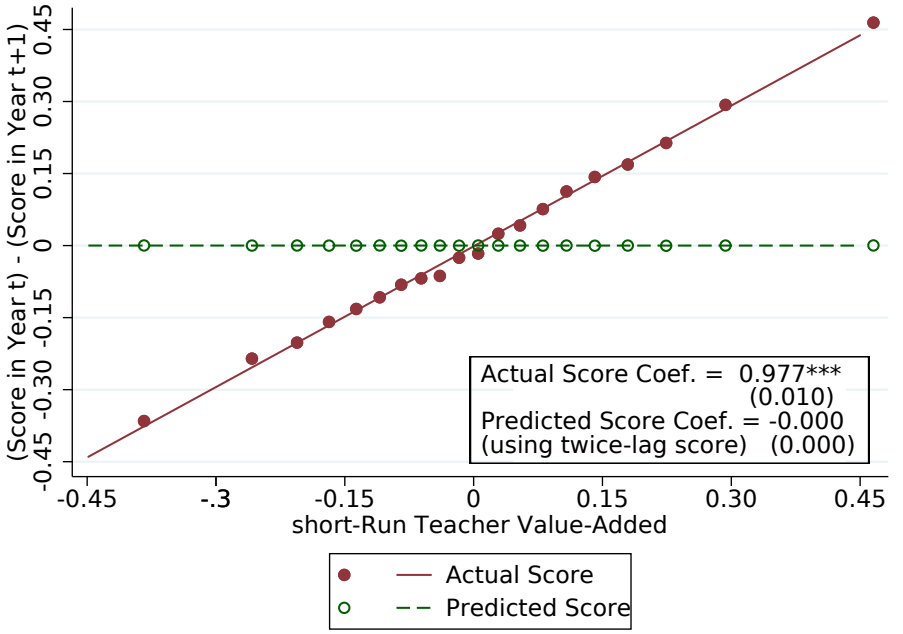
Notes: This figure shows the effect of teacher VA on long-run outcomes for non-cognitive and long-run value-added. Both VA measures are normalized by the standard deviation of their respective teacher effects. Each figure is constructed in three steps: (i) residualize the long-run outcome with respect to our control vector using within-teacher variation as described by equation (4.1), (ii) divide the normalized long- and short-run VA measures, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the long-run outcome in the estimation sample to facilitate interpretation of the scale. A line of best fit is then superimposed. Figures also report estimates of ρ from equation (4.3), which represent the effect of being assigned to a teacher whose non-cognitive or long-run VA is one standard deviation higher in a single grade on the long-run outcomes, along with its standard errors. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 4: Effects of Short- and Long-Run VA on Actual and Predicted Scores

(a) Long-Run VA



(b) Short-Run VA

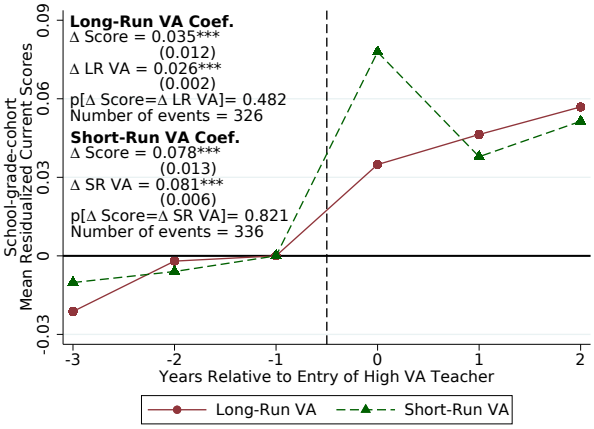


Notes: These figures assess whether students sort on variables that predict test score residuals but are omitted from the long- and short-run VA model. We predict scores based on twice-lagged outcomes that include: math scores, English scores, suspensions, absences, held back, and GPA. Third grade students are eliminated from the sample given the need for twice-lagged outcomes. These figures are constructed in three steps: (i) residualize twice-lagged outcomes $\mathbf{Y}_{it}^{-2}$ by regressing each element of $\mathbf{Y}_{it}^{*-2}$ on our control vector X_{ijt} and teacher fixed effects, as in equation (2.3), (ii) regress residualized test scores A_{ijt} on $\mathbf{Y}_{it}^{-2}$, again including teacher fixed effects, and calculate predicted values $A_{ijt}^{Y} = \hat{\boldsymbol{\rho}}\mathbf{Y}_{it}^{-2}$, (iii) divide the long- or short-run VA estimates into twenty equal-sized groups (vingtiles) and plot the means of the residuals within each bin against the mean value of the VA estimate within each bin. The actual score is also provided which nonparametrically plots test score residuals against the VA estimates. Point estimates for the slope of this line are close to one, indicating forecast unbiasedness. The lines indicate the line of best fit estimated on the underlying micro data using OLS. The coefficients show the estimated slope of the best-fit line with standard errors clustered at the school-cohort level reported in parentheses. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

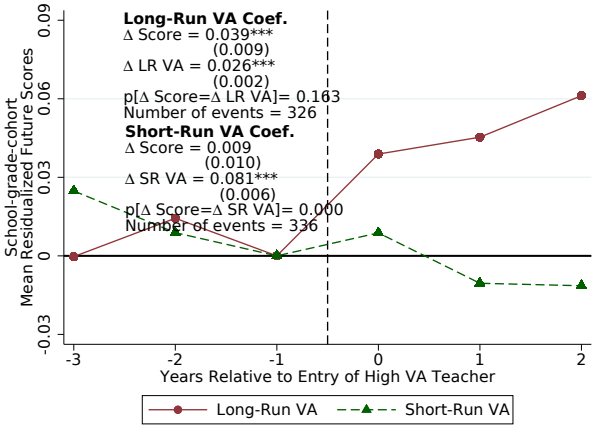
Figure 5: Impacts of High and Low Long- and Short-Run VA Teacher Entry on Test Scores

Panel A: Impacts of **High** VA Teacher Entry on Current and Future Scores

(a) Current Test Scores

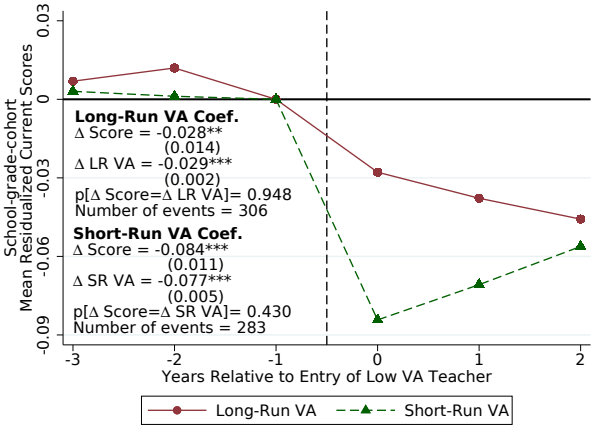


(b) Future Test Scores (i.e., Scores in $t+1$)

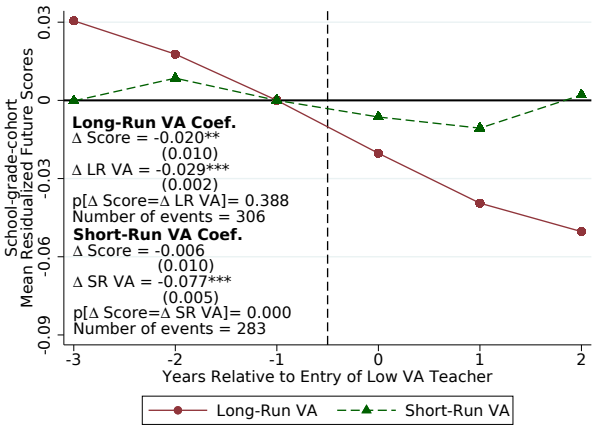


Panel B: Impacts of **Low** VA Teacher Entry on Current and Future Scores

(c) Current Test Scores

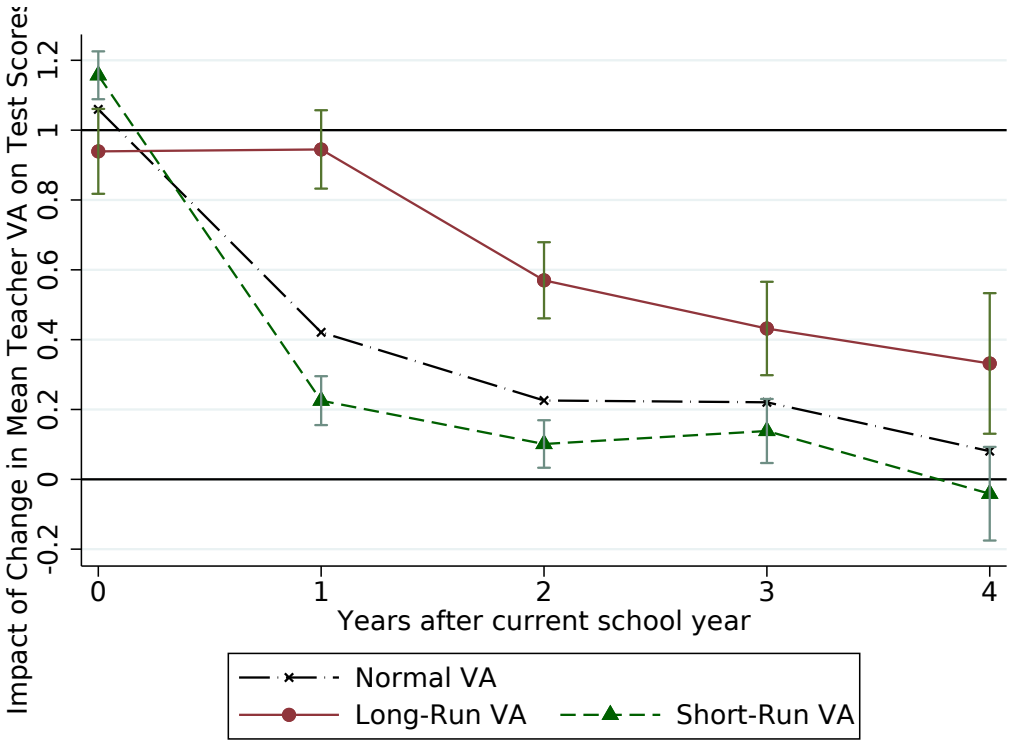


(d) Future Test Scores (i.e., Scores in $t+1$)



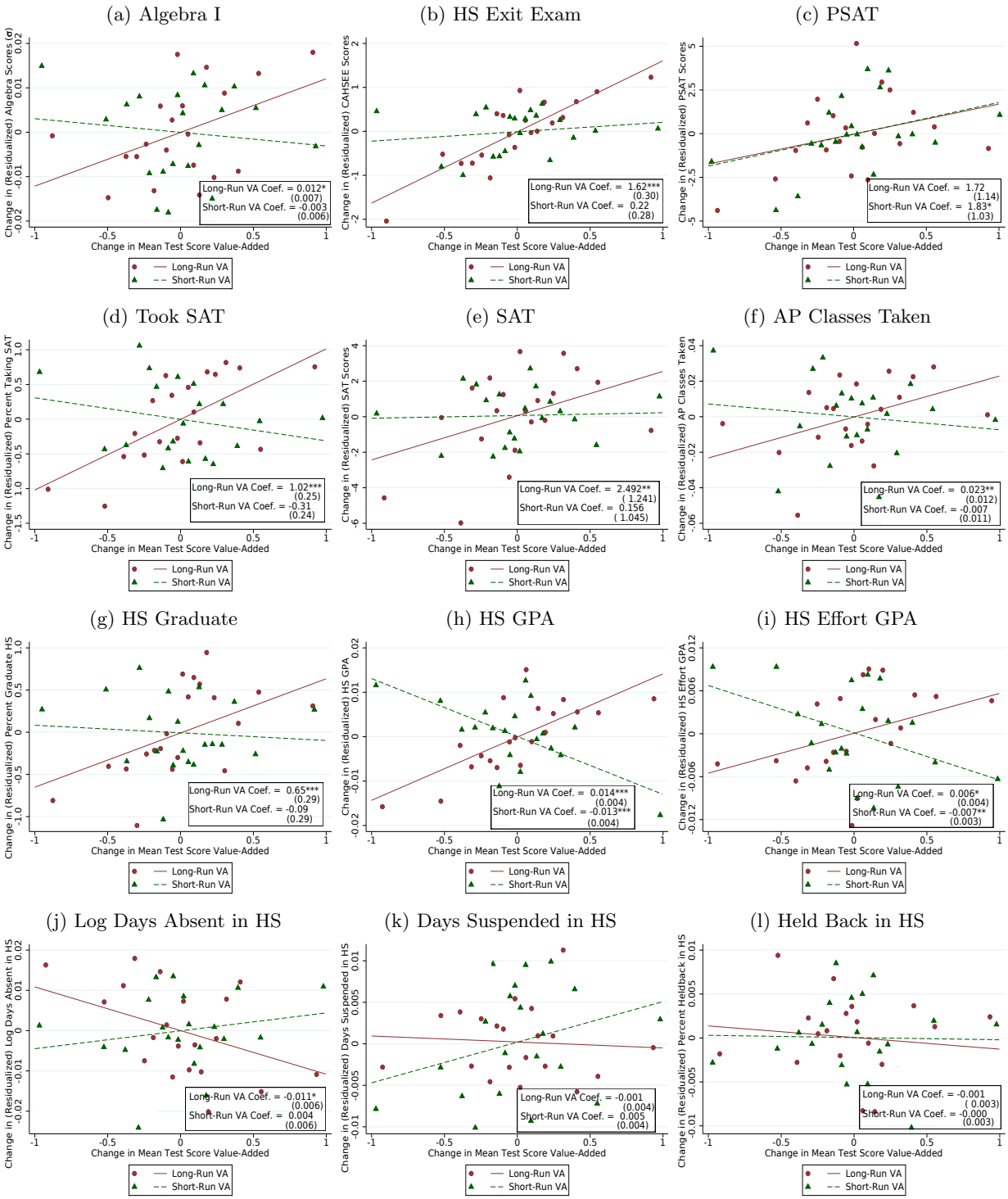
Notes: These figures plot event studies of test score changes by cohort as teachers enter a school-grade-subject cell at event-time 0. Panel A does so for high VA teachers (teachers with VA in the top 5% of the distribution), while Panel B does so for low VA teachers (teachers with VA in the bottom 5% of the distribution). Each figure then plots a series whereby VA is measured as long- and short-run VA. The left-hand side figures plotting 'current test scores' plot the residualized mean test scores for the school-grade-cohort this year, while right-hand side figures plot mean residualized 'future test scores' that the school-grade-cohort achieves in the *following* year. To construct each panel, we first identify the set of teachers who entered a school-grade-subject cell and were not teaching at the same school in the prior period (i.e., we do not use within-school switcher variation) and define event time as the school year relative to the year of entry. We then estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$, including fixed effects for the subsequent teacher. We then identify the subset of teachers with VA estimates in the top 5% of the distribution among entering teachers and then plot mean school-cohort current or future resdidualized test scores in the relevant school-grade-subject cell for the event years before and after the entry of such a teacher. We normalize the score changes to zero at event year -1 and include year fixed effects to eliminate secular time trends. 'Δ Score' reports the change in current or future test scores in the period after the teacher entered (period 0) relative to the period before (period -1). 'Δ VA' reports the change in long- or short-run VA (at the school-grade-cohort level) in the period after the teacher entered (period 0) relative to the period before (period -1). The p-value of a test of whether these coefficients are equal is then reported. Figure A.3 reports results from similar event studies that leverage teacher exit.

Figure 6: Quasi-Experimental Estimates of Short- and Long-Run Value-Added on Future Test Scores



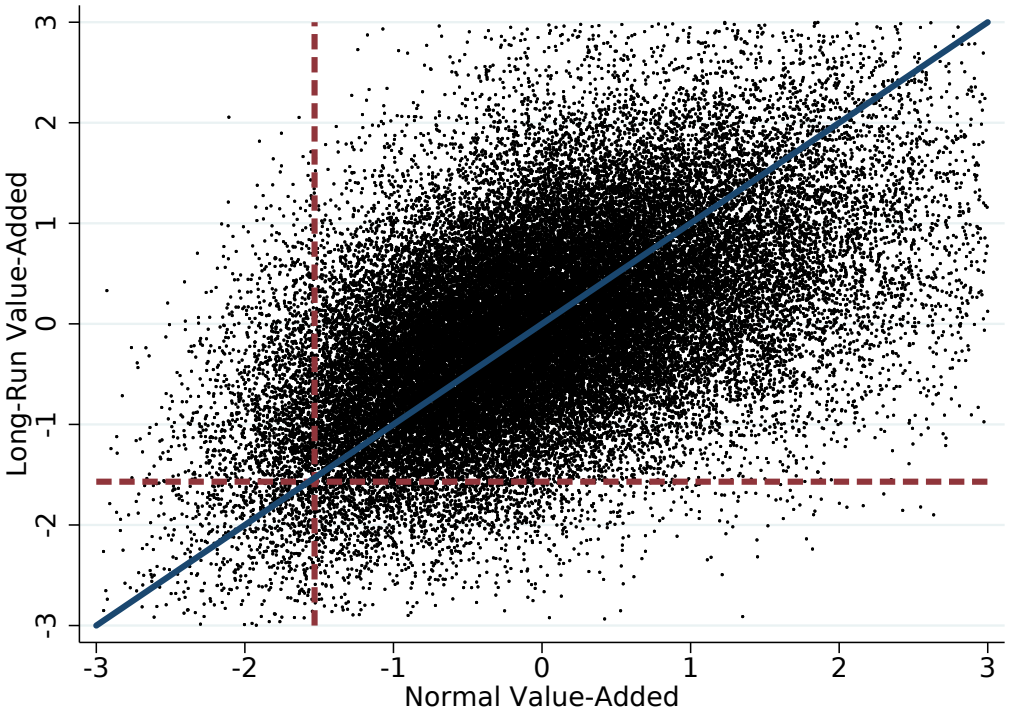
Notes: This figure reports the changes in school-grade mean raw test scores across cohorts for the current period and the next four periods against changes in mean teacher 'normal,' long-, and short-run VA as described in equation (5.1). Effectively, this figure recreates Figure 1 using quasi-experimental variation from staffing changes to estimate the impact of teachers' 'normal,' long-, and short-run VA on current and future test scores. The whiskers represent 95% confidence intervals with standard errors clustered at the school-cohort level.

Figure 7: Quasi-Experimental Estimates of Effect of Short- and Long-Run Value-Added on Long-Run Outcomes



Notes: This figure visualizes how changes in mean school-grade-year short- and long-run teacher VA affects changes in the long-run outcomes of that school-grade-year. We create these figures by dividing normalized long- and short-run VA estimates into twenty equal-sized groups (vingtiles) and plotting the school-grade-year means of the long-run outcome residuals defined by equation (4.2), Y_{it}, within each bin against the school-grade-year mean value of the VA estimate within each bin. Effectively, this figure recreates Figure 2 using quasi-experimental variation from staffing changes to estimate the impact of teachers' long- and short-run VA on high school outcomes. Points estimates of equation (5.1) where school-grade-year long-run outcomes are used as the dependent variable are also reported in the figures alongside standard errors clustered at the school-cohort level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure 8: Two-Dimensional Cross Teacher Value-Added Plots

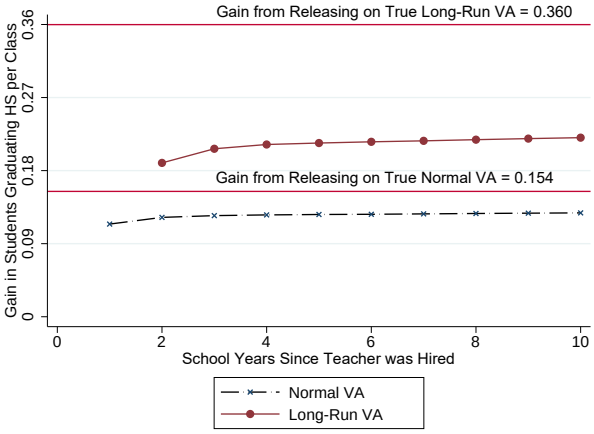


Notes: This figure plots 'normal' and long-run VA estimates with each dot representing a teacher in a given year. The dashed vertical line delineates bottom five percent teachers according to 'normal' VA with teachers left of the line being in the bottom five percent. Similarly, the horizontal line delineates bottom five percent teachers according to long-run VA with teachers below the line being in the bottom five percent. The diagonal blue line represent the 45 degree line where the two VA measures agree. Teachers in the first and fourth quadrants (as delineated by the dashed lines) are released under one VA measure, but not the other. For instance, teachers in the first quadrant are released if the 'normal' VA measure is used, but not if long-run VA were used instead.

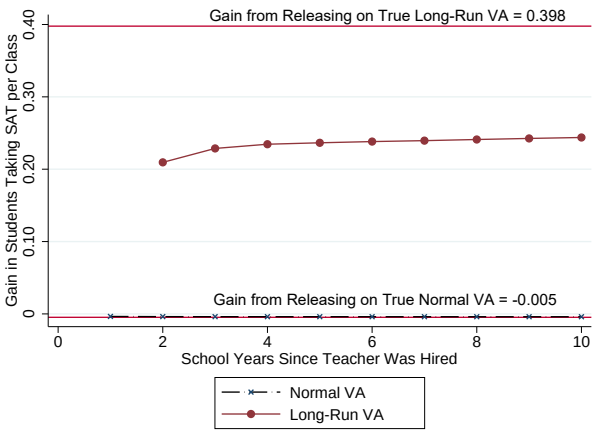
Figure 9: Impacts of Releasing Low Value-Added Teachers

Panel A: Impacts in First Year After Release

(a) High School Graduation

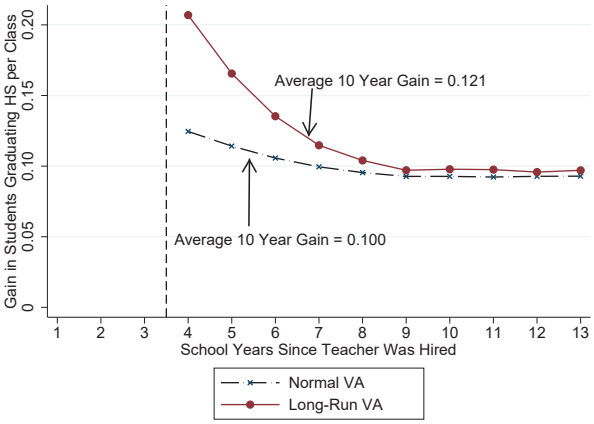


(b) SAT Taking

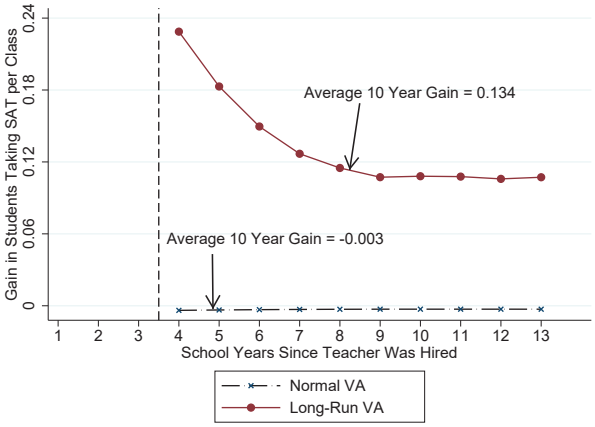


Panel B: Impact Over Time

(c) High School Graduation



(d) SAT Taking



Notes: This figure calculates the impact of replacing teachers with normal or long-run VA in the bottom 5% with teachers of average quality on the number of high school graduates and SAT takers in a single classroom of average size (24.4 students). The horizontal lines in Panel A show the hypothetical gain in the current school year from releasing the bottom 5% teachers according to their true normal (lower line) and long-run (upper line) VA. The series in each figure then plots the gains from releasing teachers based on estimated VA versus the number of years of prior data used to estimate VA. Panel B shows the gains of releasing teachers based on normal and long-run VA estimates in subsequent school years based on their estimated VA four years after they were hired. We also report the mean gain over the first 10 years after the teacher is released. The first point in the Panel B figures equals the point in the Panel A figures at school year 3 by construction. All values in these figures use mean values for math and English teachers, which are calculated separately, and are based off the estimated increase in graduation and SAT taking in Table 6.

Table 1: Summary Statistics

	Full Sample[1] (1)	Value-Added Sample[2] (2)	Value-Added Sample Linked to High School[3] (3)
Mean of Outcomes and Student Characteristics			
Cognitive Outcomes:			
Math Score (σ)[4]	0.03	0.06	0.03
Reading Score (σ)[4]	0.03	0.05	0.01
Non-Cognitive Outcomes:			
Log Days Absent	1.50	1.51	1.49
GPA	2.88	2.90	2.87
'Effort' GPA	3.14	3.16	3.12
% Suspended	2.26	2.19	2.48
% Repeating Grade	0.65	0.49	0.51
Demographics:			
% Hispanic	74.3	75.6	79.0
% Black	10.1	9.2	8.3
% White	8.9	8.6	6.6
% Asian	4.3	4.3	3.6
% Free or Reduced Price Lunch	69.2	70.8	72.7
% English Learners	30.2	30.7	31.3
Parental Education:[5]			
% High School Dropout	35.6	35.9	40.1
% High School Graduate	27.1	27.4	28.4
% College Graduate	19.6	19.1	25.9
# of Students	649,694	552,517	261,096
# of Teachers	15,155	12,975	12,025
Observations (student-year)	1,452,367	1,190,019	671,076

Notes:
[1] Data coverage: third through fifth grades from 2003-04 through 2011-12.
[2] VA sample restrictions: must be assigned to valid teacher, be in a class with between seven and forty students, and have valid current and lagged math scores.
[3] We define as linked to a high school outcome if we observe the student in the high school transcript data at any point. We require that the student's cohort to reach the end of eleventh grade by 2016-17, which eliminates the 2009-10 through 2011-12 third grade cohorts, the 2010-11 and 2011-12 fourth grade cohort and the 2011-12 fifth grade cohort. The number of observations is identical to the number of students for whom we infer high school suspension outcomes (see Table A.1).
[4] Standardized test scores are not exactly zero as standardization occurs at the test-level and we drop students whose grade cannot be determined. These students, who either have missing grade data or are coded as 'ungraded,' tend to be lower-performing.
[5] The omitted category is 'Some College,' and 'College Graduate' also incorporates those with graduate school degrees. Thirty percent of observations are missing parental education data or have parental education recorded as "Decline to Answer."

Table 2: Autocorrelation and Variance Estimates of Long- and Short-Run VA (Mathematics)

	Normal VA	Long-Run VA		Short-Run VA	
	(1)	(1)	(2)	(1)	(2)
Autocorrelation Vector					
Lag 1	0.66	0.39	0.33	0.50	0.59
Lag 2	0.61	0.30	0.26	0.44	0.56
Lag 3	0.57	0.25	0.23	0.40	0.54
Lag 4	0.53	0.20	0.20	0.37	0.52
Lag 5	0.51	0.17	0.21	0.34	0.50
Lag ≥ 6	0.48	0.14	0.20	0.32	0.49
Within-year variance components					
Total SD	0.597	0.629	0.583	0.643	0.611
Individual-level SD	0.507	0.582	0.559	0.569	0.548
Class + teacher level SD	0.315	0.238	0.165	0.300	0.271
Estimates of teacher SD					
Lower bound based on lag 1	0.270	0.169	0.117	0.230	0.227
Quadratic estimate	0.282	0.194	0.128	0.245	0.233
Future Year Teacher FE	No	No	Yes	No	Yes
Student-Year Observations	1,185,181	1,060,374	1,060,374	1,060,374	1,060,374

Notes: This table gives the drift autocorrelation estimates across years for the same teacher used to compute normal, short- and long-run VA estimates in mathematics. It does so for both when we do not control for the future teacher (columns (1)) and when we include future teacher fixed effects (columns (2)). It also reports the raw standard deviation of test score residuals and decomposes 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). In particular, we use the square root of the autocovariance across classrooms at a one year lag to estimate a lower bound and report an estimate of the standard deviation of teacher effects constructed by regressing the log of first seven autocovariances on the time lag and time lag squared and extrapolating to 0 to estimate the within-year covariance. Table A.2 reports analogous model estimates for VA using English scores as the outcome.

Table 3: Correlation of Teacher Value-Added Measures

VA Measure	'Normal' VA	Long-Run VA	Short-Run VA	Non-Cognitive VA
'Normal' VA	1			
Long-Run VA	0.524	1		
Short-Run VA	0.745	-0.164	1	
Non-Cognitive VA	0.113	0.180	-0.010	1

Notes: This table reports the correlations between our various value-added measures. Each VA measure is constructed as described in Section 3.2. In particular, our three test score VA measures ('normal,' long- and short-run VA) combine our math and English VA estimates for all of the VA measures, giving each subject equal weight (i.e., $VA_{test} = \frac{1}{2}VA_{math} + \frac{1}{2}VA_{English}$). Parameter estimates for the math and English VA models are reported in Tables 2 and A.2, respectively. The non-cognitive VA index is computed using VA for suspensions, log days absent, GPA, and not progressing to the next grade on time (i.e., held back). We compute the index by summing the standardized value-added variables, recoded so each has the same expected sign, and then standardizing the resulting index to be mean zero, standard deviation one. Table A.3 reports the full correlation matrix between all the components that make up the various VA measures.

Table 4: Impacts of Teacher Value-Added Measures on High School Outcomes

Outcome:	Algebra Score (σ) (1)	HS Exit Exam (2)	PSAT Score (3)	Took SAT (%) (4)	SAT Score (5)	# AP Courses (6)
Sample Mean	0.145	762.0	1087.5	30.6	893.8	1.14
(s.d)	(0.782)	(41.4)	(153.3)	(43.8)	(119.2)	(1.76)
Panel A. Test-Based High School Outcomes						
Long-Run VA	0.048***	3.08***	10.83***	0.53***	9.75***	0.041***
(s.e.)	(0.002)	(0.10)	(0.35)	(0.08)	(0.40)	(0.004)
Short-Run VA	0.001	0.17*	0.73**	-0.45***	-0.10	-0.007**
(s.e.)	(0.002)	(0.09)	(0.30)	(0.07)	(0.34)	(0.003)
Non-Cognitive VA	-0.001	0.10	-1.80***	0.69***	-2.61***	0.034***
(s.e.)	(0.002)	(0.09)	(0.33)	(0.08)	(0.36)	(0.004)
Observations	402,920	387,819	465,931	649,188	203,977	490,225

Outcome: VA Measure	Graduated HS (%) (7)	HS GPA (8)	HS Effort GPA (9)	Log Days Absent (10)	Days Suspended (11)	Held Back in HS (%) (12)
Sample Mean	80.1	2.32	2.23	2.93	0.171	28.2
(s.d)	(37.1)	(0.78)	(0.43)	(1.04)	(0.718)	(41.9)
Panel B. Behaviour-Based High School Outcomes						
Long-Run VA	0.63***	0.015***	0.010***	-0.026***	-0.001	-0.10
(s.e.)	(0.09)	(0.001)	(0.001)	(0.002)	(0.001)	(0.08)
Short-Run VA	-0.08	-0.003**	-0.001	0.003*	0.001	0.25***
(s.e.)	(0.07)	(0.001)	(0.001)	(0.002)	(0.001)	(0.07)
Non-Cognitive VA	0.37***	0.011***	-0.000	0.000	-0.008***	-0.35***
(s.e.)	(0.08)	(0.001)	(0.001)	(0.002)	(0.001)	(0.08)
Observations	303,635	534,413	458,932	555,201	567,867	523,770

Notes: This table reports the effect of a standard deviation increase in long-run, short-run, and non-cognitive VA on students' residualized long-run outcomes as described by equation (4.4) to check whether each VA measure independently affects long-run outcomes. Point estimates for the univariate effect of each of these VA measures are reported in Figures 2 and 3. Standard errors are clustered by student and class. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Table 5: Impacts of Teacher Value-Added Measures on Current and Future Test Scores

Period:	t	$t+1$	$t+2$	$t+3$	$t+4$
	(1)	(2)	(3)	(4)	(5)
'Normal' VA	0.994	0.338	0.216	0.182	0.166
(s.e.)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
[95% CI]	[0.985,1.002]	[0.328,0.348]	[0.206,0.226]	[0.172,0.193]	[0.154,0.179]
Observations	2,008,520	2,008,520	1,634,611	1,253,680	908,558
Long-Run VA	0.982	0.990	0.667	0.513	0.507
(s.e.)	(0.008)	(0.009)	(0.009)	(0.010)	(0.012)
[95% CI]	[0.966,0.998]	[0.973,1.007]	[0.649,0.685]	[0.494,0.532]	[0.484,0.531]
Short-Run VA	1.012	-0.018	-0.031	-0.017	-0.033
(s.e.)	(0.006)	(0.007)	(0.006)	(0.007)	(0.008)
[95% CI]	[1.000,1.023]	[-0.031,-0.004]	[-0.043,-0.018]	[-0.030,0.003]	[-0.049,-0.017]
Observations (Long- and Short-Run VA)	1,999,996	1,999,996	1,604,352	1,225,888	883,174

Notes: This table reports effect of teacher VA on test scores at the end of the current and subsequent school years. Point estimates are from a regression that regresses residualized end-of-grade math and English test scores in year $t+s$ on the teacher VA measure in year t in that subject. A subject fixed effect is also included. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (2.3). The table reports results from our three test score VA measures: normal VA, long-run VA, and short-run VA. The number of observations is the same for long- and short-run VA. Our test score based VA measures are not normalized here to check for forecast unbiasedness. Forecast unbiasedness in these measures occurs if a point estimate of one cannot be rejected for either 'normal' VA, long-run VA, short-run VA in period t *and* a point estimate of one and zero cannot be rejected for long- and short-run VA in period $t+1$, respectively. Point estimates for 'normal,' long-, and short-run VA are also plotted in Figure 1. Standard errors are two-way clustered by student and classroom. 95% confidence intervals are also reported.

Table 6: Policy Gains of Benchmark Policy Targeting Based on True 'Normal' vs. Long-Run VA

Test-Based Outcome:	Algebra Score (σ) (1)	HS Exit Exam (2)	PSAT Score (3)	Took SAT (%) (4)	SAT Score (5)	# AP Courses (6)
Sample Mean	0.150	762.4	1087.2	29.0	896.2	1.15
Average Change in VA of Released Teachers (Δm_σ)	2.06	2.06	2.06	2.06	2.06	2.06
Panel A. 'Normal' Value-Added						
Benefit ($\hat{\rho}^N$)	0.026	1.73	5.96	-0.03	4.81	0.016
Gain of Releasing Bottom 5% (G^N)	0.054	3.56	12.30	-0.06	9.92	0.033
Panel B. Long-Run Value-Added						
Benefit ($\hat{\rho}^L$)	0.048	3.00	10.00	0.58	9.08	0.043
Gain of Releasing Bottom 5% (G^L)	0.100	6.20	20.63	1.20	18.73	0.089
Policy Gain Increase from Using Long-Run VA (%)	185.2	174.2	67.8	∞	88.8	168.8

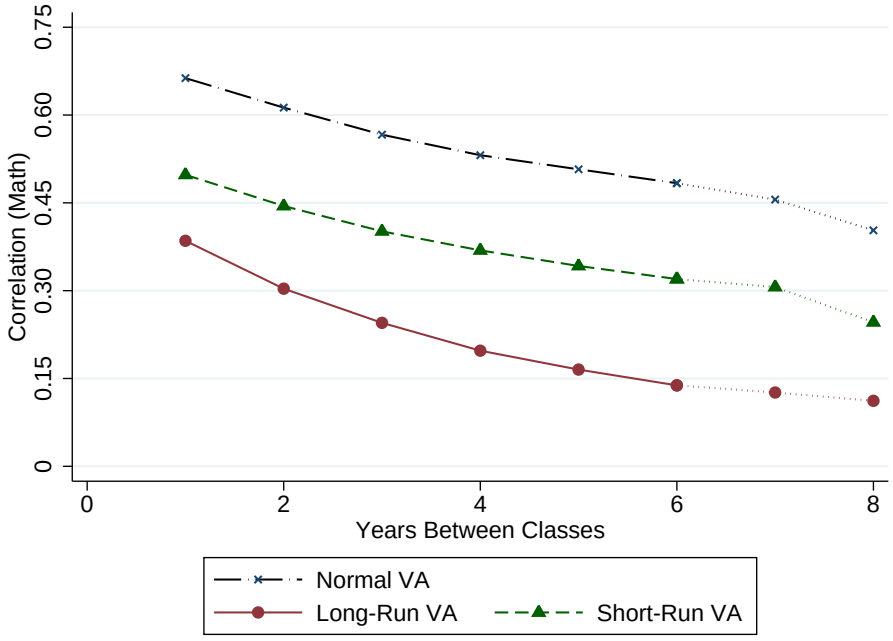
Behavioral-based Outcome: VA Measure	Graduated HS (%) (7)	HS GPA (8)	HS Effort GPA (9)	Log Days Absent (10)	Days Suspended (11)	Held Back in HS (%) (12)
Sample Mean	80.1	2.32	2.23	2.92	0.17	28.0
Average Change in VA of Released Teachers (Δm_σ)	2.06	2.06	2.06	2.06	2.06	2.06
Panel A. 'Normal' Value-Added						
Benefit ($\hat{\rho}^N$)	0.26	0.006	0.004	-0.010	-0.000	0.12
Gain of Releasing Bottom 5% (G^N)	0.54	0.012	0.008	-0.021	-0.000	0.25
Panel B. Long-Run Value-Added						
Benefit ($\hat{\rho}^L$)	0.68	0.017	0.010	-0.026	-0.003	-0.22
Gain of Releasing Bottom 5% (G^L)	1.40	0.035	0.020	-0.054	-0.006	-0.45
Policy Gain Increase from Using Long-Run VA (%)	161.5	183.3	242.6	160.0	540.8	∞

Notes: This tables calculates the policy gains for twelve high school outcomes under policies that release teachers in the bottom five percent of the true VA distribution and replace them with a mean quality teacher according to 'normal' VA (Panel A) and long-run VA (Panel B). Gains are calculated using equation (6.1) which multiplies the impact of being assigned to a teacher whose VA is one standard deviation higher ($\hat{\rho}^\kappa$) by the the average improvement in VA caused by the policy. For long-run VA, the estimated benefit ($\hat{\rho}^L$) is identical to those reported in Figure 2. Under normality, the expected value of VA conditional on being a teacher in the bottom five percent is given by 2.06, which is then just the average change in VA caused by the policy given the replacement mean quality teachers have a VA of zero. The policy gain increase in terms of the high school outcome from using long-run VA in place of 'normal' VA is reported in the last row as a percent gain; if increased 'normal' VA is estimated to cause a deterioration in that outcome (while long-run VA causes an improvement) then an ∞ percent gain is reported.

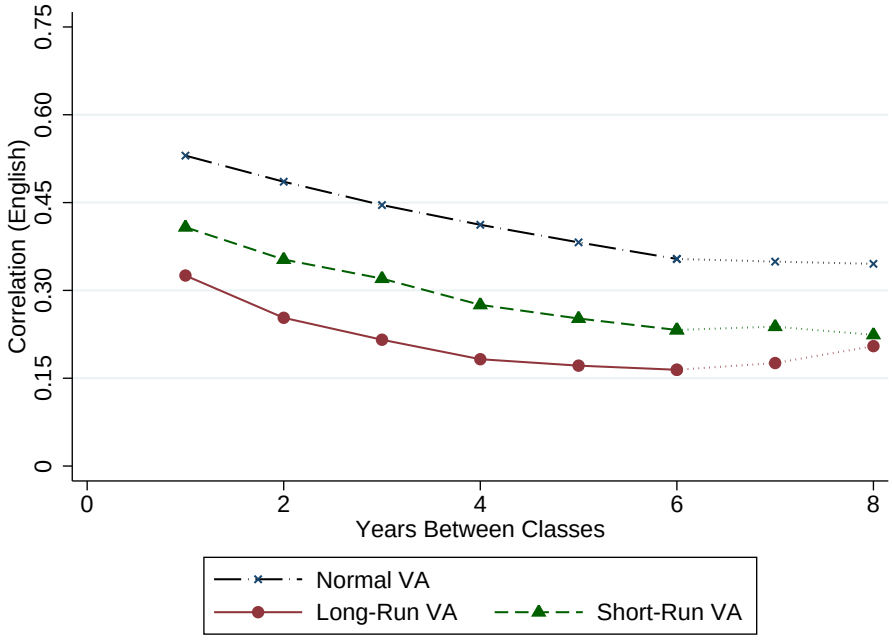
A Appendix Figures and Tables

Figure A.1: Autocorrelation for Mathematics and English Scores

(a) Mathematics Scores



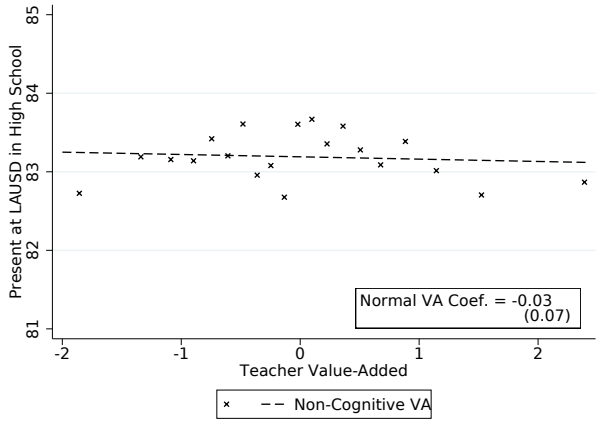
(b) English Scores



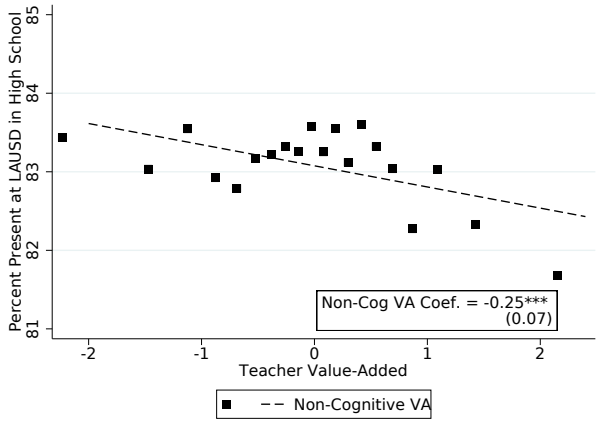
Notes: These figures show the correlation between mean test score residuals across classes taught by the same teacher for mathematics (Figure A.1(a)) and English (Figure A.1(b)). Correlations are estimated by first residualizing test scores using within-teacher variation as described by equation (2.2) then calculating a mean test score residual for each classroom. The autocorrelation coefficients are then given as the correlation across years for a given teacher, weighting by class size. See Table 2 (mathematics) and Table A.2 (English) for the point estimates underlying these figures.

Figure A.2: Effect of Value-Added Measures on High School Presence

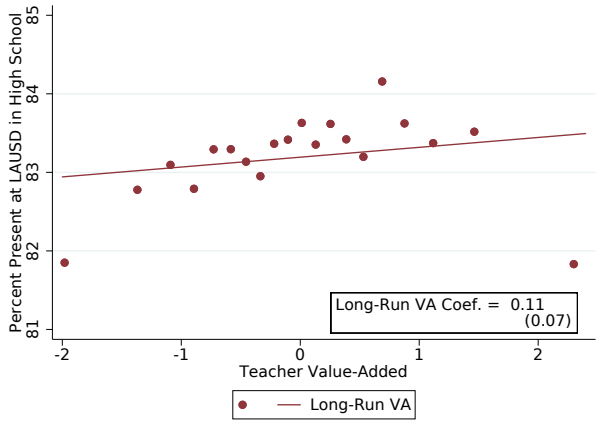
(a) 'Normal' Value-Added



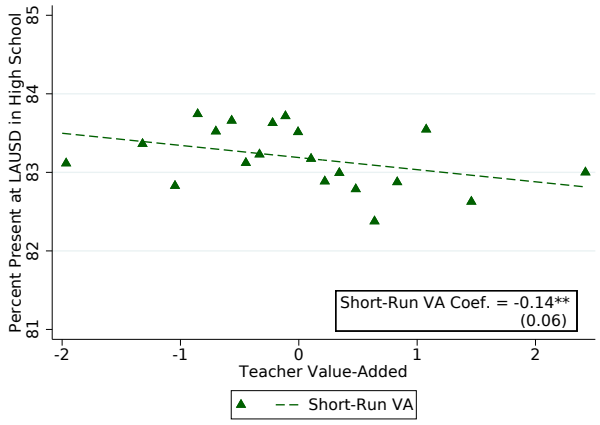
(b) Non-Cognitive Value-Added



(c) Long-Run Value-Added



(d) Short-Run Value-Added

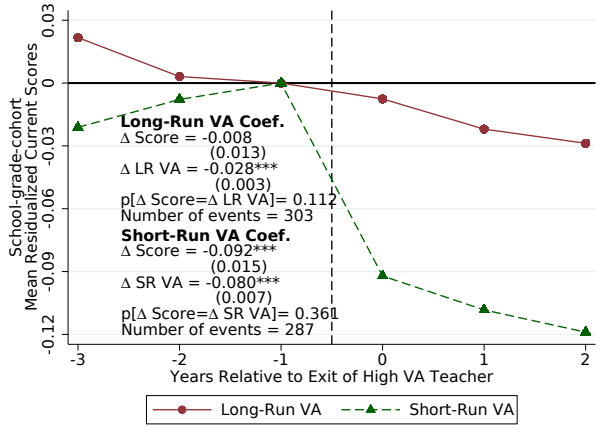


Notes: This figure shows the effect of our four measures of teacher VA on the likelihood of appearing in our high school outcomes data. All VA measures are normalized by the standard deviation of their respective teacher effects. Each figure is constructed in three steps: (i) residualize the likelihood of appearing in our high school outcomes data with respect to our control vector using within-teacher variation as described by equation (4.1), (ii) divide the normalized long- and short-run VA measures, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the long-run outcome in the estimation sample to facilitate interpretation of the scale. A line of best fit is then superimposed. Figures also report estimates of ρ from equation (4.3), which represent the effect of being assigned to a teacher whose VA is one standard deviation higher in a single grade on the likelihood of appearing in the high school data, along with its standard errors. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

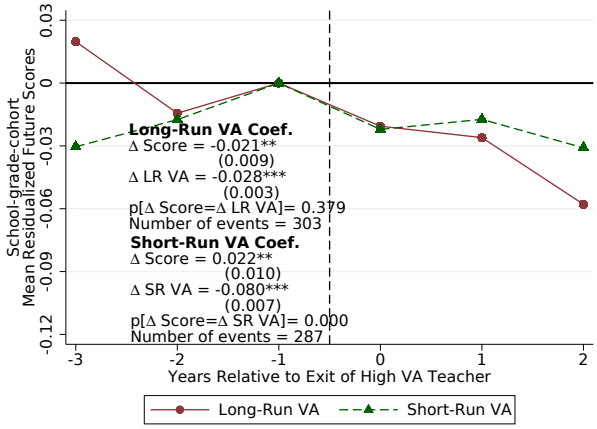
Figure A.3: Impacts of High and Low Long- and Short-Run VA Teacher Exit on Test Scores

Panel A: Impacts of **High** VA Teacher Exit on Current and Future Scores

(a) Current Test Scores

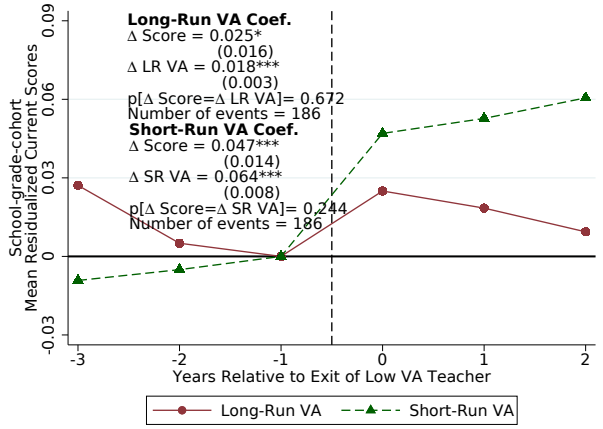


(b) Future Test Scores (i.e., Scores in $t+1$)

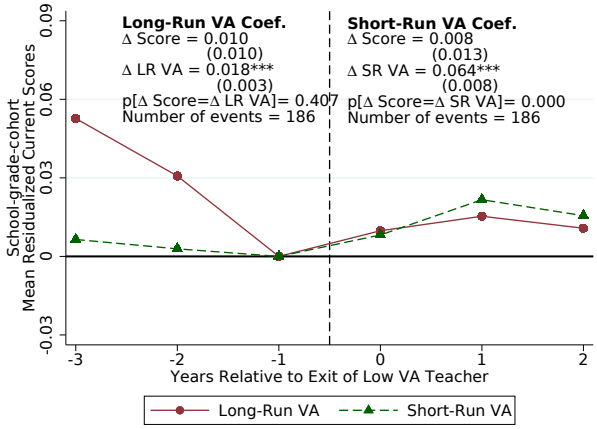


Panel B: Impacts of **Low** VA Teacher Exit on Current and Future Scores

(c) Current Test Scores



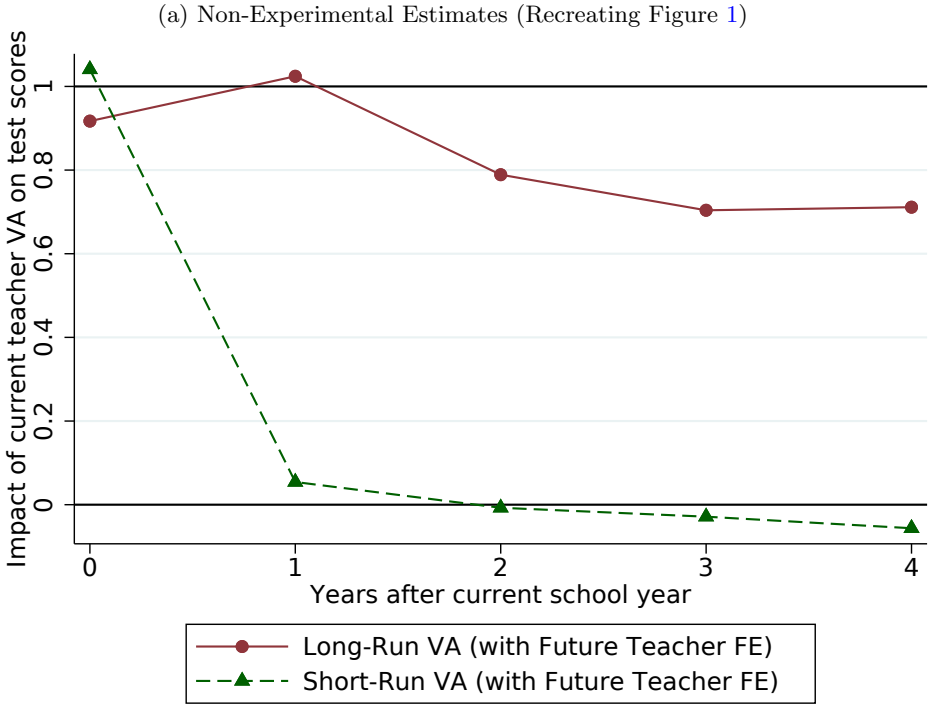
(d) Future Test Scores (i.e., Scores in $t+1$)



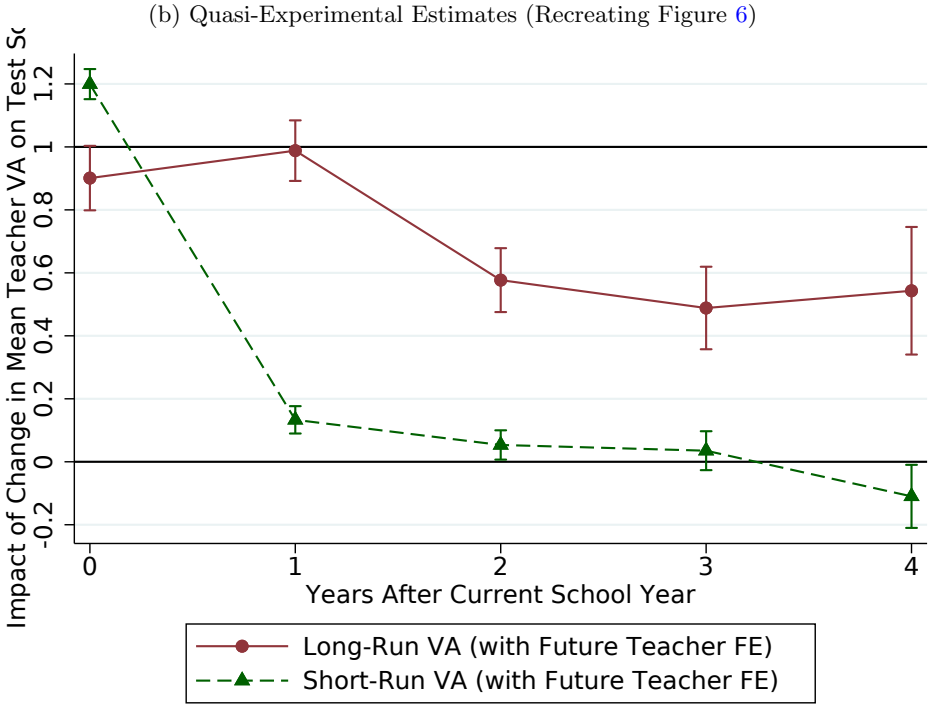
Notes: These figures plot event studies of test score changes by cohort as teachers exit a school-grade-subject cell at event-time 0. Panel A does so for high VA teachers (teachers with VA in the top 5% of the distribution), while Panel B does so for low VA teachers (teachers with VA in the bottom 5% of the distribution). Each figure then plots a series whereby VA is measured as long- and short-run VA. The left-hand side figures plotting 'current test scores' plot the residualized mean test scores for the school-grade-cohort this year, while right-hand side figures plot mean residualized 'future test scores' that the school-grade-cohort achieves in the *following* year. To construct each panel, we first identify the set of teachers who exited a school-grade-subject cell and were not teaching at the same school in the next period (i.e., we do not use within-school switcher variation) and define event time as the school year relative to the year of exit. We then estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$, including fixed effects for the subsequent teacher. We then identify the subset of teachers with VA estimates in the top or bottom 5% of the distribution among exiting teachers and then plot mean school-cohort current or future resdidualized test scores in the relevant school-grade-subject cell for the event years before and after the exit of such a teacher. We normalize the score changes to zero at event year -1 and include year fixed effects to eliminate secular time trends. 'Δ Score' reports the change in current or future test scores in the period after the teacher exited (period 0) relative to the period before (period -1). 'Δ VA' reports the change in long- or short-run VA (at the school-grade-cohort level) in the period after the teacher entered (period 0) relative to the period before (period -1). The p-value of a test of whether these coefficients are equal is then reported. Figure 5 reports results from similar event studies that leverage teacher entry.

Figure A.4: Effects of Short- and Long-Run Value-Added on Future Test Scores, Controlling for Future Teacher Fixed Effects

(a) Non-Experimental Estimates (Recreating Figure 1)

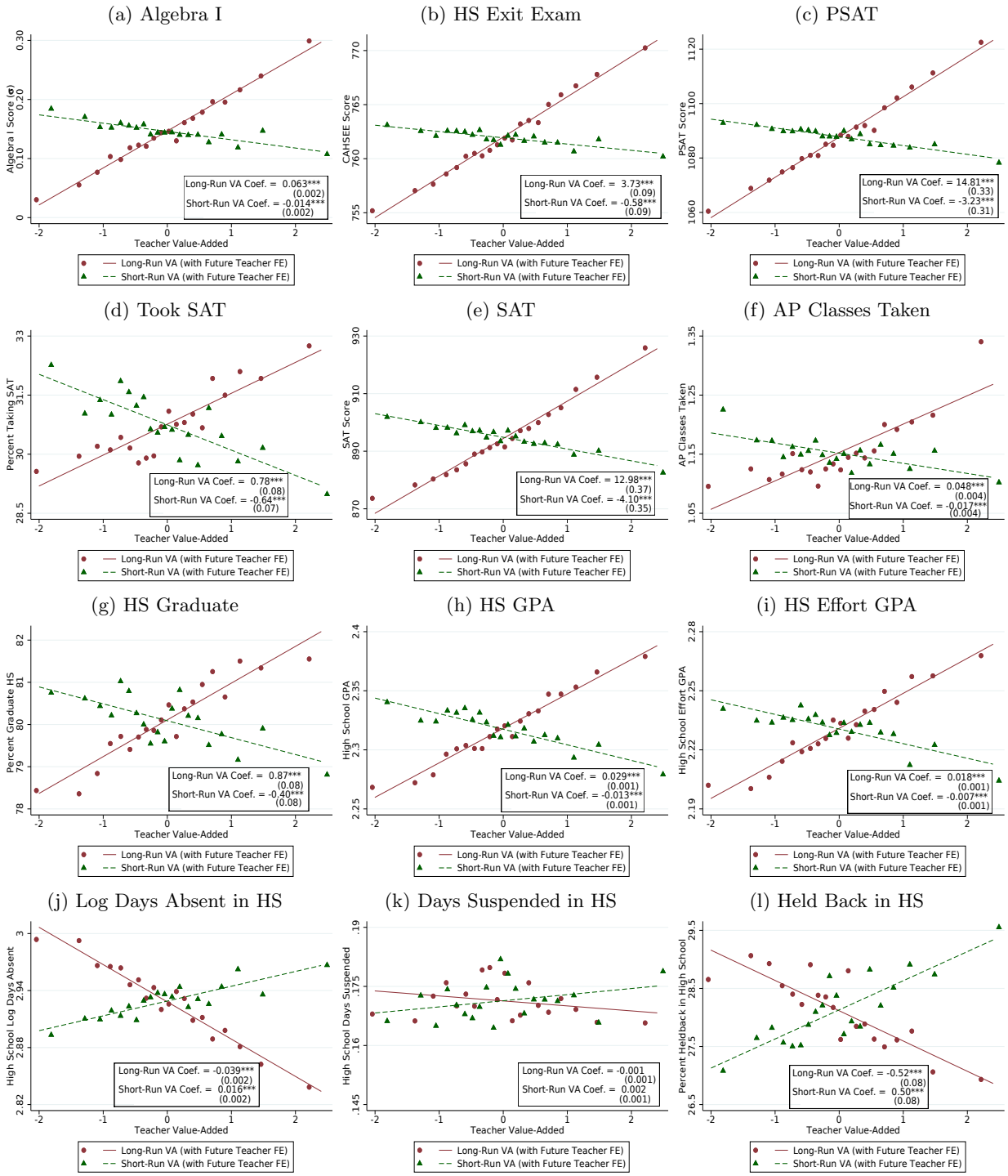


(b) Quasi-Experimental Estimates (Recreating Figure 6)



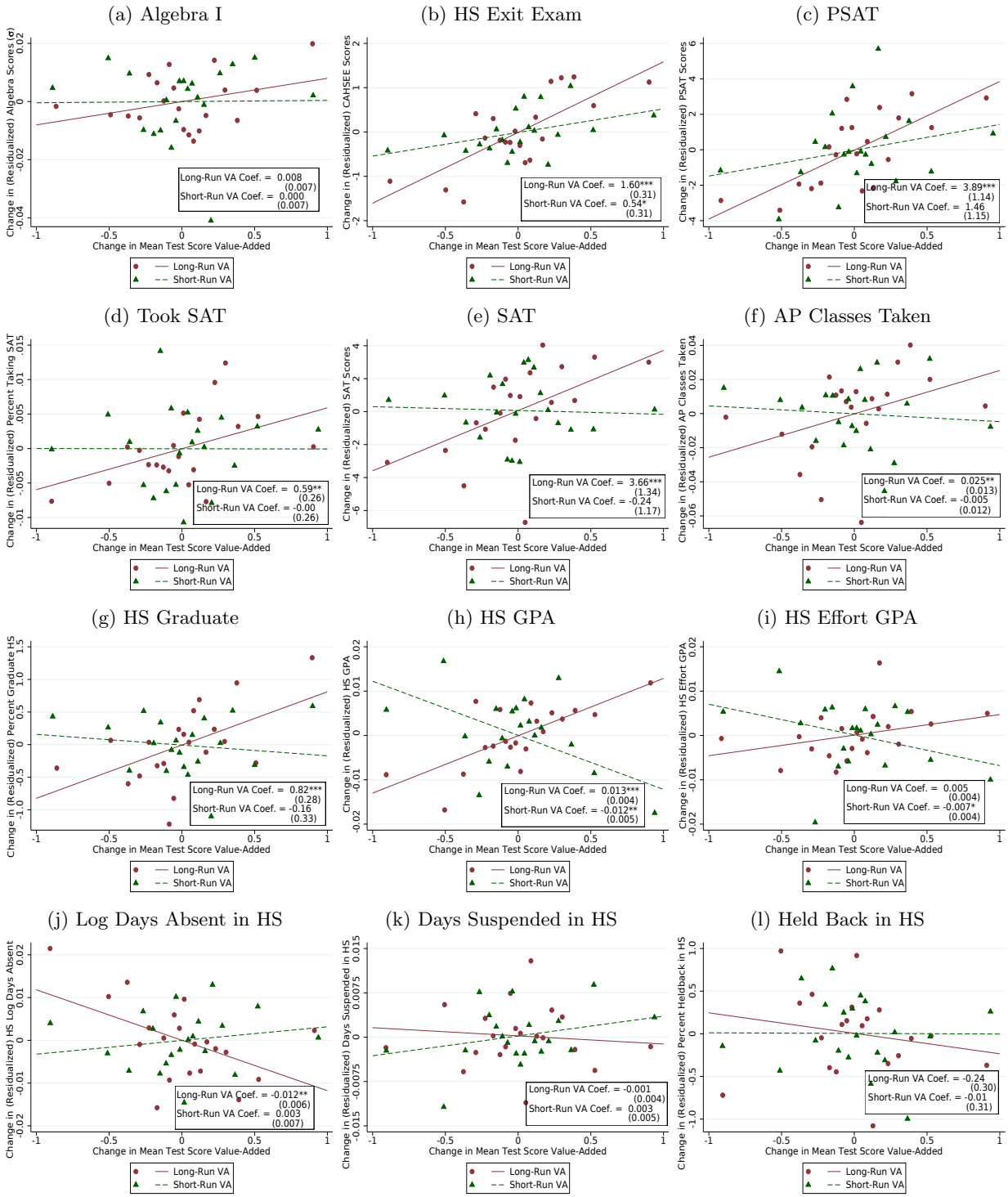
Notes: Figure A.4(a) recreates Figure 1, but includes fixed effects for the teacher in the subsequent year in the calculation of a teacher's short- and long-run VA. The figure thus shows the effect of teacher VA on test scores at the end of the current and subsequent school years for short- and long-run VA. The figure is constructed by regressing residualized end-of-grade test scores (summing math and English) in year $t+s$ on the teacher VA measure in year t. Similarly, Figure A.4(b) recreates Figure 6 including fixed effects for the teacher in the subsequent year in the calculation of a teacher's short- and long-run VA. The figure therefore reports the changes in school-grade mean raw test scores across cohorts for the current period and the next four periods against changes in mean teacher long-, and short-run VA as described in equation (5.1). Whiskers represent 95% confidence intervals with standard errors clustered at the student and classroom level in Figure A.4(a) and at the school-cohort level in Figure A.4(b).

Figure A.5: Effect of Short- and Long-Run Value-Added on Long-Run Outcomes, Controlling for Future Teacher Fixed Effects



Notes: This figure replicates Figure 2, but includes fixed effects for the teacher in the subsequent year in the calculation of a teacher's short- and long-run VA. As in Figure 2, both VA measures are normalized by the standard deviation of their respective teacher effects. Each figure is constructed in three steps: (i) residualize the long-run outcome with respect to our control vector using within-teacher variation as described by equation (4.1), (ii) divide the normalized long- and short-run VA measures, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the long-run outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the long-run outcome in the estimation sample to facilitate interpretation of the scale. A line of best fit is then superimposed. Figures also report estimates of ρ from equation (4.3), which represent the effect of being assigned to a teacher whose non-cognitive or long-run VA is one standard deviation higher in a single grade on the long-run outcomes, along with its standard errors. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A.6: Quasi-Experimental Estimates of Effect of Short- and Long-Run Value-Added on Long-Run Outcomes, Controlling for Future Teacher Fixed Effects



Notes: This figure replicate Figure 7, but includes fixed effects for the teacher in the subsequent year in the calculation of a teacher's short- and long-run VA. As in Figure 7, this figure visualizes how changes in mean school-grade-year short- and long-run teacher VA affects changes in the long-run outcomes of that school-grade-year. We create these figures by dividing normalized long- and short-run VA estimates into twenty equal-sized groups (vingtiles) and plotting the school-grade-year means of the long-run outcome residuals defined by equation (4.2), Y_{it}, within each bin against the school-grade-year mean value of the VA estimate within each bin.Points estimates of equation (5.1) where school-grade-year long-run outcomes are used as the dependent variable are also reported in the figures alongside standard errors clustered at the school-cohort level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

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

	Data Coverage (1)	Grades Usually Taken (2)	Cohorts Covered[a] (3)	Match Rate (% of covered VA sample) (4)
Algebra I	2002-03 to 2012-13	Grades 7-9	Entering 3^{rd} grade in 2006-07 or before	69% (481,041 of 699,833)
California High School Exit Exam (CAHSEE)	2002-03 to 2014-15	Grade 10	Entering 3^{rd} grade in 2007-08 or before	56% (459,447 of 826,808)
PSAT	2008-09 to 2016-17	Grade 10	Entering 3^{rd} grade in 2009-10 or before	50% (539,378 of 1,070,909)
SAT	2006-07 to 2016-17	Grades 11-12	Entering 3^{rd} grade in 2007-08 or before	29% (237,793 of 826,808)
Graduation[b]	2011-12 to 2015-16	Grade 12	Entering 3^{rd} grade in 2006-07 or before	52% (340,020 of 651,911)
AP Classes	2002-03 to 2016-17	Grade 12	Entering 3^{rd} grade in 2007-08 or before	71% (586,077 of 826,808)
High School Outcomes:[c]				
Absences	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	69% (656,293 of 950,914)
GPA	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	66% (628,713 of 950,914)
Days Suspended	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	71% (671,076 of 950,914)
Held Back	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	65% (616,405 of 950,914)

[a] For example, if third grade cohorts from 2007-08 and before are covered, then the linkage will include: third grade students 2003-04 through 2007-08, fourth grade students 2003-04 through 2008-09, and fifth grade students 2003-04 through 2009-10.

[b] Graduation is coded as a one if a student graduates or receives a special education certificate and zero if the student is coded as a dropout or still enrolled in the graduation data files. Still enrolled students are coded as graduates if they do graduate in a future year (or obtain their GED from the district). If the student is not present in these data files the student is coded as missing. Graduation also omits the 2003-04 fifth grade cohort as this cohort is not covered since the data does not start until 2011-12.

[c] For high school outcomes we require the student's cohort to reach the end of eleventh grade by 2016-17. The number of observations across these outcomes differ as some of these variables require information that is missing in the high school transcripts. To be in any of the data, we must observe a high school transcript for the student. In addition, for GPA we must also be able to construct a GPA using your transcripts. For absences, we note that there are some students have a transcript but missing absence data. Suspension data is universal conditional on having a transcript since we assume you were not suspended if you have a transcript but not a suspension record. For held back, we must observe a transcript both this year and last year to infer whether you were held back or not.

Table A.2: Autocorrelation and Variance Estimates of Long- and Short-Run VA (English)

	Normal VA	Long-Run VA		Short-Run VA	
	(1)	(1)	(2)	(1)	(2)
Autocorrelation Vector					
Lag 1	0.53	0.33	0.29	0.41	0.47
Lag 2	0.49	0.25	0.23	0.35	0.44
Lag 3	0.45	0.22	0.22	0.32	0.42
Lag 4	0.41	0.18	0.21	0.28	0.39
Lag 5	0.38	0.17	0.21	0.25	0.37
Lag ≥ 6	0.35	0.16	0.22	0.23	0.36
Within-year variance components					
Total SD	0.522	0.551	0.533	0.552	0.540
Individual-level SD	0.475	0.526	0.517	0.517	0.510
Class + teacher level SD	0.217	0.165	0.128	0.192	0.177
Estimates of teacher SD					
Lower bound based on lag 1	0.175	0.115	0.091	0.144	0.145
Quadratic estimate	0.185	0.135	0.098	0.158	0.151
Future Year Teacher FE	No	No	Yes	No	Yes
Student-Year Observations	1,181,362	1,055,506	1,055,506	1,055,506	1,055,506

Notes: This table gives the drift autocorrelation estimates across years for the same teacher used to compute normal, short- and long-run VA estimates in English. It does so for both when we do not control for the future teacher (columns (1)) and when we include future teacher fixed effects (columns (2)). It also reports the raw standard deviation of test score residuals and decomposes 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). In particular, we use the square root of the autocovariance across classrooms at a one year lag to estimate a lower bound and report an estimate of the standard deviation of teacher effects constructed by regressing the log of first seven autocovariances on the time lag and time lag squared and extrapolating to 0 to estimate the within-year covariance. Table 2 reports analogous model estimates for VA using mathematics scores as the outcome.

Table A.3: Correlation of All Components of Teacher Value-Added Measures

		Math VA		English VA		Non-Cognitive VA			
VA Measure		Long-Run	Short-Run	Long-Run	Short-Run	Absences	GPA	Suspensions	Held Back
Math VA	Long-Run	1							
	Short-Run	-0.24	1						
English VA	Long-Run	0.71	-0.07	1					
	Short-Run	-0.11	0.72	-0.16	1				
	Absences	0.11	-0.04	0.07	-0.05	1			
Non-Cog VA	GPA	0.17	0.04	0.19	0.05	0.18	1		
	Effort GPA	0.13	0.03	0.13	0.04	0.15	1		
	Suspension	0.10	-0.05	0.08	-0.05	0.12	0.10	1	
	Held Back	0.01	0.02	0.02	0.02	0.04	0.07	0.05	1

Notes: This table reports the correlations between our various value-added measure. Each VA measure is constructed as described in Section 3.2. Table 3 reports the correlation matrix between our four VA measures (which are built using the components listed in this table).