Making Teaching Last: Long-Run Value-Added[*]

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

January 18, 2023

ABSTRACT

Teacher value-added (VA) measures how teachers improve their students' contemporaneous test scores. Many teachers, however, argue that contemporaneous test scores are a poor proxy for their permanent influence on students. This paper considers a new VA measure – 'long-run VA' – that captures teachers' contributions that *persist* by replacing contemporaneous test scores with subsequent test scores in VA estimation. We find that students assigned to high long-run VA teachers fare substantially better in terms of long-term outcomes. Policy simulations indicate that the use of long-run VA improves policy effectiveness by a factor of two compared to traditional VA measures.

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

JEL Classifications: H40, I21, I22

[*] We would like to thank Kirabo Jackson, Jesse Bruhn, and Jonah Rockoff along with seminar participants at the NBER Fall Education Meetings, Arizona State University, McMaster University, Simon Fraser University, University College Dublin, University of Sydney, University of Wisconsin-Madison, 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, Friedman, and Rockoff, 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 methods that develop long-lasting cognitive or non-cognitive skills are not initially measured by tests, but are important for future learning.[1] In addition, there are concerns that incentivizing teachers' contemporaneous contributions may encourage teachers to adopt teaching practices that improve contemporaneous test scores over those that impart deeper knowledge or lead to improved non-cognitive skills – a phenomenon referred to as 'teaching to the test' in common parlance (Koretz, 2002).

In light of these concerns, this paper considers a new VA measure – 'long-run VA' – that captures teachers' contributions that *persist* to the following year by replacing contemporaneous test scores with subsequent test scores as the dependent variable in VA estimation. When doing so, we incorporate fixed effects for a student's subsequent teacher to account for that teacher's influence in the following year. We estimate long-run VA alongside the standard VA measure using over a decade of data from students in third through fifth grade in a large school district. While standard and long-run VA measures are correlated, the correlation is relatively modest at 0.5. We also estimate non-cognitive VA in light of recent research highlighting that these VA measures are better predictors of long-term behavioral outcomes such as high-school completion,

[1] A burgeoning literature has found that teaching practices are linked to the skills that students develop. These papers often delineate teaching practices into traditional methods such as lecturing and rote memorization and modern methods such as group work. Bietenbeck (2014) finds that the two teaching practices promote different cognitive skills, with traditional methods improving students' factual knowledge and routine problem solving while modern methods promote reasoning. In line with this finding, recent research has found that traditional methods are more effective at raising contemporaneous test scores (Schwerdt and Wuppermann, 2011; Lavy, 2016; Cordero and Gil-Izquierdo, 2018) – possibly as factual knowledge and routine problem solving skills are tested more frequently (Bietenbeck, 2014) – while modern teaching methods lead to improvements in students' noncognitive skills, but no change in contemporaneous test scores (Algan et al., 2013; Korbel and Paulus, 2018).

college-going, and crime (Jackson, 2018; Petek and Pope, 2018; Flèche, 2017; Rose et al., 2019). We find that long-run VA is substantially more correlated with non-cognitive VA than standard VA, suggesting that our long-run VA measure partially captures teachers' influence on student behavioral improvements that are crucial for future success.

We next compare all three VA measures and show that long-run VA is the best predictor of teachers' contributions to students' lifelong success. To do so, we calculate the impact of being assigned to high standard, long-run, and non-cognitive VA teachers in elementary school on later test scores and high school outcomes such as graduation, high school test scores, GPA, absences, and taking and performing well on the SAT. First, we show that students assigned to high long-run VA teachers continue to see large test score improvements in subsequent grades, contrary to the widely-documented test score fade out after being assigned high standard VA teachers (e.g., McCaffrey et al., 2004; Jacob et al., 2010; Rothstein, 2010; Chetty et al., 2014b, etc.). Second, we find that being assigned high long-run VA teachers improves high school outcomes far more than being assigned high standard VA teachers. Third, we show that high long-run VA teachers lead to improved high school outcomes over and above high non-cognitive VA teachers. Fourth, when all three VA measures are included as predictors of later high school outcomes, VA's ability to predict long-term outcomes comes *entirely* from long-run and non-cognitive VA, with long-run VA driving the largest improvements.

We rationalize these findings using a model that separates teachers' contributions 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 test scores as the dependent variable. 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 decompose standard VA and find that long- and short-run VA contribute approximately 40 and 60 percent to the standard VA measure, respectively. While long- and short-run VA are positively correlated with the standard teacher VA measure (by construction), they are uncorrelated with each other. In addition, long-run VA is positively correlated with non-cognitive VA measures, whereas short-run VA is not. We also find no evidence that being assigned a higher short-run VA teacher improves a student's later test scores or high school outcomes. These results provide strong support for our model that long-run VA captures teachers' contributions to a deeper understanding of the material or improved behavior that persists well into the future, while short-run VA captures contributions to fully transitory knowledge.

Next, we test whether our long- and short-run VA measures are unbiased measures of teacher's contributions to long- and short-term knowledge.[2] To do so, we follow Chetty et al. (2014a) and Bacher-Hicks et al. (2014) and 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 of a teacher in the bottom or top five percent of the long- and short-run VA distributions, finding that contemporaneous test scores change sharply across cohorts when these teachers enter a school-grade. In contrast, test scores for those cohorts in the subsequent year change sharply with the entry of a high or low long-run VA teacher, whereas high or low short-run VA teachers have no effect on these subsequent test scores. We conduct similar event studies around the departure of a teacher in the bottom or top five percent of the VA distribution and find analogous results.

We also leverage variation from the entire distribution of teacher turnover and show that the change in contemporaneous school-grade mean test scores is statistically indistinguishable 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

[2] 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 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.

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 five 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 used instead.

Adopting long-run VA leads to large improvements in policy efficiency. For every outcome we consider, policy gains are at least doubled when long-run VA is used over standard VA. In particular, we find that releasing teachers based on long-run VA estimates after three years of data leads to, on average, each affected class having an additional 0.2 students graduating high school and 0.22 students taking the SAT. These classes would also score 0.04σ greater on high school algebra and have SAT scores that are 10 points higher. 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 standard VA that predicts students' future outcomes can be easily estimated and used in lieu of standard VA for education policy. Third, because the standard teacher quality measure 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, teachers may substitute between teaching methods and so policies targeting long-run VA may be even more effective if they encourage teachers to adopt teaching practices which generate gains that persist, such as through improvements to non-cognitive skills. 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 account when assessing teachers. In particular, we measure the extent to which some teachers produce a temporary recognition that disappears the following year, while other teachers impart deep knowledge that persists. The ability to measure teachers' contributions along these two dimensions opens up new avenues of research. For instance, teacher incentive schemes aim to improve educational outcomes by increasing the VA of teachers (Biasi, 2021). Teachers, however, may also respond to incentives by substituting between teaching short- and long-term knowledge. For instance, Macartney et al. (2018) show that teachers increase VA in response to accountability incentives, but those test score gains fade out at high rates. Those results suggest that teacher incentive schemes could be further refined to take both effort responses and substitution between teaching methods into account to better drive students' lifelong success.

The rest of the paper is organized as follows: Section 2 describes the administrative data that we use and the estimation procedure for the VA measures. In Section 3, we estimate and describe the long-run, standard, and non-cognitive VA, and link these VA measures to long-run outcomes. We develop our model of long- and short-run VA in Section 4 and assess our long- and short-run VA estimates. 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 Data and VA Estimation

Our goal is to compare our long-run VA measure to the standard VA measure in its ability to predict students' later life outcomes. We will also contrast it to non-cognitive VA, which has been found to be a better predictor than traditional VA measures for some long-term behavioral outcomes such as high-school completion, college attendance, and crime (Jackson, 2018; Petek and Pope, 2018; Flèche, 2017; Rose et al., 2019). To do so, we need to estimate these VA measures and link students to later life outcomes. In that spirit, this section introduces the administrative data that we use and provides descriptive statistics. We then detail our VA estimation procedure, including the control vector that we use to residualize test scores.

2.1 Data

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

In addition to test scores, our data also contain behavioral variables which we use to construct a 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), along with achievement grades in 10 subjects (e.g., reading, mathematics, art, etc.) for each trimester, which we use to construct a student's GPA. Jackson (2018) argues that these non-test score outcomes are good proxies for non-cognitive skills given their strong association with well-known psychometric measures including the "big five" and grit which attempt to measure these non-cognitive skills. In addition, as in Petek and Pope (2018), our data contain student "effort" grades for 10 subjects, which are teachers' assessments of student effort in a given subject. We use these to build an "effort GPA" that we employ as an additional non-cognitive outcome. 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

[3] In 2013-14, the state switched to a new testing system to align their content to Common Core. Due to this switch, state standardized test scores are unavailable for 2013-14. Given the missing test score data, we restrict our analysis to pre-2012-13 cohorts as we are unable to calculate all our VA measures for post-2012-13 cohorts due to the requirement for lagged, contemporaneous, and subsequent test scores.

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

percent for all variables with the exception of parental education, which is missing for twenty-nine percent of the sample who responded "Decline to Answer." We include a missing data category for each demographic variable to deal with the (limited) missing demographic data.

We make several sample restrictions to create our final VA analysis sample. First, we drop 148,644 student-year observations that cannot be matched to a teacher or the teacher is 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 and lagged test scores in that subject, losing 11,951 student-year observations. Third, we exclude 99,729 observations that lack a valid current or lagged test score in that subject.[6] The final standard VA sample consists of 1.2 million student-year observations, covering 550,000 students and 13,000 teachers. Our long-run VA sample is near-identical, but drops 120,696 observations with missing test scores in the subsequent year.[7]

Table 1 provides summary statistics for our data. Column (1) reports these for the full sample, while column (2) does so for our standard 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 standard 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. Column (3) then reports the sample used to construct long-run VA, with only minor differences observed between the long-run and standard VA samples.

Long-Term 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

[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 of their class.

[6] As the last two restrictions are subject-specific, our VA samples are also subject-specific. Our math VA sample has 1,187,231 observations while our English VA sample has 1,181,362 observations.

[7] We also employ subsequent teacher fixed effects in the estimation of long-run VA. Fifth grade students, however, can only be matched to a subsequent teacher in K-6 or K-12 schools due to the lack of student-teacher links in middle schools. For that reason, we replace the subsequent teacher with a school-grade-year identifier in cases where a student cannot be matched to a subsequent teacher, but we observe the subsequent school-grade attended. The subsequent teacher fixed effect requirement necessitates us to drop an additional 4,564 observations lacking teacher or school-grade assignment data in the subsequent year.

scores, exit exam scores, PSAT scores, SAT-taking and scores, AP classes taken, graduation, high school GPA, 'effort GPA,' absences, suspensions, and grade repetition. For outcomes that 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-term 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-term 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 leaving the school district between elementary school (grades 3-5) and the grade that outcome occurs 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 the VA of one's elementary school teacher influences the probability a student is present for any of the long-term outcome data that occur before they are able to drop out of high school (see Figure A.3).[8]

2.2 Estimation of Value-Added

Standard Value-Added: Teachers are indexed by j and teach one class per year (as in our empirical application). 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

[8] Students in the state cannot drop out until age 16, which (roughly) equates to tenth grade. So therefore we look for differential selection based on high school outcomes that are recorded for students by ninth grade. Specifically, we code a student as appearing in high school if they have non-missing algebra and high school suspension data. Algebra is chosen as students take algebra for the first time by ninth grade, while high school suspensions are chosen as a student will have a suspension record (even if never suspended) if we observe a ninth grade report card. These outcomes, therefore, occur prior to tenth grade, before students can drop out of high school. Differential selection based on the VA of one's elementary school teacher is observed for outcomes that often occur after tenth grade such as the high school exit exam as well as outcomes that are voluntary (e.g., PSAT and SAT).

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 within-teacher variation 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. We follow Chetty et al. (2014a) and estimate μ_{jt} using a jack-knife empirical Bayes estimator. 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 $\bar{A}_{jt}$:

$$\bar{A}_{jt} = \psi \bar{\mathbf{A}}_j^{-t}\,, \tag{2.4}$$

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:

$$\hat{\mu}_{jt} = \hat{\psi} \bar{\mathbf{A}}_j^{-t}\,. \tag{2.5}$$

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 stu-*

dent achievement and so $\mathbb{E}[\epsilon_{ijt}|j] = \mathbb{E}[\epsilon_{ijt}]$*.*

Long-Run Value-Added: While the goal of the standard VA model is to capture the impact of teacher j on *contemporaneous* test scores, we want to capture the impact of teacher j that persists. We call the impact of teacher j that persists to period $t+1$ long-run value-added. We estimate long-run VA by replacing contemporaneous test scores with *subsequent* test scores, $A^*_{ij,t+1}$, as the outcome of interest in equation (2.1):

$$A^*_{ij,t+1} = \beta X_{ijt} + \mu^L_{jt} + \mu_{k,t+1} + \epsilon_{ij,t+1}\,, \tag{2.6}$$

where X_{ijt} are the same observable determinants of student achievement defined in equation (2.1), μ^L_{jt} represents teacher j's long-run VA, and $\mu_{k,t+1}$ is the effect of the subsequent teacher k on test scores in $t+1$. We then construct residualized subsequent test scores, $A_{ij,t+1}$, using the same control vector as before from equation (2.2) with the addition of *subsequent teacher fixed effects* to control for the impact of the subsequent teacher, $\mu_{k,t+1}$.

Accordingly, residualized subsequent test scores, $A_{ij,t+1}$, are given by:

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

Estimation of μ^L_{jt} then follows a near-identical method to the estimation of standard VA but redefines $\bar{A}_{jt}$ based on (residualized) test scores in period $t+1$. Specifically, $\bar{A}^{t+1}_j \equiv \frac{1}{n}\sum_{i\in j} A_{ij,t+1}$ and $\bar{\mathbf{A}}^{t+1,-t}_j \equiv (\bar{A}^{t+1}_{j1}, ..., \bar{A}^{t+1}_{jt-1}, \bar{A}^{t+1}_{jt+1}, ..., \bar{A}^{t+1}_{jT})'$. The empirical analog for ϕ^{t+1} is then estimated by regressing $\bar{A}^{t+1}_{jt} = \phi^{t+1}\bar{\mathbf{A}}^{t+1,-t}_j$. The estimate for the long-run component of teacher j's VA in period t is then:

$$\hat{\mu}^L_{jt} = \hat{\phi}^{t+1}\bar{\mathbf{A}}^{t+1,-t}_j\,. \tag{2.8}$$

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 one. This assumption is identical to that of Petek and Pope (2018). Formally:

Assumption 2 *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 come back to this assumption in Section 5, where we explicitly test its validity.

Non-Cognitive VA and Dimensionality Reduction: To estimate non-cognitive VA we follow Jackson (2018) and Petek and Pope (2018) and replace the (residualized) test score outcomes with five (residualized) non-cognitive outcomes from the year after a student was in a teacher's class (as done for long-run VA):[9] log days absent (plus one), GPA, effort GPA, indicator for getting suspended, and an indicator for not progressing to the next grade on time (i.e., held back). As we are using subsequent year outcomes, we include subsequent teacher fixed effects when residualizing the non-cognitive outcomes (as done for long-run VA). We note that the inclusion of subsequent teacher fixed effects increases the predictive power of non-cognitive VA for long-term outcomes, and so our non-cognitive VA measures are more predictive of long-term outcomes compared to the prior literature which does not include these subsequent teacher fixed effects.

To reduce dimensionality, we construct a non-cognitive VA index. 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 have mean zero and standard deviation one.[10] Similarly, we construct a test score index for both the standard and long-run VA measures by combining our math and English VA estimates, giving each subject equal weight (i.e., $VA_{test} = \frac{1}{2}VA_{math} + \frac{1}{2}VA_{English}$), and normalizing the resulting index to have mean zero and standard deviation one.

Constructing Test Score Residuals: We construct the test score residuals for contemporaneous, A^*_{ijt}, and subsequent test scores, $A^*_{ij,t+1}$, for each subject (math and English) along with non-cognitive outcome residuals by regressing the raw standardized test score (or non-cognitive outcome) on a vector of covariates, X_{ijt}, and teacher fixed effects, as described by equation (2.3) (subsequent test score and non-cognitive residuals also include subsequent teacher fixed effects in the residualization). Our baseline control vector is similar to that of Chetty et al. (2014a),

[9] We use the subsequent rather than contemporaneous year for non-cognitive outcomes to avoid potential teacher manipulation as these non-test-score variables are often teacher reported.

[10] Alternatively, one could use exploratory factor analysis to choose the factor load on each VA variable as in Jackson (2018) and Petek and Pope (2018). The factor loadings that we find are near-identical to those in Petek and Pope (2018) and so we refer interested readers to their paper. Similar to them, we find only small differences when we use the estimated factor loadings to compute the index instead of equal-weighting the index.

although we also include lagged non-cognitive measures 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 English, interacted with grade dummies,[11] (ii) a cubic polynomial of the five lagged non-cognitive measures (log absences, GPA, effort GPA, suspended, and held back) interacted with grade dummies,[12] (iii) demographics, including: parental education (five education groups), 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 (log absences, GPA, effort GPA, suspended, and held back) interacted with grade dummies, (iii) class and school-grade means of all the demographic covariates, (iv) class size, and (v) grade-by-year dummies.

3 Results

We estimate standard, long-run, and non-cognitive VA using the methodology described above. Here, we describe our VA model estimates and the correlation between the three VA measures. We then turn to capturing the relative ability of the three VA measures to predict impacts on students' future test score performance and high school outcomes. Throughout, standard errors are adjusted for two-way clustering at the classroom and student levels to account for the fact that students face common class-level shocks and our stacked data often features multiple entries per student (Cameron et al., 2012).

[11] We exclude observations if own-subject prior scores are missing. If other-subject prior scores are missing, 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.

[12] Once again, we omit observations if the lagged own-non-cognitive measure is missing. If data is missing for other-non-cognitive measures, we set the non-cognitive measure equal to its mean for that student's grade and include an indicator for missing data.

3.1 VA Model Estimates

Table 2 presents parameter estimates for standard and long-run VA in mathematics and English. The first six rows 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 that number of years later. Unsurprisingly, reliability decays over time and so more recent test scores are better predictors of current teacher performance. The reliability of long-run VA is lower than that of standard VA. Intuitively, this arises because additional factors influence the subsequent test score which reduces the reliability of our long-run VA measure. Such differences in reliability will be incorporated into the policy analysis later.

Table 2 also reports the estimated standard deviation of teacher effects. Since teachers teach only one class per year, the standard deviation of teacher effects cannot be point identified 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. We find that the standard deviation of teacher effects for standard VA is 0.28 and 0.19 in mathematics and English, respectively. For long-run VA, the estimated standard deviation of teacher effects for mathematics and English is 0.12 and 0.09, respectively. The estimated variance of teacher effects makes clear that the distribution of long-run teacher effects is far less disperse than that of standard VA.

Relationships Between VA Measures: Table 3 reports the relationship between our various VA measures (with VA measures being combined into indices as described in Section 2.2). Here, some interesting patterns emerge. First, long-run VA is correlated to the standard VA measure. This is expected as teachers that raise test contemporaneous test scores also likely raise subsequent test scores. Perhaps surprisingly, however, the correlation between standard and long-run VA is only 0.51. This indicates that there are many teachers who are highly-effective at raising subsequent test scores, but not contemporaneous test scores (and vice versa). In addition, long-run VA is substantially more correlated with non-cognitive VA than standard VA, suggesting that long-run VA is picking up components of teaching that not just raise test scores in the following period but *also* improve student behavior. Table A.2 further reports the correlations of all components of the various VA measures.

3.2 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

$$Y_i^* = \alpha_j + \alpha_k + \beta^Y X_{it}\,, \tag{3.1}$$

where Y_i^* is the future outcome of interest, α_j a fixed effect for the contemporaneous teacher, α_k is a fixed effect for the subsequent teacher, and X_{it} are observed characteristics of the student. Using the estimates from equation (3.1), the future outcome residuals, Y_{it}, are defined as:[13]

$$Y_{it} = Y_i^* - \hat{\beta}^Y X_{it} - \hat{\alpha}_k\,. \tag{3.2}$$

When future outcomes are defined as a test score (e.g., test scores in the subsequent year), we pool across all grades and subjects and regress these subject-specific test score residuals on our jack-knifed teacher VA estimate. Formally, we regress:

$$Y_{ist} = \delta + \rho^\kappa \hat{\mu}^\kappa_{jst} + \lambda_s + \eta_{ist}\,, \quad \kappa \in \{\textit{Standard}, \textit{Long-Run}\}, \tag{3.3}$$

where λ_s is a fixed effect for subject s and $\hat{\mu}^\kappa_{jst}$ is our estimate of a teacher's VA in subject s, with the superscript κ denoting the VA measure we are using (e.g., standard or long-run VA).

For high school outcomes,[14] we pool across all grades and regress high school outcome resid-

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

[14] We also use the same methodology for future test scores when assessing the impact of non-cognitive VA, using mean mathematics and English residual test scores as the outcome.

uals on teachers' jack-knifed normalized VA *index*. We consider three VA indices: standard, long-run, and non-cognitive. These indices reduce dimensionality by combining multiple outcomes measures (e.g., math and English), and are always normalized to have mean zero and variance one to account for differences in the dispersion of the VA measures (see Section 2.2). Formally,

$$Y_{it} = \delta + \rho^{\kappa}\hat{m}^{\kappa}_{jt} + \eta_{ij}\ , \quad \kappa \in \{\mathit{Standard}, \mathit{Long\text{-}Run}, \mathit{Non\text{-}Cognitive}\}, \tag{3.4}$$

where $\hat{m}^{\kappa}_{jt}$ denotes teacher VA index κ.

Future Test Scores: Figure 1 plots the impacts of our three VA measures on test scores in the current year and subsequent years.[15] (Table A.3 reports the point estimates along with standard errors that underlie Figure 1.) In the current year, a one-unit higher standard or long-run VA teacher raises contemporaneous test scores by one-unit. In the year following teacher j's class, however, the impact of the two VA measures diverge: a one-unit higher long-run VA teacher increases test scores by one, while a one-unit higher standard VA teacher only raises test scores by 0.34 of a standard deviation. Four years later, the one-unit higher long-run VA teacher continues to substantially raise scores by 0.74 of a standard deviation. In contrast, the one-unit higher standard VA teacher has a relatively small effect, only raising scores by 0.16 of a standard deviation four years later. The small impact of standard VA on test scores in subsequent years has been well-noted in the VA literature, generating concern about whether VA can be used to improve long-term student performance (Jacob et al., 2010; Rothstein, 2010). Only small test score impacts are observed for non-cognitive VA.

Impacts on High School Outcomes: Figure 2 displays the impact of standard and long-run VA on twelve high school outcomes by plotting (residualized) high school outcomes for students in school year t versus the (jack-knifed) standard and long-run VA indices. We construct the binned scatter plots in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the standard or long-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot

[15] When regressing long-run VA on contemporaneous test scores we also control for what we will later call 'short-run VA' in the model we develop in Section 4. This is done because both long-run and short-run VA influence contemporaneous test scores and are (slightly) correlated (see Table B.1) and so both must be included to generate the one-to-one relationship between long-run VA and contemporaneous test scores.

the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school outcome in the estimation sample to facilitate interpretation of the scale. Each panel of Figure 2 also reports the point estimates and standard errors from equation (3.4).

In terms of the slope coefficients underlying each panel, we find that being assigned to a teacher whose *long-run* VA index is one standard deviation higher in a single grade increases algebra scores by 0.04 standard deviations, raises exit exam scores by 3 points, improves PSAT scores by 11 points (on the 1600 scale), boosts SAT scores by 9 points (on the 1600 scale) *in addition to* increasing SAT-taking rates by 0.6 percentage points, raises the number of AP classes taken by 0.06, improves high school graduation rates by 0.6 percentage points, boosts high school GPA by 0.03, raises high school 'effort GPA' by 0.01, *decreases* days absent in high school by 2.4 percent, reduces days suspended in high school by 0.005, and lowers high school grade repetition by 0.5 percentage points. All of these improvements in long-run outcomes are statistically significant at the one percent level. In contrast, the estimated improvements in terms of high school outcome from being assigned a one standard deviation higher standard VA teacher are substantially lower. Given that the correlation between long-run and standard VA is 0.51, this indicates the possibility that the *entire* impact of being assigned high standard VA teachers on high school outcomes is driven by long-run VA.

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. We find that being assigned to a teacher whose *non-cognitive* VA is one standard deviation higher in a single grade improves test score based outcomes (i.e., algebra scores), but to a lesser extent than long-run VA. For behavioral-based high school outcomes (i.e., suspensions, absences, GPA), we find that one standard deviation higher non-cognitive and long-run VA teachers affect outcomes similarly. Together these results suggests that long-run VA is a better predictor of high school outcomes than non-cognitive VA, although non-cognitive VA is likely to have some predictive power that is independent of long-run VA, especially for more behavioral-based outcomes.

Multivariate VA Effects: The prior analysis looked at how each dimension of teacher quality affected high school outcomes. Given that these VA measures are correlated, however, we want to determine whether each VA measure independently affects long-term outcomes. As foreshadowed, it could be that the impact of standard VA on high school outcomes is solely caused by the fact that it is positively correlated to long-run VA. We investigate the extent to which different dimensions of teacher quality matter for high school student outcomes by regressing:

$$Y_{it} = \delta + \rho^{Stan}\hat{m}_{jt}^{Stan} + \rho^{Long}\hat{m}_{jt}^{Long} + \rho^{NC}\hat{m}_{jt}^{NC} + \eta_{ij}\,, \tag{3.5}$$

where the superscripts *Stan*, *Long*, and *NC* represent standard, long-run, and non-cognitive VA, respectively.

Table 4 reports the results of the multivariate regression. As before, long-run VA is an strong predictor of high school outcomes, with non-cognitive VA also being able to independently predict high school outcomes, especially more behaviorally-oriented outcomes such as graduation, GPA, and absences. The estimates make clear that long-run VA is a superior predictor of high school outcomes, although non-cognitive VA contains additional information that influences high school outcomes. Therefore, it is likely optimal for policymakers to utilize long-run and non-cognitive VA jointly in decision-making.

In contrast, the standard VA measure is, if anything, negatively related to high school outcomes once we control for long-run VA. Intuitively, long-run VA captures the contemporaneous test score gains measured by standard VA that *persist*. Once the persistent test score gains are accounted for, the residual variation in standard VA that remains is the contemporaneous test score gains imparted by the teacher that cease to impact test scores in the following period. To interpret this residual variation in the standard VA measure, Section 4 develops a model where knowledge has both short- and long-term components. While this requires us to impose some additional structure on the education production function, it allows us to capture teachers' contribution to long- and short-term knowledge, which we call 'long-' and 'short-run' VA, respectively.

4 Mechanisms and 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 standard VA can be decomposed into teachers' contributions to long- and short-term knowledge. Teachers' contributions to long-term knowledge will be our long-run VA measure which we have showed contains the portion of standard VA that predicts students' later life outcomes. The model will then interpret the residual component of standard VA as teachers' contributions to short-term knowledge, which we call 'short-run VA.'

Long- and Short-Run VA Model: We depart from the standard VA model and consider knowledge to consist of transitory and permanent components. 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 separated into a short-, μ_{jt}^S, and long-term, μ_{jt}^L, knowledge component:

$$\mu_{jt} = \mu_{jt}^L + \mu_{jt}^S \,. \tag{4.1}$$

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

We start with our standard VA model introduced in Section 2.2. There, residualized test scores in year t are given by the sum of the contemporaneous impact of teacher j on test scores (i.e., standard VA), μ_{jt}, and idiosyncratic student variation, ϵ_{ijt}:

$$\begin{aligned} A_{ijt} &= \mu_{jt} + \epsilon_{ijt} \\ &= \mu_{jt}^L + \mu_{jt}^S + \epsilon_{ijt} \,, \end{aligned} \tag{4.2}$$

where the second line follows from equation (4.1). Unfortunately, the econometrician can only estimate the sum of the long- and short-run teacher components (i.e., standard VA), μ_{jt}, and not teachers' contributions to the individual components.

To find teachers' contributions to the long- and short-run components, we consider achievement the following year when student i is assigned to teacher k. In this period, achievement will be given by:

$$A^*_{ijk,t+1} = \delta^L \mu^L_{jt} + \delta^S \mu^S_{jt} + \mu_{k,t+1} + \beta X_{ijt} + \epsilon_{ijk,t+1}\,, \tag{4.3}$$

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) and so 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.[16] Formally, our conceptualization assumes:

Assumption 3 *Short-term knowledge completely fades out and so* $\delta^S = 0$.

Assumption 4 *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, achievement in $t+1$ is:

$$A^*_{ijk,t+1} = \mu^L_{jt} + \mu_{k,t+1} + \beta X_{ijt} + \epsilon_{ijk,t+1}\,. \tag{4.4}$$

We now residualize $t+1$ achievement as described in Section 2.2, by estimating $\hat{\beta}$ and the subsequent teacher fixed effects, $\hat{\mu}_{k,t+1}$, using within teacher variation and then removing the effect of these observable characteristics.[17] Residualized $t+1$ test scores, $A_{ijk,t+1}$, are then given

[16] 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). To capture these components of knowledge, we can replace period $t+1$ test scores with period $t+2$ test scores. We do this exercise in Section 6.1, but find that the gains of using $t+2$ test scores to capture long-term knowledge components missed by $t+1$ test scores are limited.

[17] Residualized $t+1$ test scores are therefore given by $A_{ijk,t+1} = A^*_{ijk,t+1} - \hat{\mu}_{k,t+1} - \hat{\beta} X_{ijt}$.

by:

$$A_{ijk,t+1} = \mu_{jt}^{L} + \epsilon_{ijk,t+1}\,. \tag{4.5}$$

Under Assumption 2, we have that the expected VA of teacher k is zero and so $\mathbb{E}[\epsilon_{ijk,t+1}] = 0$. This is identical to our long-run VA estimating equation in Section 2.2 (equation (2.7)) and so is just the long-run VA measure we estimated previously.

With an estimate of the long-run VA component, μ_{jt}^{L}, in hand, we can estimate the short-run VA component using equation (4.1) (i.e., $\mu_{jt}^{S} = \mu_{jt} - \mu_{jt}^{L}$). Equivalently, since $\delta^{L} = 1$, we could sub equation (4.5) into (4.2) and rearrange:

$$A_{ijt} - A_{ijk,t+1} = \mu_{jt}^{S} + \tilde{\epsilon}_{ijkt}\,. \tag{4.6}$$

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

To recap, we have constructed three different VA measures: standard VA, μ_{jt}, long-run VA, μ_{jt}^{L}, and short-run VA, μ_{jt}^{S}. The estimating equations for each are as follows:

1. **Standard Value-Added:** $A_{ijt} = \mu_{jt} + \epsilon_{ijt}$
2. **Long-Run Value-Added:** $A_{ijk,t+1} = \mu_{jt}^{L} + \epsilon_{ijk,t+1}$.
3. **Short-Run Value-Added:** $A_{ijt} - A_{ijk,t+1} = \mu_{jt}^{S} + \tilde{\epsilon}_{ijkt}$.

4.1 Assessing the Model

We now estimate both long- and short-run VA and assess whether these estimated VA measures fit our theoretical model above.

Model Estimates and Relationships: Table A.4 reports the parameter estimates of short-run VA (alongside standard and long-run VA which we presented previously in Table 2). We see that the reliability of short-run VA is higher than that of long-run VA, implying that teachers' contributions to short-run knowledge are easier to predict than long-run knowledge contributions. Intuitively, we would expect this as there is a delay in the econometrician observing the long-run knowledge gains, decreasing our predictive power due to the presence of an additional shock.[18] Table A.4 also decomposes the standard deviation of teacher effects into their short- and long-run knowledge contributions. Here, we see that over sixty percent of the variation in standard VA comes from short-run VA and so is driven by short-term knowledge gains. This reinforces the possibility that a large portion of the variation in standard VA measures does not contribute to students' lifelong success.

Table B.1 then reports the relationship between our various VA measures. First, standard VA is highly correlated to long- and short-run VA, as expected since they are both components of the standard VA measure. Second, long- and short-run VA are uncorrelated with each other, in line with our model as each component captures a different component of knowledge. Third, the entirety of the correlation between standard VA and non-cognitive VA is driven by long-run VA; short-run VA is completely uncorrelated with non-cognitive VA. This highlights that long-run VA partially picks up components of teaching that *also* improve student behavior, while short-run VA does not.

Future Test Scores: Figure A.4 plots the impacts of long- and short-run VA measures (along with standard VA) on test scores in the current and subsequent years. (Table A.3 reports the point estimates along with standard errors that underlie Figure A.4.) We do so following the methodology set out in Section 3.2 where we regress test score residuals in year t on the VA measure.[19] In line with equation (4.1) of our model, we find that a one-unit higher long- and short-run VA teacher increases *contemporaneous* test scores by one standard deviation.

In the subsequent year, a one-unit higher long-run VA teacher continues to raise those students' test scores by one standard deviation, while a one-unit higher short-run VA teacher has no effect on the subsequent year's test scores. These results align with the fade out assumptions we made in our model (Assumptions 3 and 4). Being assigned to a high long-run VA continues to substantially affect test scores: four years out a one-unit higher long-run VA teacher raises test scores by 0.74 of a standard deviation. In contrast, short-run VA teachers continue to have no impact on future test scores. Despite using no future test scores past $t+1$ in the construction

[18] For example, a student assigned to a high long-run VA teacher may experience a shock the following period that unwinds the long-run knowledge gain. The teacher's long-term knowledge contribution is then not observed (as it does appear in subsequent test scores), worsening our predictive power.

[19] For contemporaneous test scores, we include both long- and short-run measures since they both influence contemporaneous test scores according to equation (4.2). For future test scores, adding the other VA measure makes little difference given that only long-run VA affects future test scores (as predicted by the model).

of our VA measures, long-run VA strongly impacts these test scores while short-run VA does not. This provides strong support for our model that long-run VA captures teachers' contributions to a deeper understanding of the material or improved behavior that persists well into the future, while short-run VA captures contributions to fully transitory knowledge.

Long-Term Outcomes: Figure A.5 is constructed analogously to Figure 2 and displays the impact of long- and short-run VA on our twelve high school outcomes. The long-run VA estimates are identical to those in Figure 2 and show that being assigned to a teacher whose long-run VA index is one standard deviation higher substantially improves high school outcomes. In contrast, being assigned to a teacher whose short-run VA index is one standard deviation higher has little effect on high school outcomes (if anything, it *worsens* high school outcomes). Therefore, the *entire* impact of being assigned high standard VA teachers on high school outcomes is driven by its long-run VA component.

5 Testing for Bias

All of our results to this point rely on the assumption that unobservable determinants of students' long-term outcomes are uncorrelated with teacher quality conditional on observables (Assumptions 1 and 2). In this section we verify the impact of teachers' impacts in three ways: (i) test for sorting based on twice-lagged outcomes, (ii) conduct event studies on the entry or exit of high or low VA teachers, and (iii) use quasi-experimental variation that leverages staffing changes among the full distribution of teachers.

5.1 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 test scores.

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[20] on $\mathbf{Y}_{it}^{-2}$, again including teacher fixed effects, and calculate predicted values $A_{ijt}^{Y} = \hat{\boldsymbol{\rho}}\mathbf{Y}_{it}^{-2}$. We do this procedure for both mathematics and English and stack the results, including subject fixed effects. 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 standard, long-run, and 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. The figures also plot test score residuals against the VA estimates. These plots have a slope near one indicating the one-to-one relationship between their respective test score residuals and our VA measures, mimicking our prior findings (e.g., see Table A.3).[21] Similar to Chetty et al. (2014a), we find a small positive relationship between predicted scores based on twice-lagged outcomes and standard VA: the coefficient is 0.018 (s.e. 0.002). We uncover a similar small positive relationship for long-run VA (coefficient of 0.028). Both of these coefficients are similar in magnitude to Chetty et al. (2014a) who find a coefficient of 0.022 (s.e. 0.002) in their analysis. Regardless, the relationship between long-run VA and predicted scores, A_{ijt}^{Y}, is small relative to the relationship between long-run VA and test score residuals indicating limited sorting based on twice-lagged test scores.

5.2 Event Studies

While 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 affect test scores for those students in the following year.

Methodology: Let event year '0' denote the school year a teacher enters or exits a school-

[20] The residualized test scores we use are those employed to construct the respective VA measure. Therefore, standard, long-run, and short-run VA use the residualized test scores A_{ijt}, $A_{ij,t+1}$, $A_{ijt} - A_{ij,t+1}$, respectively.

[21] In the terminology of Chetty et al. (2014a), all of our VA measures are 'forecast unbiased' since there is a one-to-one relationship between VA and the test score residuals.

grade 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. We therefore restrict our event-study to school-switchers (in comparison to Chetty et al. (2014a) who also use school-grade switchers). We do so because in our setting school-grade switchers could induce spurious event study results when we use the subsequent test score as an outcome since the teacher could switch to or have switched from the subsequent grade.[22]

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.[23] We estimate the VA of each entering teacher by excluding event years $t \in [-3, 2]$ from their VA calculation, ensuring that VA is calculated using data from students outside the six year school-grade event window.[24]

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 year (left-side figure) and for those same students in the subsequent year (right-side figure) before and after a high long-run VA teacher enters the school-grade. Similarly, the dashed series 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

[22] For example, suppose a high short-run VA teacher switches from fourth to fifth grade within the same school. Then, the exit of the teacher is associated with a decrease in fourth grade test scores *and* an increase in those students subsequent test scores in fifth grade. The increase in subsequent test scores, however, is being driven by those students being taught by the same high short-run VA teacher rather than short-run VA teachers influencing subsequent test scores.

[23] 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).

[24] 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.

before the teacher's arrival. The arrival of a high long-run VA teacher raises mean long-run VA in that school-grade by 0.02, while the arrival of a high short-run VA teacher raises mean short-run VA in that school-grade by 0.07. The higher increase in mean VA caused by the arrival of a high short- relative to long-run VA teacher is due to the fact that the standard deviation of 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 0.06σ, respectively. Both of these test score increases are 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 unbiased measures of teacher quality.

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 0.03σ, which is statistically indistinguishable from the 0.02 increase in teacher long-run VA. 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 0.003σ). 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 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.6 then conducts the teacher exit event studies. These event studies yield similar results with the hypotheses that the observed impact on subsequent test scores equals the increase in mean VA being always rejected for short-run VA, but never for long-run VA.

5.3 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: As for the event studies, we alter the methodology of Chetty et al. (2014a) to ensure that we only use variation from teachers who switch schools. Once again, failing to do so may cause teacher switchers to impact subsequent test scores (and thereby affect long-run VA) if they switched to the current grade from the subsequent grade within the same school. To ensure that only teachers who switch schools are used, we implement an instrumental variable strategy that instruments for the change in value-added at the school-grade level with the average value-added of teachers who entered and exited that school-grade from *another school*.

Formally, let Q^k_{sgt} denote the (student-weighted) mean of value-added $\hat{\mu}^k$ for VA measure $k \in \{\textit{Long-Run, Short-Run, Standard}\}$ across teachers in school s in grade g, respectively. We then define the change in mean teacher VA from year $t-1$ to year t in grade g in school s as $\Delta Q^k_{sgt} = Q^k_{sgt} - Q^k_{sg,t-1}$ for VA measure $k \in \{\textit{Long-Run, Short-Run, Standard}\}$. 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 $\Delta A^*_{sgt} = A^*_{sgt} - A^*_{sg,t-1}$. Our coefficient of interest then comes from regressing changes in mean test scores across cohorts on changes in mean teacher VA:

$$\Delta A^*_{sgt} = a + b\Delta Q^k_{sgt} + \Delta\xi_{sgt}\,, \quad k \in \{\textit{Long-Run, Short-Run, Standard}\}\,. \tag{5.1}$$

If our VA measures are unbiased, these coefficients should equal those found using the cross-sectional approach. Unfortunately, equation (5.1) captures changes in teacher VA coming from across-school and within-school across grade teacher switchers.We therefore instrument ΔQ^k_{sgt} with the change in teacher quality coming solely from teachers who switch schools.

Let $\hat{\mu}^{k,-\{t,t-1\}}_{jt}$ denote the VA estimates for teacher j in year t constructed as described in Section 2.2 using data from all years except t and $t-1$ for VA measure $k \in \{\textit{Long-Run, Short-Run, Standard}\}$. Leaving out both years t and $t-1$ when estimating VA eliminates the correlation between changes in mean test scores across cohorts t and $t-1$ and estimation error our instrument. Let n_{jt} denote the enrollment of teacher j's class in period t. We then take all teachers who enter school s in period t from another school s' in $t-1$[25] and find the enrollment-weighted VA, $\hat{Z}^{enter}_{sgt}$, of these

[25] The set s' also includes the option of not teaching. We therefore include teachers who enter school s but did

teachers in school-grade $s-g$:

$$\hat{Z}^{enter}_{sgt} = \frac{\sum_j n_{jt}\hat{\mu}^{k,-\{t,t-1\}}_{jt}\mathbb{1}\{st \neq s', t-1\}}{\sum_j n_{jt}} \,. \tag{5.2}$$

Analogously, we take all teachers who exited school s in period $t-1$ and find the enrollment-weighted VA, $\hat{Z}^{exit}_{sgt}$, that these teachers would have contributed to school-grade $s-g$ in period t:

$$\hat{Z}^{exit}_{sgt} = \frac{\sum_j n_{j,t-1}\hat{\mu}^{k,-\{t,t-1\}}_{jt}\mathbb{1}\{s't \neq st-1\}}{\sum_j n_{jt}} \,. \tag{5.3}$$

Our instrument, $\hat{Z}_{sgt}$, is then given as the change in VA in school-grade $s-g$ at period t coming from teachers that enter and exit school s: $\hat{Z}_{sgt} = \hat{Z}^{enter}_{sgt} - \hat{Z}^{exit}_{sgt}$. We then use $\hat{Z}_{sgt}$ as an instrument for ΔQ^k_{sgt} in equation (5.1).

We use the same methodology to find quasi-experimental estimates of the impact of our three VA measures on other outcomes. For instance, define $A^*_{s,g+1,t+1}$ as 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 $\Delta A^*_{s,g+1,t+1} = A^*_{s,g+1,t+1} - A^*_{s,g+1,t}$. Quasi-experimental estimates of teacher VA on subsequent test scores are then found by regressing changes in *subsequent* test scores across cohorts on changes in mean teacher VA:

$$\Delta A^*_{sg,t+1} = a + b\Delta Q^k_{sgt} + \Delta\xi_{sgt}\,, \quad k \in \{\textit{Long-Run, Short-Run, Standard}\}\,, \tag{5.4}$$

where we once again use $\hat{Z}_{sgt}$ as an instrument for ΔQ^k_{sgt}. Similarly, we replace the change in subsequent test scores in equation (5.1) with changes in high school outcome residuals to check for potential bias in our estimated impacts of the VA measures on long-run outcomes.

Results for Test Scores: Figure 6 displays the quasi-experimental estimates for the teacher VA measures for the current period and the next four periods alongside the cross-sectional estimates for comparison. The quasi-experimental estimates come from equations (5.1) and (5.4), using changes in school-grade VA coming from teachers who switch schools ($\hat{Z}_{sgt}$) as an instrument for the change in teacher quality. Whiskers in the figure indicate 95% confidence intervals. Our

not teach in the prior year as part of our identifying variation for $\hat{Z}^{enter}_{sgt}$.

quasi-experimental estimates in the current period (i.e., $t = 0$) are statistically indistinguishable from one for standard, long-run, and short-run VA. For the *subsequent* year (i.e., $t = 1$), the quasi-experimental estimates cannot reject that long-run VA leads to a one standard deviation improvement in test scores, while short-run VA has no effect. These results align precisely with our cross-sectional findings. Similarly, the quasi-experimental point estimates in future periods mirror our cross-sectional results demonstrating that being assigned to high long-run VA teachers lead to large improvements in future test scores, whereas assignment to high short-run VA teachers do not. In fact, for nearly all point estimates in Figure 6 one cannot reject that the quasi-experimental and cross-sectional estimates are identical.

Results for High School Outcomes: Table 5 reports the effect of teacher VA on high school outcomes using the quasi-experimental variation, with cross-sectional estimates provided for comparison. While the quasi-experimental estimates are noisy, they remain quite close to their cross-sectional counterparts and for most outcomes one cannot reject that the quasi-experimental and cross-sectional estimates are identical. These results therefore indicate that our teacher VA measures do not feature substantial bias and that long-run VA leads to large improvements in high school 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 released from teaching, 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 standard VA. Figure 7 displays a scatter plot of teacher quality as measured by standard 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 there are many teachers who are released under standard VA who would not be released if long-run VA were used instead (and vice versa). 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 standard 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 standard VA are not released under a policy based on long-run VA.

Policy Efficiency: Since using long-run VA in place of standard VA affects who is released, the quality of teachers in terms of long-run VA will improve if policymakers use long-run over standard VA in high-stakes decision-making. Given that high long-run VA teachers are better at improving long-run outcomes, 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 standard and long-run VA following the methodology in Chetty et al. (2014b). In particular, we ignore general equilibrium effects and assume that replacement teachers are of mean quality. We do not, however, require the distribution of teacher quality to be normally distributed by leveraging recent computational advances in nonparametric MLE (Koenker and Mizera, 2014; Gilraine et al., 2020). 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 VA. Given that teacher VA is not observed in practice, we then calculate the gains from feasible policies that select teachers based on VA estimates.

Our estimates from Section 3.2 of the impact of being assigned to a teacher whose standard 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:

$$G^{\kappa} = \Delta m_{\sigma}^{\kappa} \times \rho^{\kappa}\ , \quad \kappa \in \{\mathit{Long\text{-}run}, \mathit{Standard}\}, \tag{6.1}$$

where ρ^{κ} is our estimated impact of a one standard deviation higher teacher, with the superscript k representing which VA measure is used. $\Delta m_{\sigma}^{\kappa}$ represents the average improvement in VA (normalized by the standard deviation and measured in terms of test scores) of the policy, which is given by $\mathbb{E}[m^k | m^k < F_k^{-1}(0.05)]$, where $F_k(\cdot)$ is the cdf of the teacher VA distribution.[26]

[26] Under normality, the expected value of VA conditional on being a teacher in the bottom five percent would

Table A.6 reports the policy gains in terms of twelve high school outcomes. The policy gains from using long-run VA in place of standard VA are substantial: point estimates indicate three- to five-fold improvement to policy effectiveness in terms of high school outcomes.[27] 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 over one million dollars.

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 as VA is estimated with error. Given that long-run VA features more estimation error (see Table 2), we expect that policy gains when using estimated long-run VA will underperform their true VA benchmarks more than those of standard 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:

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

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 calculate the expected value using Monte Carlo simulations.[28] 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 $t+1$.

Figure 8 plots the mean gain per classroom of releasing bottom five percent teachers according to standard and long-run VA in terms of four outcomes that are particularly salient to policymakers: PSAT and SAT scores, high school graduation and SAT-taking (as PSAT and

be 2.06, since $\Delta m^{k}_{\sigma} = \mathbb{E}[m^{k} | m^{k} < \Phi^{-1}(0.05)] = 2.06$ (where $\Phi(\cdot)$ is the cdf of the normal distribution).

[27] A small portion of the improvement comes from the fact that while the distribution of long-run VA is roughly normal (as $\Delta m^{\textit{Long-run}}_{\sigma}$ roughly equals 2.06, as one would expect under normality), the standard VA measure has a thin left tail and so the average change in teacher quality is lower (1.79) compared to a normal distribution (2.06).

[28] We calculate the conditional expectation in equation (6.2) in three steps. First, we sample 250,000 teacher observations from the estimated distribution F. Second, for each sample teacher observation, we generate noisy data $A_j = \mu_j + \epsilon_j$ assuming $\epsilon_j \sim \mathcal{N}(0, \sigma^2_{\epsilon}/(k \cdot n))$, where n represents yearly class sizes (set at 24) and k the number of years of data for each teacher (which we vary). Third, we estimate teacher VA and calculate $\frac{1}{250000 \times 0.05} \sum_j m^{\kappa}_{j,n+1} 1\{m^{\kappa}_{j,n+1} \leq \hat{F}^{-1}_{\hat{m}^{\kappa}_{j,n+1}}(0.05)\}$ as an estimator for $\mathbb{E}\left[m^{\kappa}_{j,n+1} | \hat{m}^{\kappa}_{j,n+1} < F^{-1}_{\hat{m}^{\kappa}_{j,n+1}}(0.05)\right]$. By the law of large numbers, this produces a consistent estimate.

SAT scores along with SAT-taking are highly-related to college-going).[29] The gains from releasing teachers based on their true standard and long-run VA are shown by the horizontal lines in the figures. The gains from releasing teachers based on estimated VA are then shown by the series and 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 the four outcomes relative to using standard VA, even when the policymaker can observe true standard 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.

The gains of using long-run VA over standard VA are substantial. After three years of data, an additional 0.26 students per classroom graduates high school when a bottom five percent teacher according to long-run VA is released. In comparison, a policymaker releasing teachers based on standard VA would only see an additional 0.05 students per classroom graduating. We see similar substantial gains from using long-run over standard VA for SAT-taking as well as SAT and PSAT scores. The relative gains of using long-run VA become more pronounced as more years of data are incorporated into the teacher release decision.

6.1 Incorporating Longer-Run Value-Added Measures

While policies based on long-run VA outperform those using standard VA, a natural question is whether alternative VA measures can do even better. One such VA measure that comes to mind is a longer-run measure that is based on $t+2$ outcomes, which we call '$t+2$ VA.' To assess its performance, we estimate '$t+2$ VA' by replacing the $t+1$ test scores used to estimate long-run VA with their $t+2$ counterparts and including an additional fixed effect for the student's teacher two years into the future.

We find that '$t+2$ VA' is highly-correlated with our long-run VA measure. In particular, we find that the correlation between the two measures is 0.75, which is a much higher correlation than the 0.51 correlation found between long-run and standard VA. The high degree of correlation suggests that '$t+2$ VA' and long-run VA are capturing similar contributions made by

[29] While we restrict our attention to these four outcomes in our discussion here, the gains for other outcomes can be found in Table A.7 which reports policy the gains for all outcomes using three years of data per teacher.

teachers. Indeed, our conceptual framework in Section 4 would suggest exactly this: both '$t+2$ VA' and long-run VA capture teachers' contributions to long-term knowledge.

Figure A.7 assesses the policy gains from using '$t+2$ VA' over long-run VA. The figure plots the mean gain per classroom of releasing bottom five percent teachers according to the two VA measures in terms of four outcomes (PSAT and SAT scores, high school graduation and SAT-taking). As before, the horizontal lines indicate policy gains if policymakers could release based on a teacher's true VA, while the series indicate gains when policymakers' decisions are based on estimated VA. While using '$t+2$ VA' does yield some gains for the SAT-taking outcome, long-run VA leads to higher policy gains for the PSAT and SAT scores outcomes. Overall, the policy gains are similar whether '$t+2$ VA' or long-run VA are used.

Table A.7 then compares the policy gains when using each of the three VA measures – standard, long-run, and '$t+2$ VA' – for each of the 12 high school outcomes. In this table, we consider a teacher who has just finished her third year at a school and report the policy gain when making the teacher release decision using each of the VA measures. For all outcomes, using long-run VA leads to significant policy gains over standard VA and similar policy gains compared to using 't+2 VA.' While the long-run and '$t+2$ VA' lead to significant policy gains, additional policy gains may also be gleaned by combining the information from all three of these measures (as well as non-cognitive VA) from each year of available data instead of using each of these measures individually.[30]

7 Conclusion

This paper decomposes VA into 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 main goal of educators and policymakers is to improve the lifelong success

[30] Work by Mulhern and Opper (2021) looks at how to best combine multiple dimensions of teacher effectiveness into metrics that can be used for personnel decisions.

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 high school outcomes, on average, increase by more than 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.

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. This might encourage the adoption of teaching methods that are more effective at imparting long-lasting cognitive and non-cognitive skills that will better influence students' test scores in the following year. Investigating teachers' response to incentive schemes in terms of teaching practices is something that we are exploring in related work in designing optimal dynamic incentive schemes (Macartney, 2016; Gilraine, 2018).

References

Algan, Yann, Pierre Cahuc, and Andrei Shleifer (2013), "Teaching practices and social capital." *American Economic Journal: Applied Economics*, 5, 189–210.

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 (2021), "The labor market for teachers under different pay schemes." *American Economic Journal: Economic Policy*, 13, 63–102.

Bietenbeck, Jan (2014), "Teaching practices and cognitive skills." *Labour Economics*, 30, 143–153.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller (2012), "Robust inference with multiway clustering." *Journal of Business & Economic Statistics*.

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.

Cordero, Jose M. and María Gil-Izquierdo (2018), "The effect of teaching strategies on student achievement: An analysis using TALIS-PISA-link." *Journal of Policy Modeling*, 40, 1313–1331.

Flèche, Sarah (2017), "Teacher quality, test scores and non-cognitive skills: Evidence from primary school teachers in the UK." Discussion Paper 1472, Centre for Economic Performance, URL http://eprints.lse.ac.uk/83602/1/dp1472.pdf.

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.

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.

Koenker, Roger and Ivan Mizera (2014), "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association*, 109, 674–685.

Korbel, Václav and Michal Paulus (2018), "Do teaching practices impact socio-emotional skills?" *Education Economics*, 26, 337–355.

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

Lavy, Victor (2016), "What makes an effective teacher? Quasi-experimental evidence." *CESifo Economic Studies*, 62, 88–125.

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.

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

Mulhern, Christine and Isaac M. Opper (2021), "Measuring and summarizing the multiple dimensions of teacher effectiveness." EdWorkingPaper 21-451, URL https://www.edworkingpapers.com/sites/default/files/ai21-451.pdf.

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.

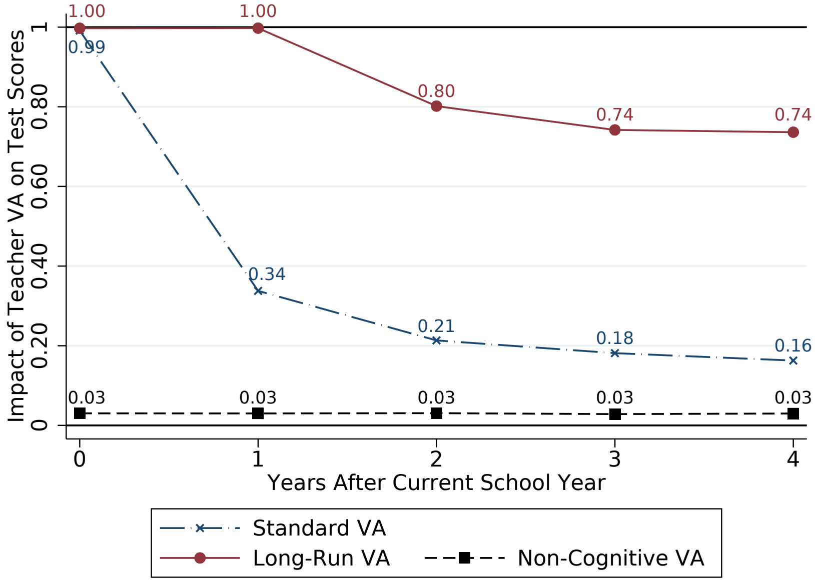
Rothstein, Jesse (2010), "Teacher quality in educational production: Tracking, decay, and student achievement." *Quarterly Journal of Economics*, 125, 175–214.

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

Schwerdt, Guido and Amelie C. Wuppermann (2011), "Is traditional teaching really all that bad? A within-student between-subject approach." *Economics of Education Review*, 30, 365–379.

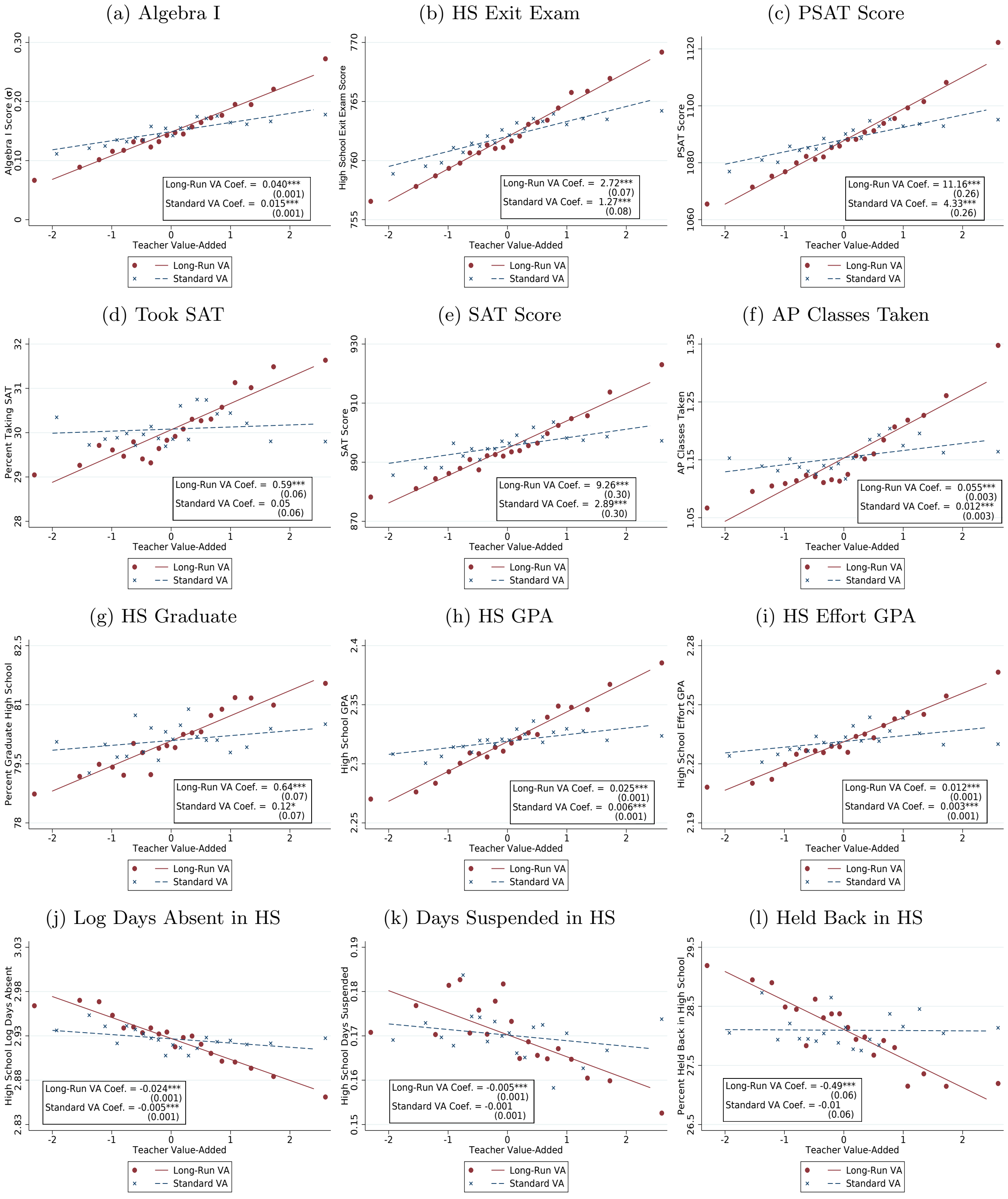
Stepner, Michael (2013), "Vam: Stata module to compute teacher value-added measures." URL http://fmwww.bc.edu/RePEc/bocode/v/vam.ado.

Figure 1: Effects of Standard, Long-Run, and Non-Cognitive Value-Added on Future Test Scores



Notes: This figure shows the effect of teacher standard, long-run, and non-cognitive VA on future test scores. The figure is constructed by regressing residualized end-of-grade math and English test scores in t years after being with teacher j on teacher j's VA measure in that subject as described by equation (3.3). (Since non-cognitive VA is not subject-specific, it is just regressed on the mean of residualized end-of-grade math and English test scores as described by equation (3.4).) When regressing long-run VA on contemporaneous test scores we also control for short-run VA. Point estimates from our regressions are reported above each point. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (3.2). The coefficients and standard errors of the point estimates underlying the figure are reported in Table A.3.

Figure 2: Effect of Standard and Long-Run Value-Added on High School Outcomes



(a) Algebra I

(b) HS Exit Exam

(c) PSAT Score

(d) Took SAT

(e) SAT Score

(f) AP Classes Taken

(g) HS Graduate

(h) HS GPA

(i) HS Effort GPA

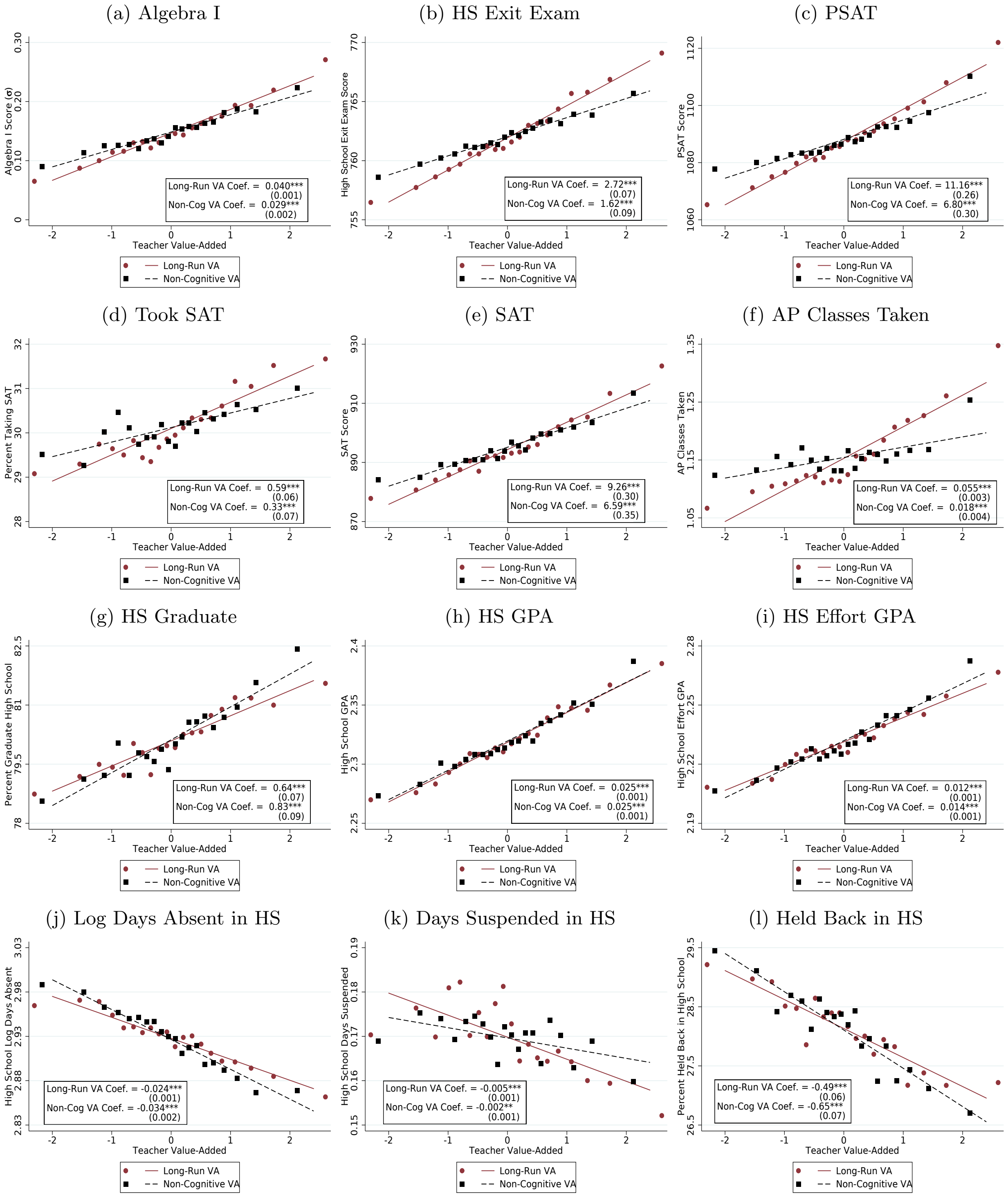
(j) Log Days Absent in HS

(k) Days Suspended in HS

(l) Held Back in HS

Notes: This figure shows the effect of teacher VA on high school outcomes for the standard and long-run value-added indices. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the standard or long-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school 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^{κ} from equation (3.4), which represent the effect of being assigned to a teacher whose standard or long-run VA is one standard deviation higher in a single grade on high school outcomes, along with its standard errors in brackets below. Effects for long-run VA are the same as those reported in Figures 3 and A.5. 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 High School Outcomes


(a) Algebra I

(b) HS Exit Exam

(c) PSAT

(d) Took SAT

(e) SAT

(f) AP Classes Taken

(g) HS Graduate

(h) HS GPA

(i) HS Effort GPA

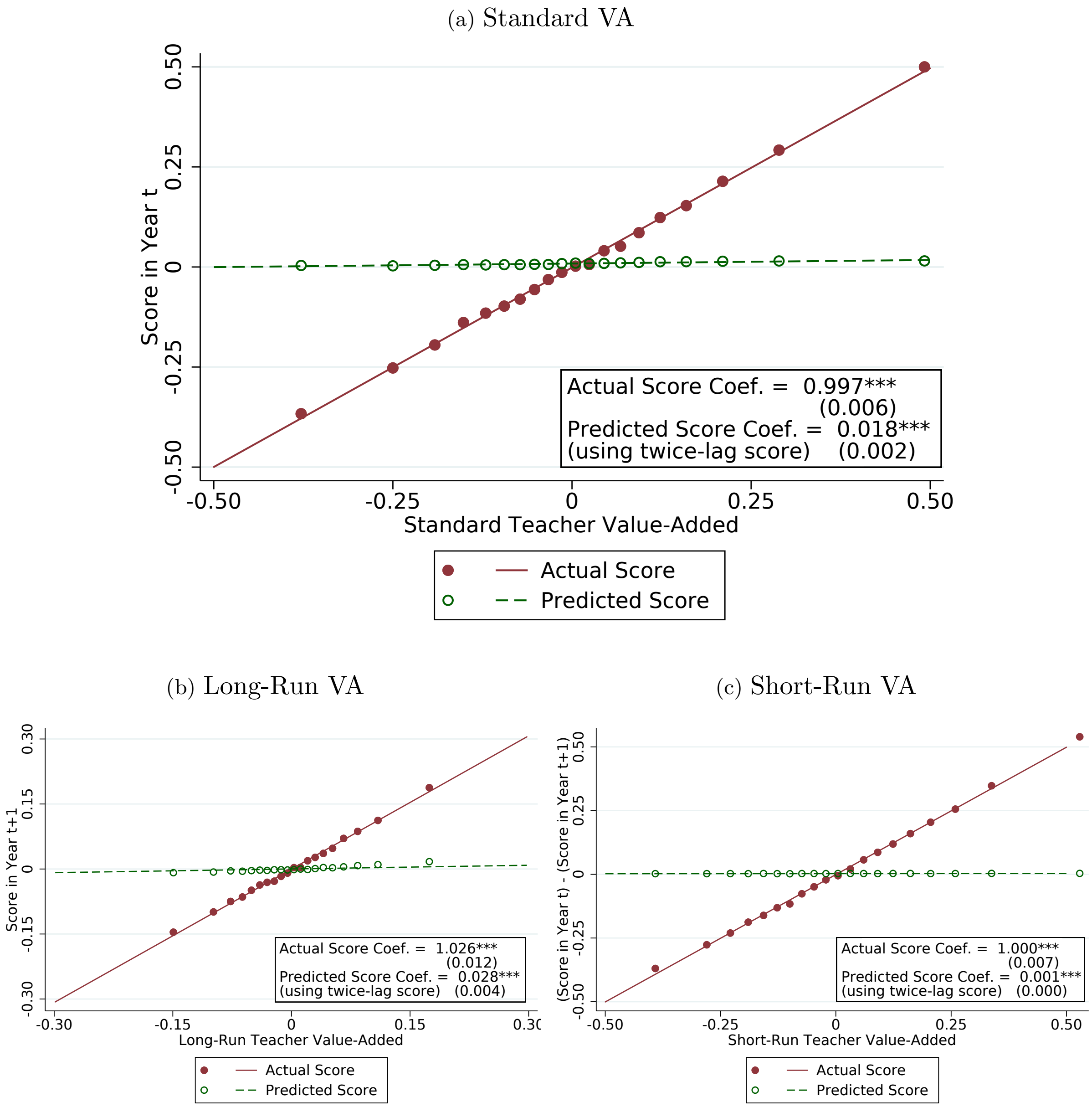
(j) Log Days Absent in HS

(k) Days Suspended in HS

(l) Held Back in HS

Notes: This figure shows the effect of teacher VA on high school outcomes for the non-cognitive and long-run value-added indices. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the standard or long-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school 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^{κ} from equation (3.4), 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 high school outcomes, along with its standard errors in brackets below. Effects for long-run VA are the same as those reported in Figures 2 and A.5. 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 Standard, Long-Run and Short-Run VA on Actual and Predicted Scores

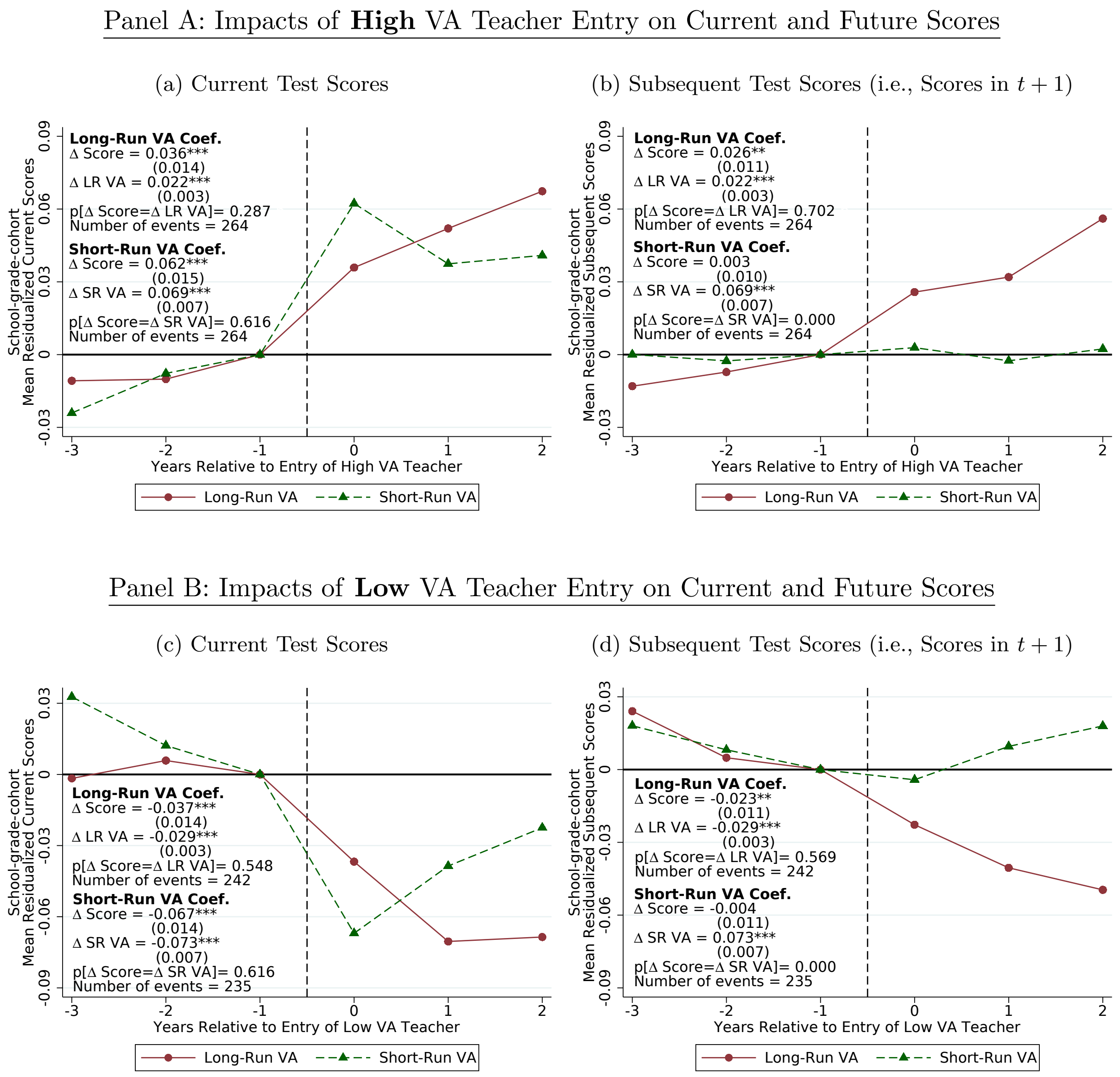

(a) Standard VA

(b) Long-Run VA

(c) Short-Run VA

Notes: These figures assess whether students sort on variables that predict test score residuals but are omitted from the VA models. We predict scores based on twice-lagged test score outcomes separately by subject. 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 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 – A_{ijt}, $A_{ij,t+1}$, $A_{ijt} - A_{ij,t+1}$ for standard, long-run, and short-run VA, respectively – against the VA estimates. Point estimates for the slope of this line are close to one, indicating the one-to-one relationship between their respective test score residuals and our VA measures, mimicking our prior findings (e.g., see Table A.3). 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 student and classroom 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) Subsequent Test Scores (i.e., Scores in $t + 1$)

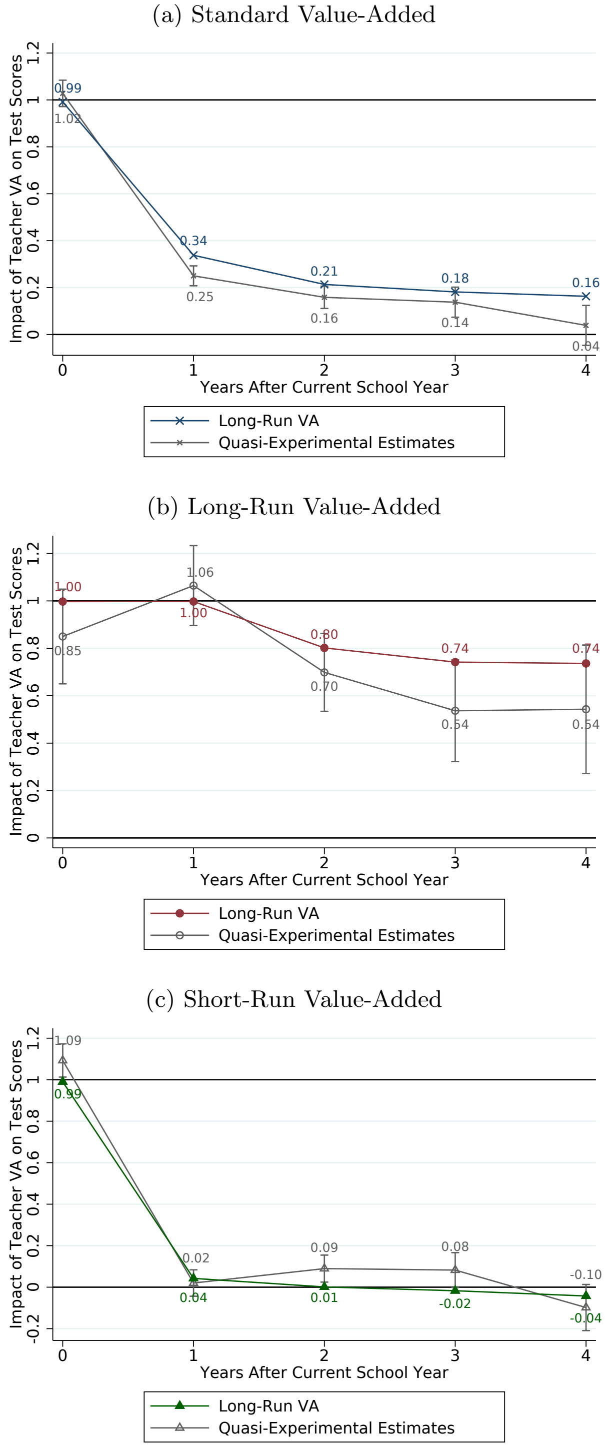
Panel B: Impacts of **Low** VA Teacher Entry on Current and Future Scores

(c) Current Test Scores

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

Notes: These figures plot event studies of (residualized) test scores 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 consists of two series whereby VA is measured as long-run (solid series) and short-run VA (dashed series). The y-axis of the left-hand side figures is 'current scores,' the residualized mean contemporaneous test scores of the school-grade-cohort, while right-hand side figures y-axis is 'subsequent scores,' the residualized mean test scores of the school-grade-cohort in the *following* year. Test scores are residualized using the control vector defined in Section 2.2 and include subsequent teacher fixed effects. To construct each panel we: (i) identify the set of teachers who entered a school-grade-subject cell and did not teach at that school for the preceding three years (i.e., we use only within-school switcher variation) and define event time as the school year relative to the year of entry, (ii) estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$ from their VA calculation, (iii) classify high- and low-VA teachers as those with VA estimates in the top or bottom 5% of the distribution among teachers who entered schools in that year, (iv) 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. The test score changes at event year -1 are normalized to zero and we include grade-by-year fixed effects to eliminate secular time trends. 'Δ Score' reports the change in current or future test scores from 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 from 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.6 reports results from similar event studies that leverage teacher exit.

Figure 6: Cross-Sectional vs. Quasi-Experimental Estimates of Standard, Long-Run and Short-Run Value-Added on Future Test Scores



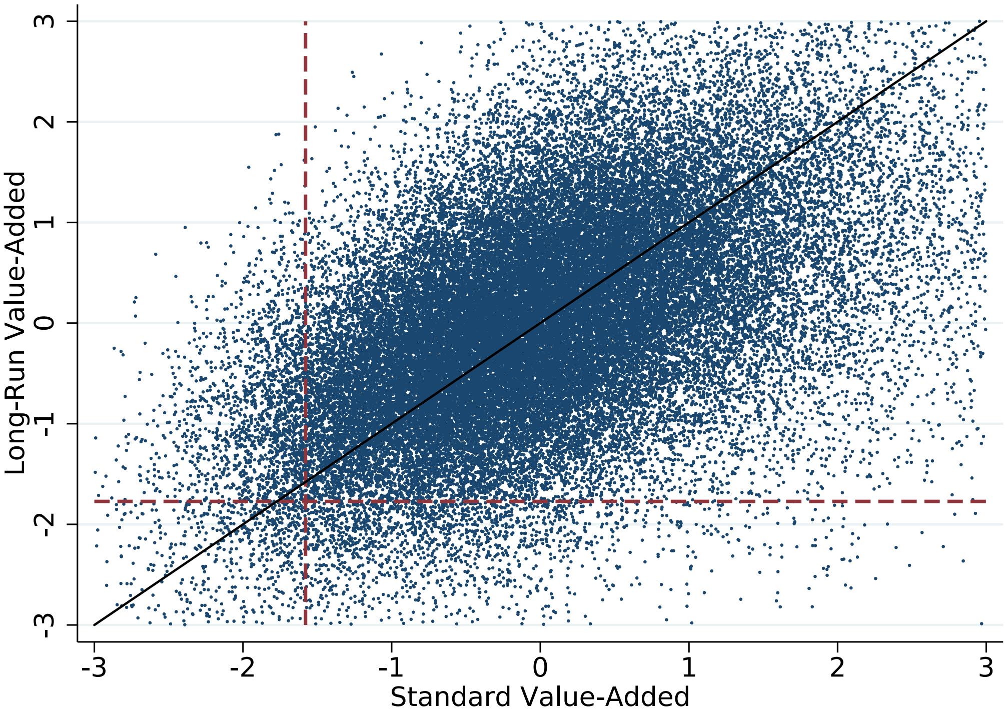
(a) Standard Value-Added

(b) Long-Run Value-Added

(c) Short-Run Value-Added

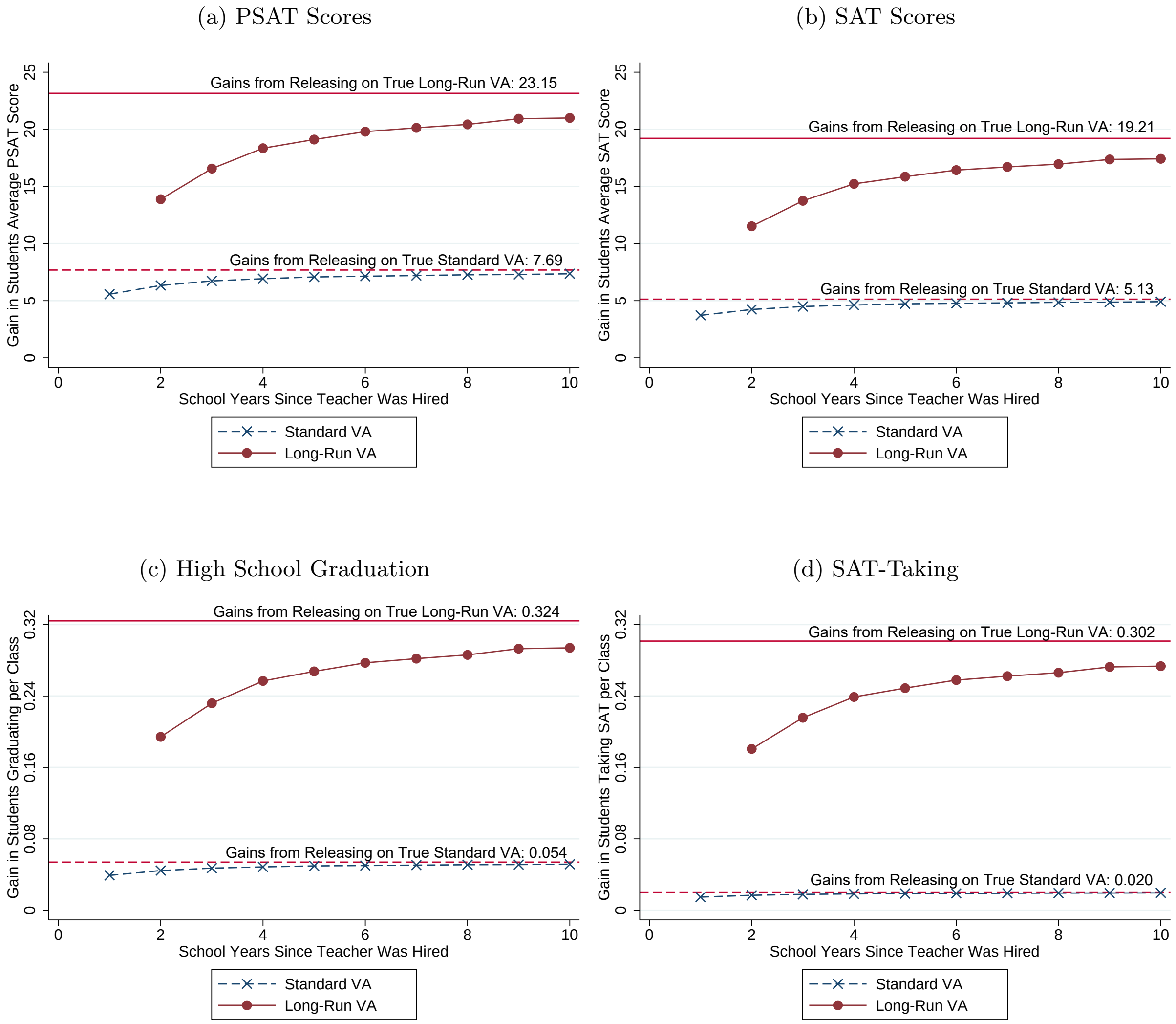
Notes: These figures compare the cross-sectional estimates of VA on future test scores from Figure 1 (and A.4 for short-run VA) to quasi-experimental estimates that leverage teacher school-switchers. The cross-sectional estimates are identical to those shown in Figures 1 and A.4 reported in Table A.3. The quasi-experimental estimates come from instrumental variable regressions that regress changes in school-grade mean residualized test scores across cohorts against changes in mean teacher VA, instrumenting for the change in mean teacher VA using the change in teacher VA coming from school-switchers as described in equation (5.1). Point estimates are reported above or below each point. The whiskers represent 95% confidence intervals for our quasi-experimental estimates with standard errors clustered at the school-cohort level. (Standard errors for the cross-sectional estimates are negligible.)

Figure 7: Two-Dimensional Cross Teacher Value-Added Plots



Notes: This figure plots standard 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 standard 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 standard VA measure is used, but not if long-run VA were used instead.

Figure 8: Impacts of Releasing Low Value-Added Teachers


(a) PSAT Scores

(b) SAT Scores

(c) High School Graduation

(d) SAT-Taking

Notes: This figure calculates the impact of replacing teachers with standard or long-run VA in the bottom 5% of the VA distribution with teachers of average quality on average PSAT and SAT scores and the number of high school graduates and SAT takers from a single classroom of average size (24.5 students). The horizontal lines show the hypothetical gain in the current school year from releasing the bottom 5% teachers according to their true normal (dashed) and long-run (solid) VA, with the value of this line reported above it. 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. All values in these figures use VA that equal weights teachers' mathematics and English contributions, which are calculated separately, and are based off the estimated increase in PSAT scores, SAT scores, graduation, and SAT-taking in Table A.6.

Table 1: Summary Statistics

	Full Sample[1]	Standard Value-Added Sample[2]	Long-Run Value-Added Sample[3]
	(1)	(2)	(3)
Mean of Outcomes and Student Characteristics			
Cognitive Outcomes:			
Math Score (σ)[4]	0.03	0.06	0.07
Reading Score (σ)[4]	0.03	0.05	0.06
Non-Cognitive Outcomes:			
Log Days Absent	1.50	1.51	1.49
GPA	2.88	2.90	2.90
'Effort' GPA	3.14	3.16	3.16
% Suspended	2.26	2.19	2.07
% Repeating Grade	0.65	0.49	0.46
Demographics:			
% Hispanic	74.3	75.6	76.4
% Black	10.1	9.2	8.5
% White	8.9	8.6	8.4
% Asian	4.3	4.3	4.3
% Free or Reduced Price Lunch	69.2	70.8	71.2
% English Learners	30.2	30.7	31.2
Parental Education:[5]			
% High School Dropout	35.6	35.9	36.5
% High School Graduate	27.1	27.4	27.5
% College Graduate	19.6	19.1	18.7
# of Students	649,694	552,517	501,212
# of Teachers	15,155	12,975	12,975
Observations (student-year)	1,452,367	1,192,043	1,066,783

Notes:

[1] Data coverage: third through fifth grades from 2003-04 through 2011-12.

[2] Standard 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 or English scores.

[3] Long-run VA sample restrictions: Same as standard VA, but also must be assigned to valid subsequent teacher or school-grade and have valid subsequent math or English test scores.

[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.' '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 Standard and Long-Run VA

Sample:	Mathematics		English	
Value-Added Measure:	Standard	Long-Run	Standard	Long-Run
	(1)	(2)	(3)	(4)
Autocorrelation Vector				
Lag 1	0.66	0.31	0.53	0.27
Lag 2	0.61	0.24	0.49	0.21
Lag 3	0.57	0.22	0.44	0.20
Lag 4	0.53	0.19	0.41	0.19
Lag 5	0.51	0.19	0.38	0.19
Lag ≥ 6	0.48	0.19	0.35	0.20
Within-year variance components				
Total SD	0.597	0.580	0.522	0.531
Individual-level SD	0.507	0.558	0.475	0.517
Class + teacher level SD	0.315	0.158	0.217	0.123
Estimates of teacher SD				
Lower bound based on lag 1	0.270	0.111	0.175	0.086
Quadratic estimate	0.282	0.121	0.185	0.092
Student-Year Observations	1,187,231	1,061,125	1,183,484	1,054,042

Notes: This table gives the autocorrelation estimates across years for the same teacher used to compute standard and long-run VA for both mathematics and English. 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.

Table 3: Correlation of Teacher Value-Added Measures

VA Measure	Standard VA	Long-Run VA	Non-Cognitive VA
Standard VA	1		
Long-Run VA	0.507	1	
Non-Cognitive VA	0.205	0.394	1

Notes: This table reports the correlations between our various value-added measures. Each VA measure is constructed as described in Section 2.2. In particular, our test score VA measures (standard and long-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. 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.2 reports the full correlation matrix between all the components that make up the various VA measures.

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

Outcome:	Algebra	HS Exit	PSAT	Took	SAT	# AP
	Score (σ)	Exam	Score	SAT (%)	Score	Courses
	(1)	(2)	(3)	(4)	(5)	(6)
Sample Mean	0.149	762.1	1088.2	30.4	895.2	1.15
(s.d.)	(0.761)	(40.5)	(149.6)	(43.1)	(115.2)	(1.76)
Panel A. High School Outcomes I						
Long-Run VA	0.039***	2.66***	11.46***	0.71***	9.57***	0.067***
(s.e.)	(0.002)	(0.09)	(0.32)	(0.08)	(0.37)	(0.004)
Standard VA	-0.009***	-0.32***	-2.48***	-0.38***	-3.15***	-0.024***
(s.e.)	(0.002)	(0.09)	(0.30)	(0.07)	(0.35)	(0.004)
Non-Cognitive VA	0.014***	0.55***	2.33***	0.09	2.87***	-0.006*
(s.e.)	(0.002)	(0.09)	(0.31)	(0.07)	(0.37)	(0.004)
Observations	402,151	387,076	465,125	648,303	203,414	489,343

Outcome:	Graduated	HS	HS Effort	Log Days	Days	Held Back
	HS (%)	GPA	GPA	Absent	Suspended	in HS (%)
	(7)	(8)	(9)	(10)	(11)	(12)
Sample Mean	80.1	2.32	2.23	2.93	0.17	28.1
(s.d.)	(36.7)	(0.77)	(0.42)	(1.02)	(0.71)	(41.1)
Panel B. High School Outcomes II						
Long-Run VA	0.59***	0.025***	0.011***	-0.020***	-0.005***	-0.51***
(s.e.)	(0.09)	(0.001)	(0.001)	(0.002)	(0.001)	(0.08)
Standard VA	-0.32***	-0.011***	-0.005***	0.011***	0.002*	0.36***
(s.e.)	(0.09)	(0.001)	(0.001)	(0.002)	(0.001)	(0.07)
Non-Cognitive VA	0.64***	0.016***	0.011***	-0.027***	-0.000	-0.50***
(s.e.)	(0.09)	(0.001)	(0.001)	(0.002)	(0.001)	(0.08)
Observations	302,946	533,507	457,951	554,291	566,954	522,883

Notes: This table reports the effect of a standard deviation increase in long-run, standard, and non-cognitive VA on students' residualized long-run outcomes as described by equation (3.5) to check whether each VA measure independently affects long-run outcomes. Outcomes are residualized with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2). 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 on High School Outcomes: Cross-Sectional vs. Quasi-Experimental Estimates

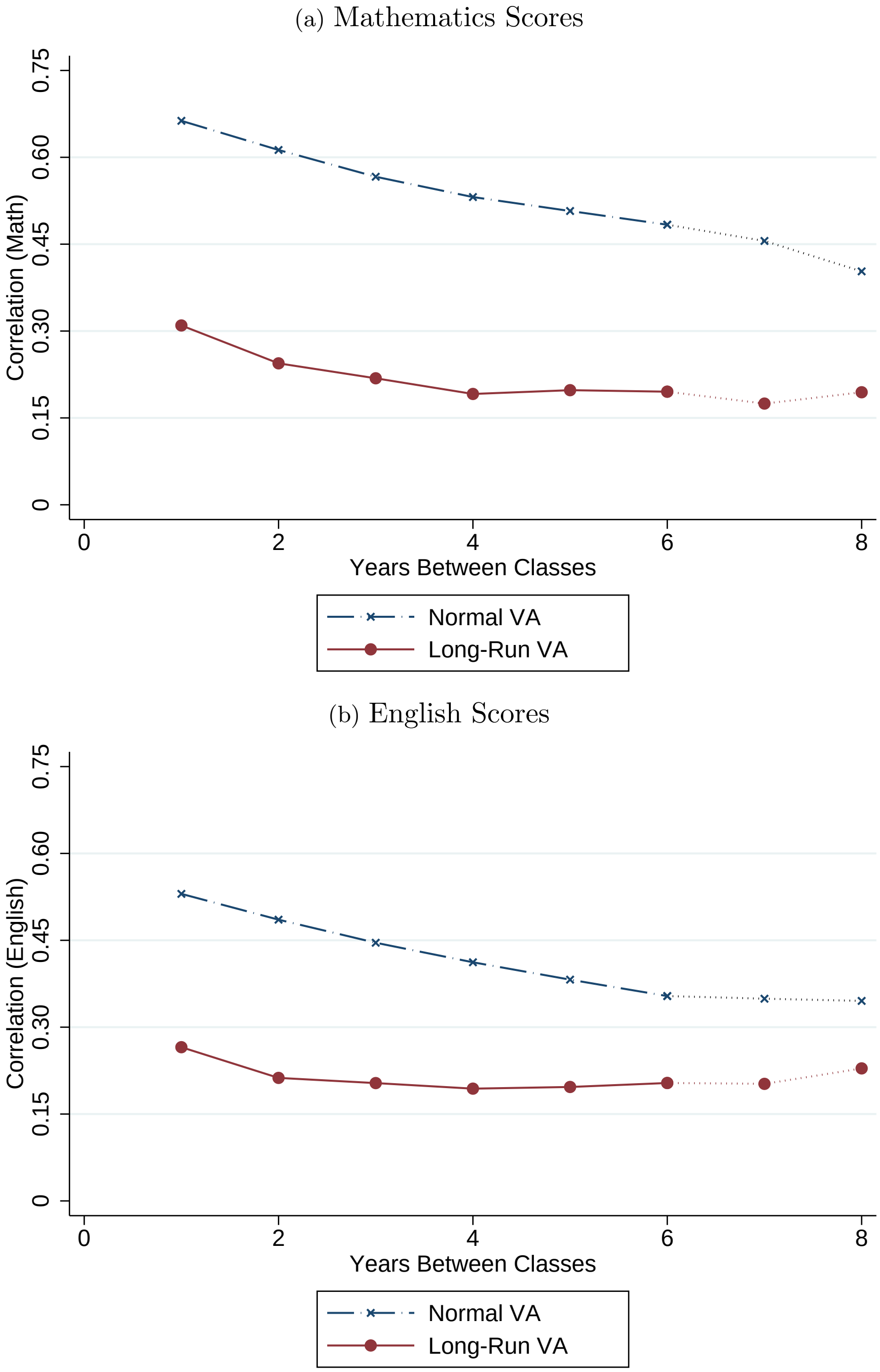
Outcome:	Algebra Score (σ)	HS Exit Exam	PSAT Score	Took SAT (%)	SAT Score	# AP Courses
A. Standard Value-Added						
Cross-Sectional	0.015***	1.27***	4.33***	0.05	2.89***	0.012***
Quasi-Experimental	0.015*	1.33***	3.02**	0.13	4.18**	0.015
(s.e.)	(0.009)	(0.41)	(1.46)	(0.32)	(1.67)	(0.017)
[95% CI]	[-0.002,0.032]	[0.52,2.15]	[0.16,5.87]	[-0.49,0.76]	[0.90,7.45]	[-0.018,0.047]
B. Long-Run Value-Added						
Cross-Sectional	0.040***	2.72***	11.16***	0.59***	9.26***	0.055***
Quasi-Experimental	0.020**	2.07***	6.71***	0.51	5.80***	0.006
(s.e.)	(0.010)	(0.48)	(1.75)	(0.41)	(1.94)	(0.019)
[95% CI]	[-0.000,0.041]	[1.13,3.00]	[3.28,10.14]	[-0.30,1.33]	[2.00,9.60]	[-0.032,0.044]
C. Short-Run Value-Added						
Cross-Sectional	-0.008***	-0.24***	-1.96***	-0.36***	-2.77***	-0.020***
Quasi-Experimental	0.006	0.47	-0.08	-0.18	1.95	0.014
(s.e.)	(0.009)	(0.43)	(1.49)	(0.34)	(1.76)	(0.018)
[95% CI]	[-0.012,0.024]	[-0.38,1.32]	[-3.00,2.85]	[-0.84,0.48]	[-1.55,5.40]	[-0.021,0.048]

Outcome:	Graduated HS (%)	HS GPA	HS Effort GPA	Log Days Absent	Days Suspended	Held Back in HS (%)
A. Standard Value-Added						
Cross-Sectional	0.12*	0.006***	0.003***	-0.005***	-0.001	-0.01
Quasi-Experimental	-0.26	-0.002	-0.002	-0.007	-0.005	-0.25
(s.e.)	(0.45)	(0.006)	(0.005)	(0.008)	(0.006)	(0.38)
[95% CI]	[-1.14,0.62]	[-0.014,0.010]	[-0.012,0.007]	[-0.023,0.008]	[-0.016,0.006]	[-1.00,0.049]
B. Long-Run Value-Added						
Cross-Sectional	0.64***	0.025***	0.012***	-0.024***	-0.005***	-0.49***
Quasi-Experimental	0.67	0.013*	0.003	-0.013*	-0.008*	-0.92**
(s.e.)	(0.46)	(0.007)	(0.005)	(0.009)	(0.005)	(0.40)
[95% CI]	[-0.23,1.57]	[-0.001,0.027]	[-0.008,0.013]	[-0.030,0.004]	[-0.018,0.002]	[-1.71,-0.13]
C. Short-Run Value-Added						
Cross-Sectional	-0.29***	-0.009***	-0.004***	0.009***	0.002**	0.32***
Quasi-Experimental	-0.65	-0.011*	-0.006	-0.003	0.000	0.13
(s.e.)	(0.49)	(0.006)	(0.005)	(0.009)	(0.006)	(0.39)
[95% CI]	[-1.61,0.31]	[-0.023,0.001]	[-0.015,0.004]	[-0.020,0.014]	[-0.012,0.012]	[-0.64,0.89]

Notes: This table reports the effect of teacher VA on high school outcomes using both cross-sectional and quasi-experimental variation. The cross-sectional point estimates report the results of equation (3.4) which regresses high school outcome residuals on teacher VA. The cross-sectional estimates for standard and long-run VA are identical to those reported in Figure 2 (and Figure A.5 for short-run VA). The quasi-experimental point estimates come from regressing changes in school-grade high school outcome residuals on changes in school-grade VA, instrumenting for the change in school-grade VA with the VA of entering and exiting teachers (i.e., equation (5.4) with high school outcome residuals as the dependent variable, instrumenting ΔQ_{sgt}^{k} with $\hat{Z}_{sgt}$). As standard errors for the cross-sectional results are small, we do not report them (they are reported in Figures 2 and A.5). Standard errors for the quasi-experimental estimates are clustered at the school-cohort level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

A Appendix Figures and Tables

Figure A.1: Autocorrelation for Mathematics and English Scores

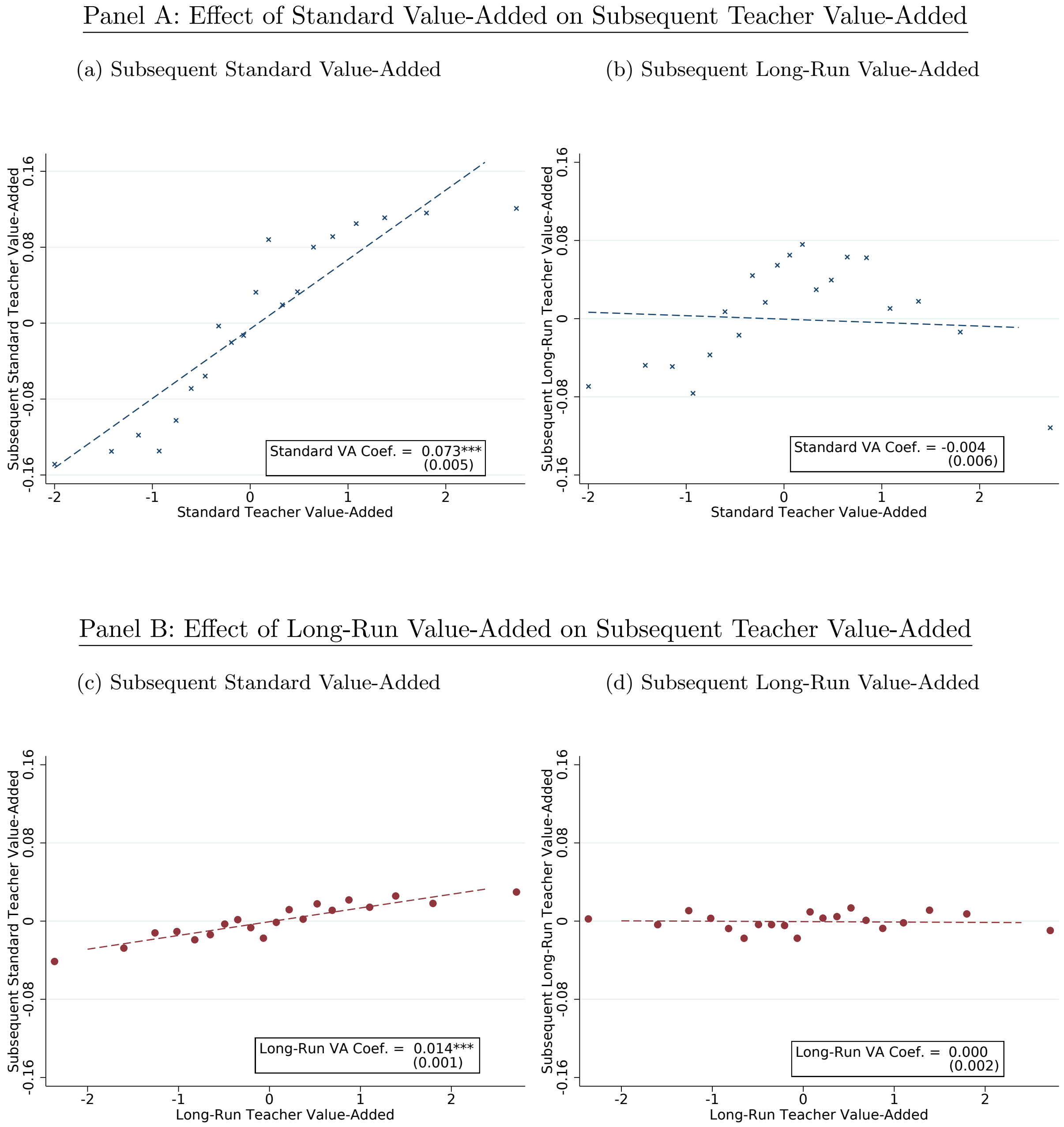


(a) Mathematics Scores

(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 for the point estimates underlying these figures.

Figure A.2: Effect of Standard and Long-Run Value-Added on Subsequent Teacher Value-Added



Panel A: Effect of Standard Value-Added on Subsequent Teacher Value-Added

(a) Subsequent Standard Value-Added

(b) Subsequent Long-Run Value-Added

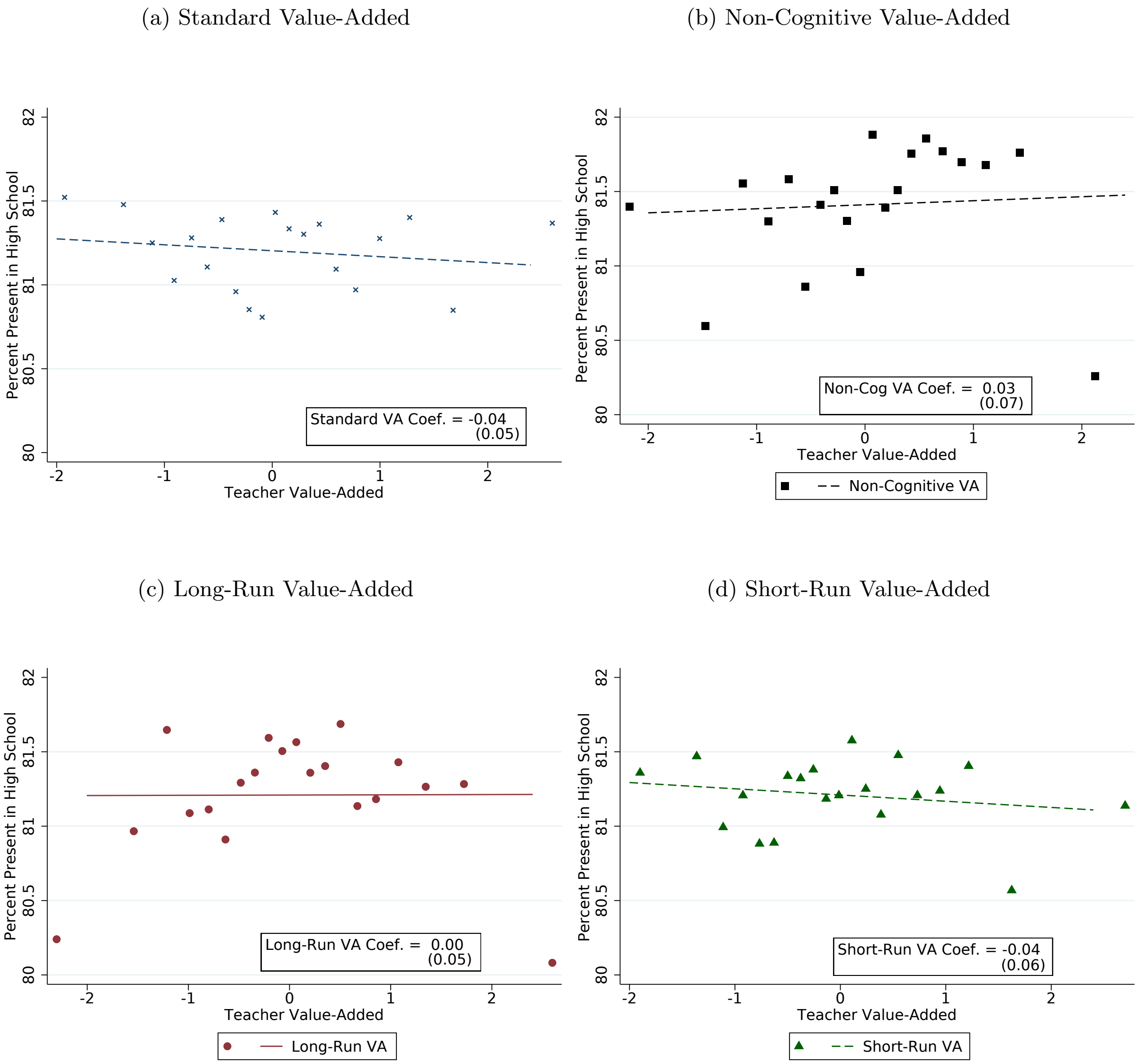
Panel B: Effect of Long-Run Value-Added on Subsequent Teacher Value-Added

(c) Subsequent Standard Value-Added

(d) Subsequent Long-Run Value-Added

Notes: This figure shows the effect of being assigned a higher standard (Panel A) and long-run (Panel B) teacher on the standard and long-run VA of your *subsequent* teacher. Each figure is constructed in three steps: (i) residualize the subsequent teacher value-added with respect to our control vector using within-teacher variation as described by equation (3.1), (ii) divide the normalized 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 subsequent teacher value-added 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 (3.4), which represent the effect of being assigned to a teacher whose VA is one standard deviation higher in a single grade on the VA of your teacher in the subsequent grade, 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: Effect of Value-Added Measures on High School Presence



(a) Standard Value-Added

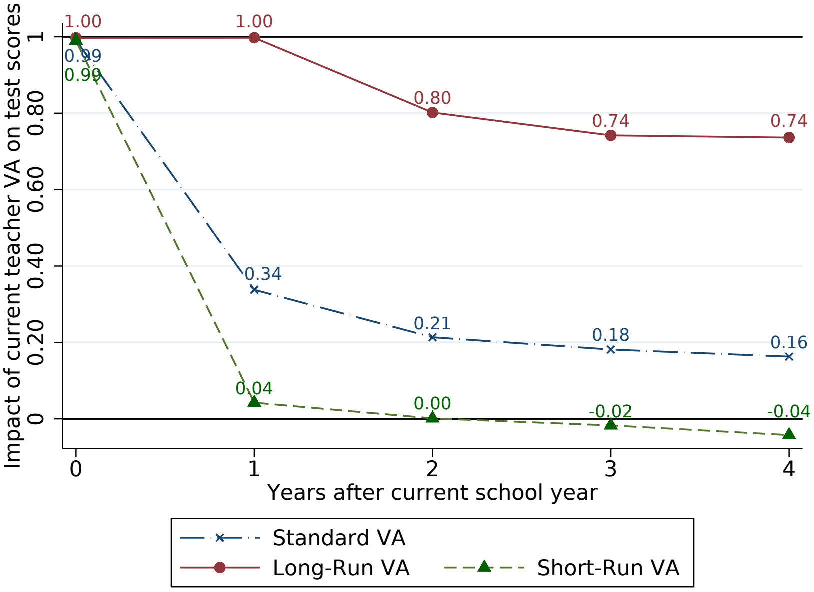
(b) Non-Cognitive Value-Added

(c) Long-Run Value-Added

(d) Short-Run Value-Added

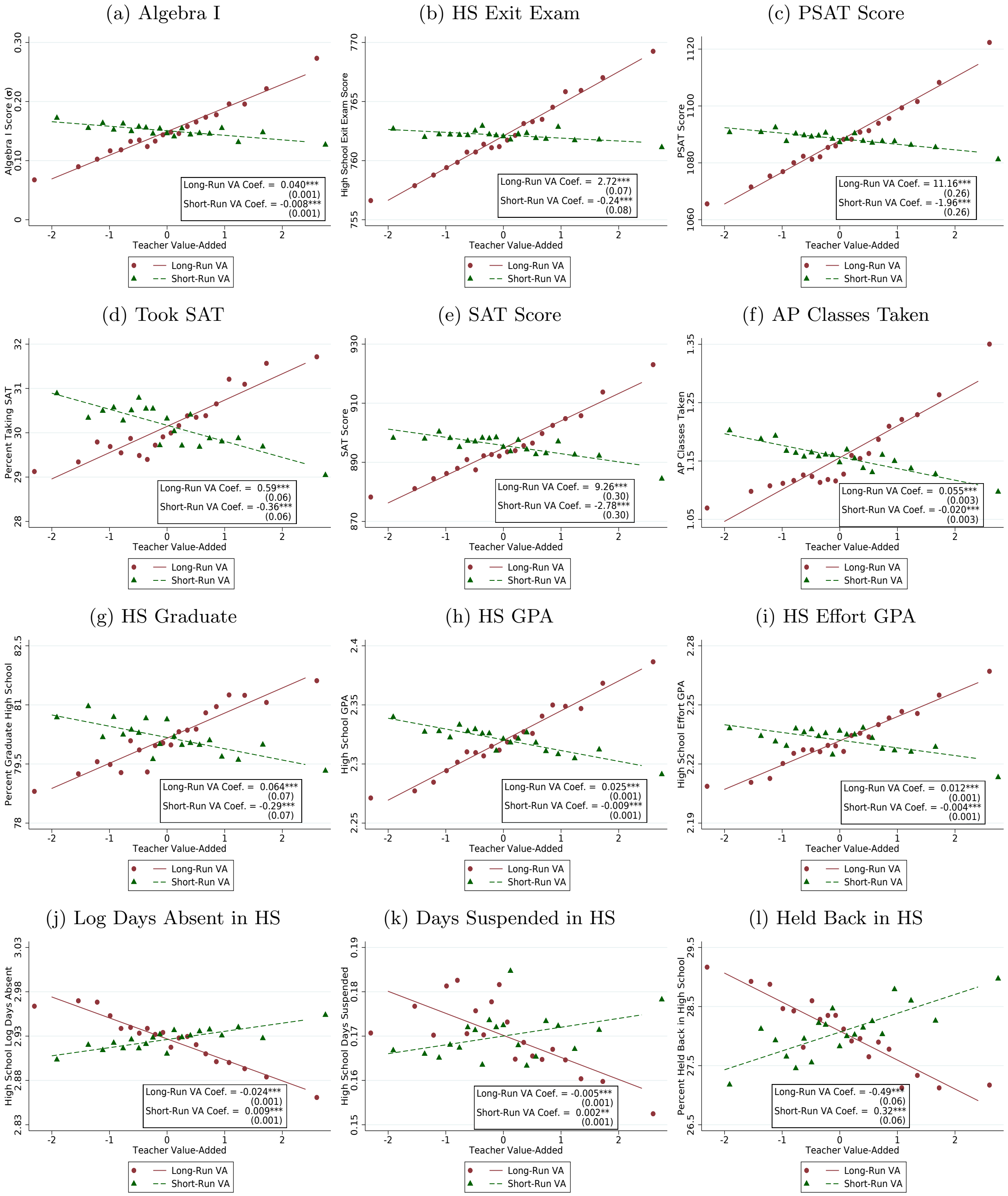
Notes: This figure shows the effect of our four measures of teacher VA on the likelihood of appearing in our high school outcomes data. We code a student as appearing in high school if they have non-missing data for the following two outcomes: algebra and high school suspensions. Algebra is chosen as students take algebra for the first time by ninth grade, while high school suspensions are chosen as a student will have a suspension record (even if never suspended) if we observe a ninth grade report card. These outcomes therefore occur prior to tenth grade, before students can drop out of high school. 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 (3.1), (ii) divide the normalized 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 (3.4), 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.4: Effects of Standard, Long-Run, and Short-Run Value-Added on Future Test Scores



Notes: This figure shows the effect of teacher standard, long-, and short-run VA on contemporaneous and future test scores. For future test scores the figure is constructed by regressing residualized end-of-grade math and English test scores in t years after being with teacher j on teacher j's VA measure in that subject as described by equation (3.3). The same method is used for contemporaneous test scores, but we include both long- and short-run VA in the regression since they both influence contemporaneous test scores according to equation (4.2). Point estimates from our regressions are reported for each point. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (3.2). The coefficients and standard errors of the point estimates underlying the figure are reported in Table A.3.

Figure A.5: Effect of Long- and Short-Run Value-Added on High School Outcomes



(a) Algebra I

(b) HS Exit Exam

(c) PSAT Score

(d) Took SAT

(e) SAT Score

(f) AP Classes Taken

(g) HS Graduate

(h) HS GPA

(i) HS Effort GPA

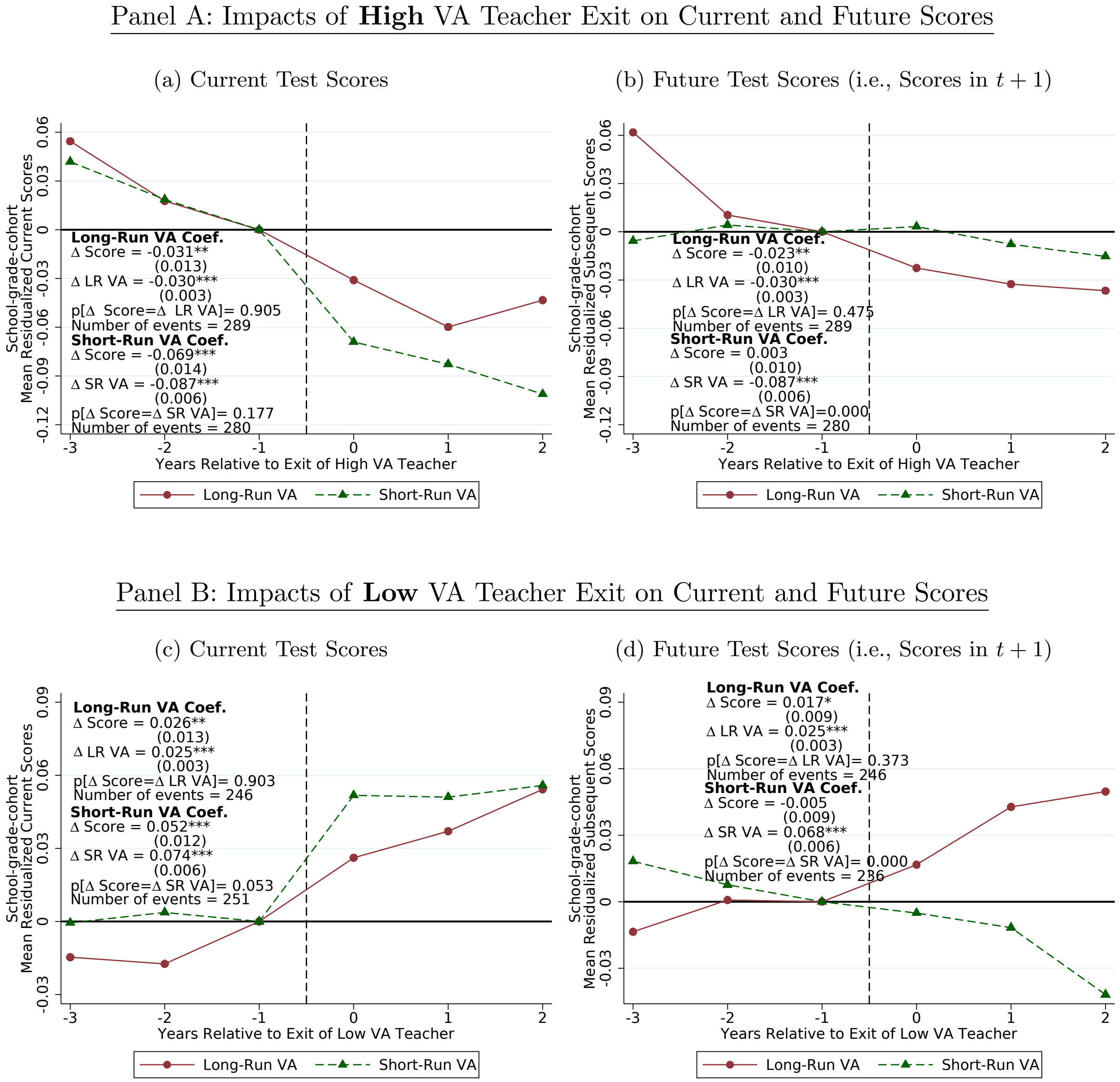
(j) Log Days Absent in HS

(k) Days Suspended in HS

(l) Held Back in HS

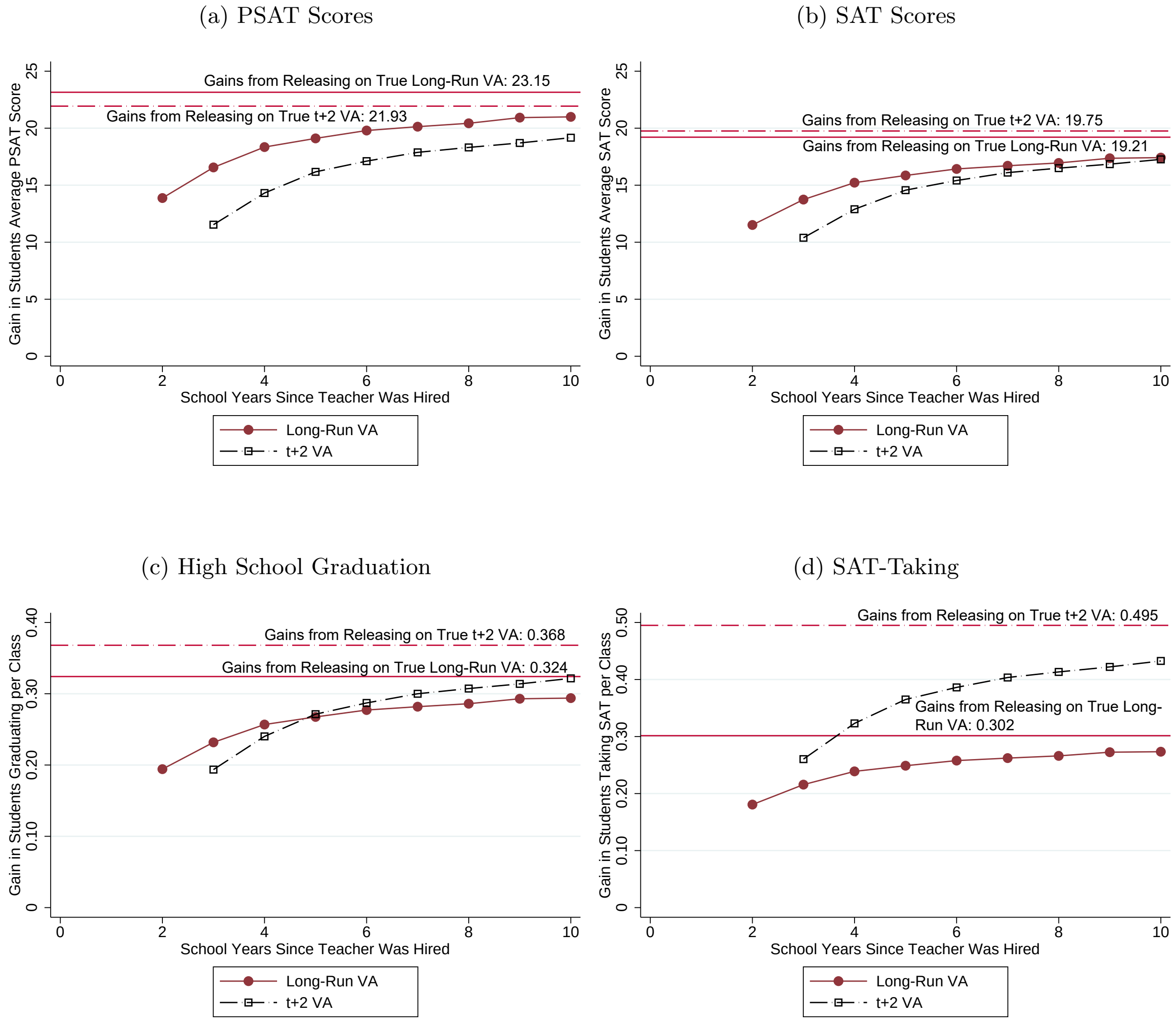
Notes: This figure shows the effect of teacher VA on high school outcomes for the long- and short-run value-added indices. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the long- or short-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school 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^{κ} from equation (3.4), 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 high school outcomes, along with its standard errors in brackets below. Effects for long-run VA are the same as those reported in Figures 2 and 3. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure A.6: Impacts of High and Low Long- and Short-Run VA Teacher Exit on Test Scores



Notes: These figures plot event studies of (residualized) test scores 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 consists of two series whereby VA is measured as long-run (solid series) and short-run VA (dashed series). The y-axis of the left-hand side figures is 'current scores,' the residualized mean contemporaneous test scores of the school-grade-cohort, while right-hand side figures y-axis is 'subsequent scores,' the residualized mean test scores of the school-grade-cohort in the *following* year. Test scores are residualized using the control vector defined in Section 2.2 and include subsequent teacher fixed effects. To construct each panel we: (i) identify the set of teachers who exit a school-grade-subject cell and do not return to that school for at least three years (i.e., we use only within-school switcher variation) and define event time as the school year relative to the year of exit, (ii) estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$ from their VA calculation, (iii) classify high- and low-VA teachers as those with VA estimates in the top or bottom 5% of the distribution among teachers who exited schools in that year, (iv) 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. The test score changes at event year -1 are normalized to zero and we include grade-by-year fixed effects to eliminate secular time trends. 'Δ Score' reports the change in current or future test scores from 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 from the period after the teacher exited (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.7: Impacts of Releasing Low Value-Added Teachers: Long-Run VA Relative to 't+2 VA'


(a) PSAT Scores

(b) SAT Scores

(c) High School Graduation

(d) SAT-Taking

Notes: This figure calculates the impact of replacing teachers with long-run VA or 't+2 VA' in the bottom 5% of the VA distribution with teachers of average quality on average PSAT and SAT scores and the number of high school graduates and SAT takers from a single classroom of average size (24.5 students). The horizontal lines show the hypothetical gain in the current school year from releasing the bottom 5% teachers according to their true long-run (solid) and 't+2' (long-dash dot) VA, with the value of this line reported above it. 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. All values in these figures use VA that equal weights teachers' mathematics and English contributions, which are calculated separately. Values for long-run VA are identical to those in Figure 8.

Table A.1: Coverage of Long-Term 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	73% (472,494 of 651,285)
High School Exit Exam	2002-03 to 2014-15	Grade 10	Entering 3^{rd} grade in 2007-08 or before	58% (448,980 of 769,545)
PSAT	2008-09 to 2016-17	Grade 10	Entering 3^{rd} grade in 2009-10 or before	53% (527,423 of 996,784)
SAT	2006-07 to 2016-17	Grades 11-12	Entering 3^{rd} grade in 2007-08 or before	30% (230,974 of 769,545)
Graduation[b]	2011-12 to 2015-16	Grade 12	Entering 3^{rd} grade in 2006-07 or before	55% (332,643 of 607,826)
AP Classes	2002-03 to 2016-17	Grade 12	Entering 3^{rd} grade in 2007-08 or before	74% (570,233 of 769,545)
High School Outcomes:[c]				
Absences	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	72% (639,035 of 885,235)
GPA	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	69% (612,478 of 885,235)
‘Effort’ GPA	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	56% (490,983 of 885,235)
Days Suspended	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	74% (653,005 of 885,235)
Held Back	2003-04 to 2016-17	N/A	Entering 3^{rd} grade in 2008-09 or before	68% (601,338 of 885,235)

[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 one if a student graduates or receives a special education certificate and zero if the student is coded as a dropout in the graduation data files. 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. To be in these data we must observe a high school transcript (grades 9-12) for the student for at least one term. Days suspended are total days suspended during a student’s high school career, with these data being universal (conditional on observing at least one transcript) since we assume you were not suspended if you have a transcript but no suspension record. Absences similarly record the total number of absences over a student’s high school career, although we note that there are some students who have a transcript but are missing absence data. GPA is the average GPA over a student’s high school career, although we cannot construct GPA for some student-term observations due letter grades not being recorded in the transcript (e.g., grades coded as pass/fail); ‘effort’ GPA is constructed analogously. Held back is constructed as an indicator variable when a student’s high school grade is the same in two consecutive years; this requires us to observe valid transcripts for two consecutive years and so data coverage is not universal for this outcome.

Table A.2: 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	Effort GPA	Suspended	Held Back
Math VA	Long-Run	1								
	Short-Run	0.02	1							
English VA	Long-Run	0.67	0.06	1						
	Short-Run	0.03	0.72	-0.07	1					
	Absences	0.19	-0.04	0.16	-0.05	1				
Non-Cog VA	GPA	0.36	0.07	0.40	0.06	0.14	1			
	Effort GPA	0.32	0.03	0.33	0.03	0.13	0.79	1		
	Suspension	0.08	-0.04	0.11	-0.03	0.12	0.13	0.13	1	
	Held Back	0.07	0.01	0.09	0.02	0.07	0.09	0.05	0.05	1

Notes: This table reports the correlations between our various value-added measure. Each VA measure is constructed as described in Section 2.2. Table 3 reports the correlation matrix between our four VA measures (which are built using the components listed in this table).

Table A.3: 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)
Standard VA	0.992	0.338	0.213	0.181	0.163
(s.e.)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)
[95% CI]	[0.984,1.001]	[0.328,0.347]	[0.204,0.223]	[0.171,0.192]	[0.151,0.175]
Observations	2,238,956	2,009,797	1,635,447	1,254,326	909,098
Long-Run VA	0.997	0.998	0.802	0.742	0.736
(s.e.)	(0.012)	(0.010)	(0.012)	(0.012)	(0.015)
[95% CI]	[0.974,1.020]	[0.978,1.017]	[0.779,0.824]	[0.717,0.767]	[0.706,0.766]
Short-Run VA	0.990	0.042	0.001	-0.017	-0.042
(s.e.)	(0.005)	(0.004)	(0.005)	(0.006)	(0.007)
[95% CI]	[0.979,1.001]	[0.034,0.050]	[-0.009,0.011]	[-0.028,-0.006]	[-0.056,-0.029]
Observations (Long- and Short-Run VA)	1,998,367	1,996,959	1,602,174	1,223,900	881,256
Non-Cognitive VA	0.030	0.030	0.031	0.028	0.030
(s.e.)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
[95% CI]	[0.028,0.033]	[0.028,0.032]	[0.029,0.032]	[0.026,0.030]	[0.027,0.033]
Observations	972,531	943,869	748,574	488,955	263,202

Notes: This table reports the 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 four test score VA measures: standard VA, long-run VA, short-run VA, and non-cognitive VA. The number of observations is the same for long- and short-run VA. The number of observations are roughly halved for non-cognitive VA as it is not subject-specific and so point estimates come from a regression of the non-cognitive VA index on averaged mathematics and English scores. The point estimates for standard, long-run, and non-cognitive VA are the same as those plotted in Figure 1. Similarly, the point estimates for standard, long-run, and short-run VA are the same as those shown in Figure A.4. Standard errors are two-way clustered by student and classroom. 95% confidence intervals are also reported.

Table A.4: Autocorrelation and Variance Estimates of Standard, Long-, and Short-Run VA

Sample:	Mathematics			English		
Value-Added Measure:	Standard	Long-Run	Short-Run	Standard	Long-Run	Short-Run
	(1)	(2)	(3)	(4)	(5)	(6)
Autocorrelation Vector						
Lag 1	0.66	0.31	0.57	0.53	0.27	0.46
Lag 2	0.61	0.24	0.54	0.49	0.21	0.42
Lag 3	0.57	0.22	0.50	0.44	0.20	0.39
Lag 4	0.53	0.19	0.48	0.41	0.19	0.35
Lag 5	0.51	0.19	0.46	0.38	0.19	0.32
Lag ≥ 6	0.48	0.19	0.44	0.35	0.20	0.31
Within-year variance components						
Total SD	0.597	0.580	0.617	0.522	0.531	0.546
Individual-level SD	0.507	0.558	0.549	0.475	0.517	0.513
Class + teacher level SD	0.315	0.158	0.280	0.217	0.123	0.188
Estimates of teacher SD						
Lower bound	0.270	0.111	0.229	0.175	0.086	0.150
Quadratic estimate	0.282	0.121	0.238	0.185	0.092	0.156
Observations (Student-Year)	1,187,231	1,061,125	1,061,125	1,183,484	1,054,042	1,054,042

Notes: This table gives the autocorrelation estimates across years for the same teacher used to compute standard, long, and short-run VA for both mathematics and English. 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. The estimates for standard and long-run VA in this table are identical to those in Table 2.

Table A.5: Correlation of Teacher Value-Added Measures Including Short-Run Value-Added

VA Measure	Standard VA	Long-Run VA	Short-Run VA	Non-Cognitive VA
Standard VA	1			
Long-Run VA	0.507	1		
Short-Run VA	0.858	0.014	1	
Non-Cognitive VA	0.205	0.394	0.011	1

Notes: This table reports the correlations between our various value-added measures. Each VA measure is constructed as described in Section 2.2. In particular, our test score VA measures (standard and long-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. 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. The correlations between standard, long-run, and non-cognitive VA in this table are identical to those in Table 3. Table A.2 reports the full correlation matrix between all the components that make up the various VA measures.

Table A.6: Policy Gains of Benchmark Policy Targeting Based on True Standard 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	0.149	762.1	1088.2	30.4	895.2	1.16
Panel A. Standard Value-Added						
Benefit ($\hat{\rho}^{Standard}$)	0.015	1.73	4.32	0.05	2.89	0.012
Average Change in VA of Released Teachers ($\Delta m_{\sigma}^{Standard}$)	1.79	1.79	1.79	1.79	1.79	1.79
Gain of Releasing Bottom 5% ($G^{Standard}$)	**0.027**	**1.27**	**7.69**	**0.08**	**5.13**	**0.022**
Panel B. Long-Run Value-Added						
Benefit ($\hat{\rho}^{Long\text{-}Run}$)	0.040	2.25	11.16	0.59	9.26	0.055
Average Change in VA of Released Teachers ($\Delta m_{\sigma}^{Long\text{-}Run}$)	2.08	2.08	2.08	2.08	2.08	2.08
Gain of Releasing Bottom 5% ($G^{Long\text{-}Run}$)	**0.083**	**5.64**	**23.15**	**1.23**	**19.21**	**0.113**

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	80.2	2.32	2.23	2.93	0.17	28.1
Panel A. Standard Value-Added						
Benefit ($\hat{\rho}^{Standard}$)	0.12	0.006	0.003	-0.005	-0.001	-0.01
Average Change in VA of Released Teachers ($\Delta m_{\sigma}^{Standard}$)	1.79	1.79	1.79	1.79	1.79	1.79
Gain of Releasing Bottom 5% ($G^{Standard}$)	**0.22**	**0.010**	**0.005**	**-0.008**	**-0.002**	**-0.02**
Panel B. Long-Run Value-Added						
Benefit ($\hat{\rho}^{Long\text{-}Run}$)	0.64	0.025	0.012	-0.024	-0.005	-0.49
Average Change in VA of Released Teachers ($\Delta m_{\sigma}^{Long\text{-}Run}$)	2.08	2.08	2.08	2.08	2.08	2.08
Gain of Releasing Bottom 5% ($G^{Long\text{-}Run}$)	**1.32**	**0.052**	**0.026**	**-0.049**	**-0.010**	**-1.02**

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 standard VA (Panel A) and long-run VA (Panel B). The estimated benefits of being assigned to a teacher whose VA is one standard deviation higher ($\hat{\rho}^k$) are the same as those reported in Figure 2. The expected value of VA conditional on being a teacher in the bottom five percent is calculated using nonparametric MLE (see Gilraine et al. (2020) for details). The policy gains are bolded and are calculated using equation (6.1), which multiplies the impact of being assigned to a teacher whose VA is one standard deviation higher ($\hat{\rho}^k$) by the average improvement in VA caused by the policy.

Table A.7: Policy Gains of Benchmark Policy Targeting Based on All VA Measures Using Three Years of Data per Teacher

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	0.149	762.1	1088.2	30.4	895.2	1.16
Panel A. Standard Value-Added						
Gain of Releasing Bottom 5% ($G^{Standard}$)	0.025	2.03	6.92	0.07	4.62	0.020
Panel B. Long-Run Value-Added						
Gain of Releasing Bottom 5% ($G^{Long\text{-}Run}$)	0.066	4.47	18.34	0.98	15.22	0.090
Panel C. 't+2' Value-Added						
Gain of Releasing Bottom 5% (G^{t+2})	0.053	3.77	14.32	1.32	12.89	0.078

Behavioral-Based Outcome:	Graduated HS (%) (7)	HS GPA (8)	HS Effort GPA (9)	Log Days Absent (10)	Days Suspended (11)	Held Back in HS (%) (12)
Sample Mean	80.2	2.32	2.23	2.93	0.17	28.1
Panel A. Standard Value-Added						
Gain of Releasing Bottom 5% ($G^{Standard}$)	0.20	0.009	0.005	-0.008	-0.002	-0.01
Panel B. Long-Run Value-Added						
Gain of Releasing Bottom 5% ($G^{Long\text{-}Run}$)	1.05	0.041	0.020	-0.039	-0.008	-0.81
Panel C. 't+2' Value-Added						
Gain of Releasing Bottom 5% (G^{t+2})	0.98	0.037	0.019	-0.035	-0.008	-1.01

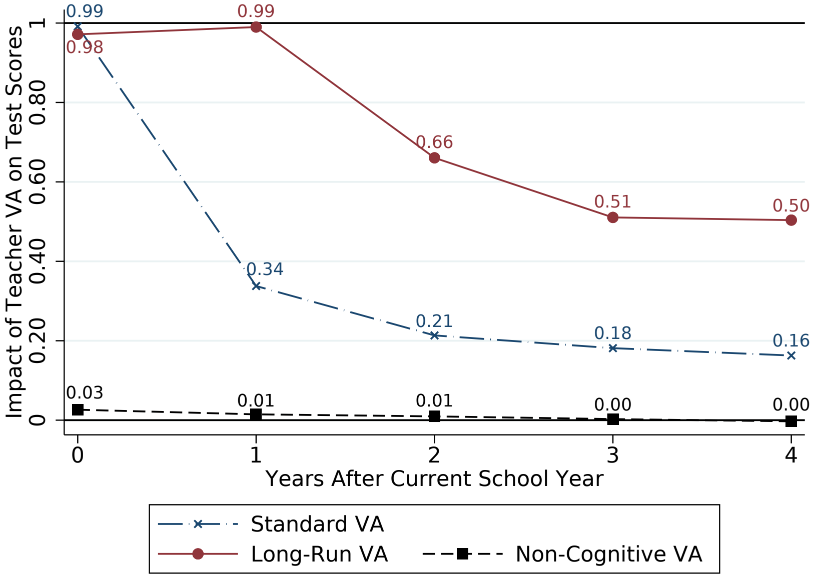
Notes: This tables calculates the policy gains for twelve high school outcomes under policies that release teachers in the bottom five percent of the true VA distribution and replace them with a mean quality teacher according to standard VA (Panel A), long-run VA (Panel B), and '$t+2$ VA' (Panel C). The estimated benefits of being assigned to a teacher whose VA is one standard deviation higher ($\hat{\rho}^k$) for standard and long-run VA are the same as those reported in Figure 2. The expected value of VA conditional on being a teacher in the bottom five percent is calculated using nonparametric MLE (see Gilraine et al. (2020) for details). The policy gains are then calculated using equation (6.1), which multiplies the impact of being assigned to a teacher whose VA is one standard deviation higher ($\hat{\rho}^k$) by the average improvement in VA caused by the policy.

B Results without Fixed Effects for Subsequent Teacher

This appendix section effectively replicates all of our results when subsequent teacher fixed effects are not included in either the estimation of teacher value-added or the residualization of high school outcomes.

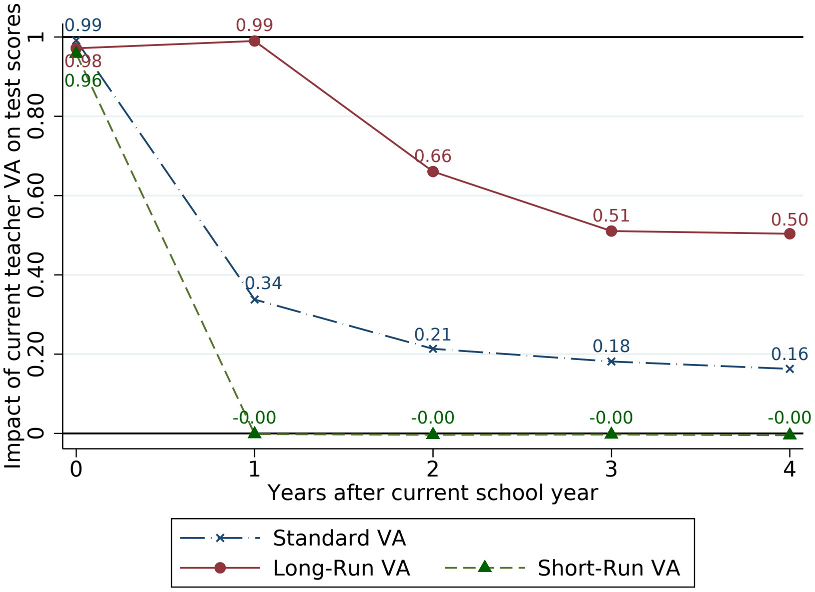
[XX: say more]

Figure B.1: Effects of Standard, Long-Run, and Non-Cognitive Value-Added on Future Test Scores



Notes: This figure shows the effect of teacher standard, long-run, and non-cognitive VA on future test scores. The figure is constructed by regressing residualized end-of-grade math and English test scores in t years after being with teacher j on teacher j's VA measure in that subject as described by equation (3.3). (Since non-cognitive VA is not subject-specific, it is just regressed on the mean of residualized end-of-grade math and English test scores as described by equation (3.4).) When regressing long-run VA on contemporaneous test scores we also control for short-run VA. Point estimates from our regressions are reported for each point. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (3.2). The coefficients and standard errors of the point estimates underlying the figure are reported in Table A.3.

Figure B.2: Effects of Standard, Long-Run, and Short-Run Value-Added on Future Test Scores



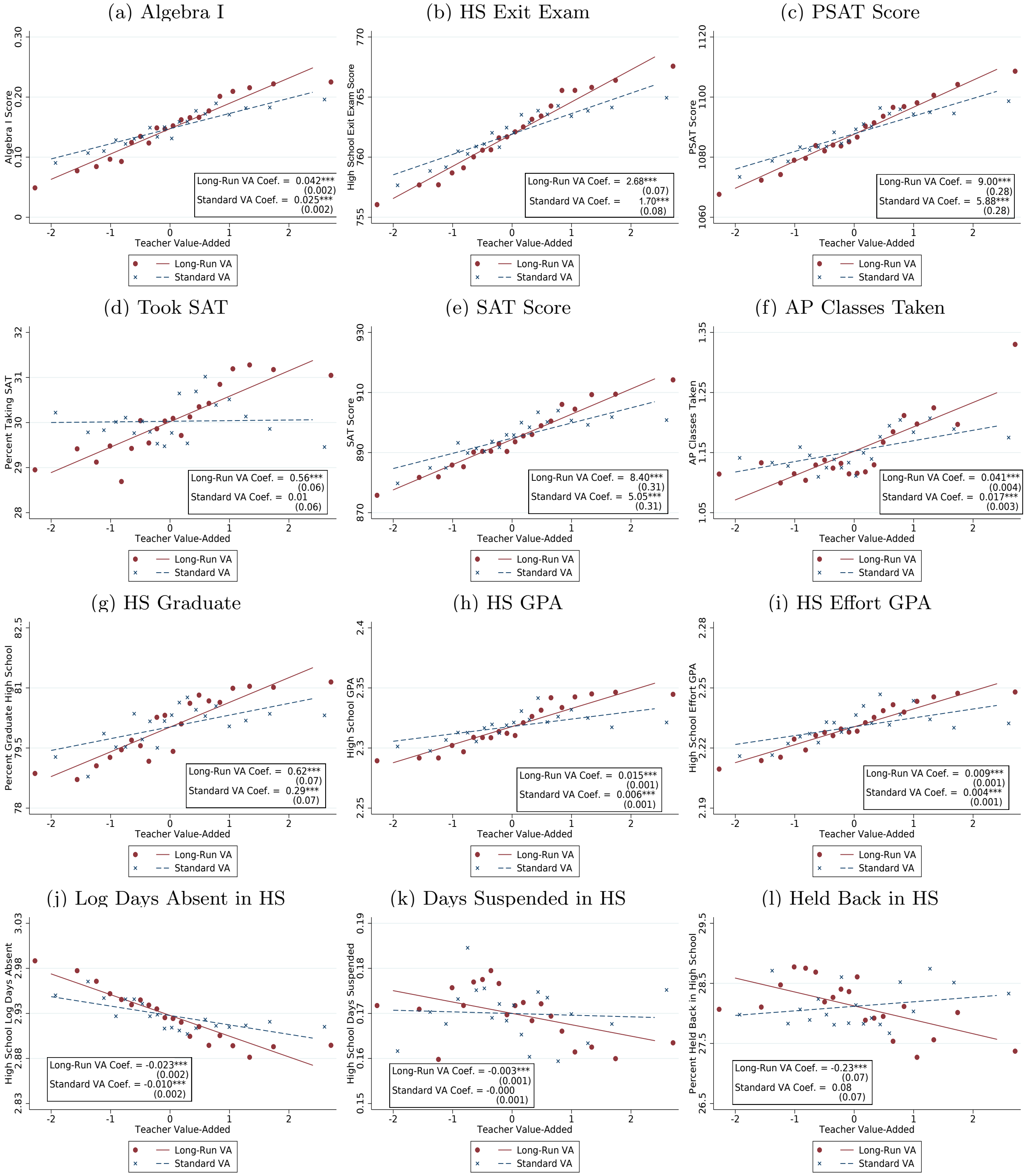
Notes: This figure shows the effect of teacher standard, long-, and short-run VA on contemporaneous and future test scores. For future test scores the figure is constructed by regressing residualized end-of-grade math and English test scores in t years after being with teacher j on teacher j's VA measure in that subject as described by equation (3.3). The same method is used for contemporaneous test scores, but we include both long- and short-run VA in the regression since they both influence contemporaneous test scores according to equation (4.2). Point estimates from our regressions are reported for each point. Test scores are residualized using our baseline control vector using within-teacher variation to identify the coefficients as described in equation (3.2). The coefficients and standard errors of the point estimates underlying the figure are reported in Table A.3.

Table B.1: Correlation of Teacher Value-Added Measures Including Short-Run Value-Added (No Subsequent Teacher Fixed Effects)

VA Measure	Standard VA	Long-Run VA	Short-Run VA	Non-Cognitive VA
Standard VA	1			
Long-Run VA	0.531	1		
Short-Run VA	0.741	-0.163	1	
Non-Cognitive VA	0.162	0.167	0.058	1

Notes: This table reports the correlations between our various value-added measures. Each VA measure is constructed as described in Section 2.2. In particular, our test score VA measures (standard and long-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. 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. The correlations between standard, long-run, and non-cognitive VA in this table are identical to those in Table 3. Table A.2 reports the full correlation matrix between all the components that make up the various VA measures.

Figure B.3: Effect of Standard and Long-Run Value-Added on High School Outcomes (No Subsequent Teacher Fixed Effects)


(a) Algebra I

(b) HS Exit Exam

(c) PSAT Score

(d) Took SAT

(e) SAT Score

(f) AP Classes Taken

(g) HS Graduate

(h) HS GPA

(i) HS Effort GPA

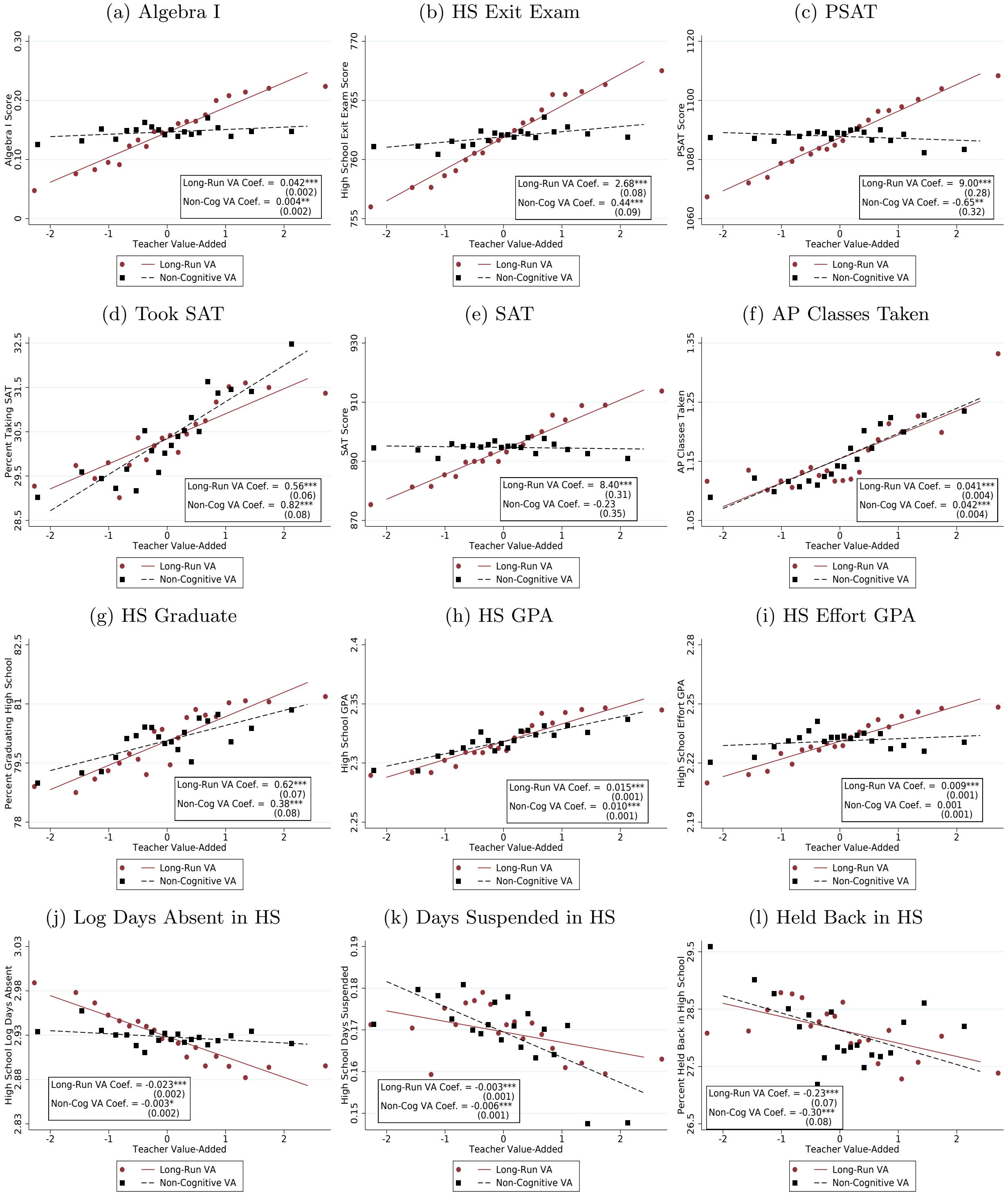
(j) Log Days Absent in HS

(k) Days Suspended in HS

(l) Held Back in HS

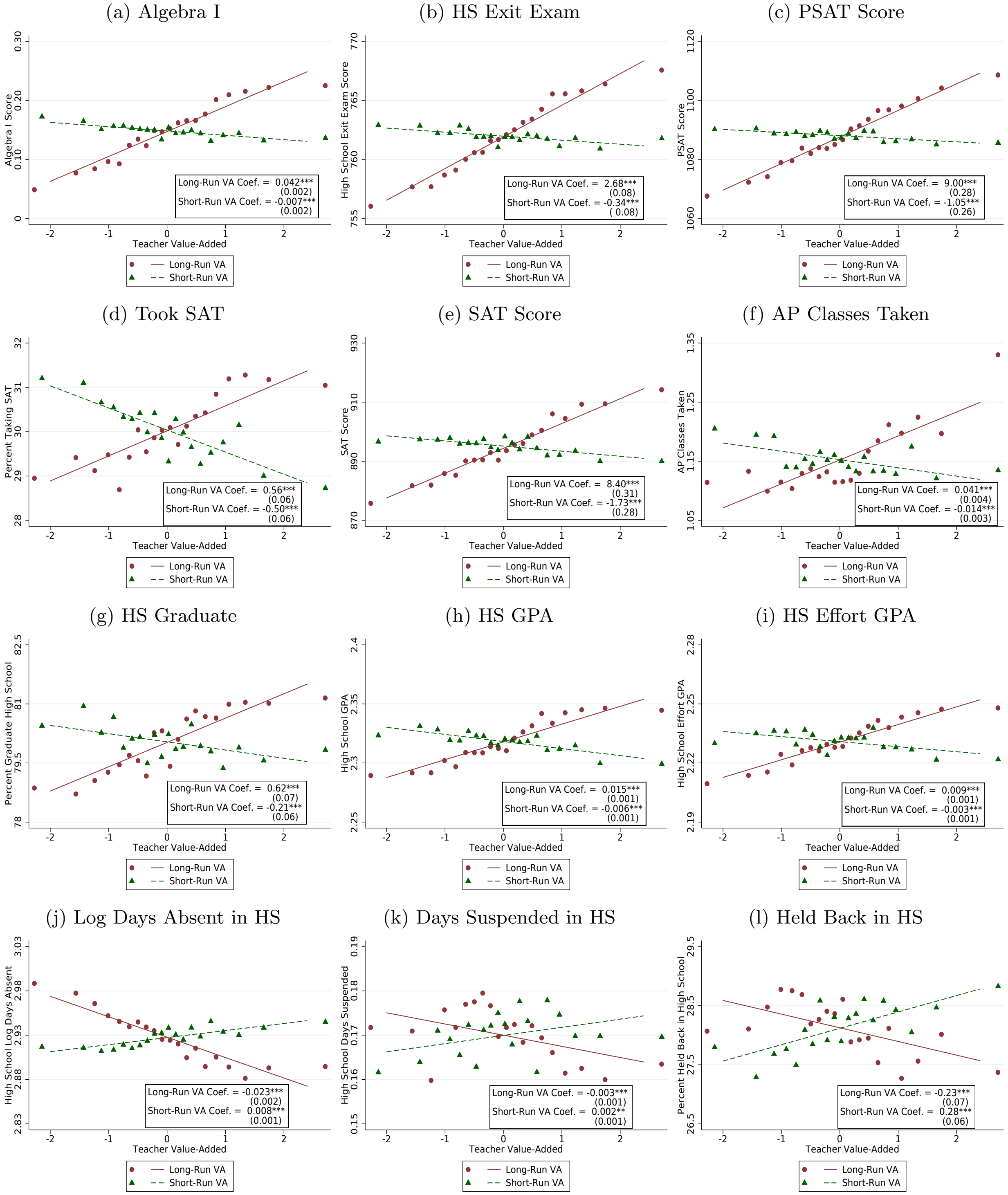
Notes: This figure replicates Figure 2 when subsequent teacher fixed effects are not included in either the estimation of teacher value-added or the residualization of high school outcomes. Note that the lack of subsequent teacher fixed effects in the outcome residualization makes the standard VA coefficients differ from those in Figure 2. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector *excluding subsequent teacher fixed effects* using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the standard or long-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school outcome in the estimation sample to facilitate interpretation of the scale. A line of best fit is then superimposed with point estimates of being assigned to a teacher whose standard or long-run VA is one standard deviation higher in a single grade also reported. Standard errors clustered at the student and classroom level are in the brackets below. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure B.4: Effect of Non-Cognitive and Long-Run Value-Added on High School Outcomes (No Subsequent Teacher Fixed Effects)



Notes: This figure shows the effect of teacher VA on high school outcomes for the non-cognitive and long-run value-added indices. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the standard or long-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school 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^{κ} from equation (3.4), 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 high school outcomes, along with its standard errors in brackets below. Effects for long-run VA are the same as those reported in Figures 2 and A.5. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure B.5: Effect of Long- and Short-Run Value-Added on High School Outcomes (No Subsequent Teacher Fixed Effects)


(a) Algebra I

(b) HS Exit Exam

(c) PSAT Score

(d) Took SAT

(e) SAT Score

(f) AP Classes Taken

(g) HS Graduate

(h) HS GPA

(i) HS Effort GPA

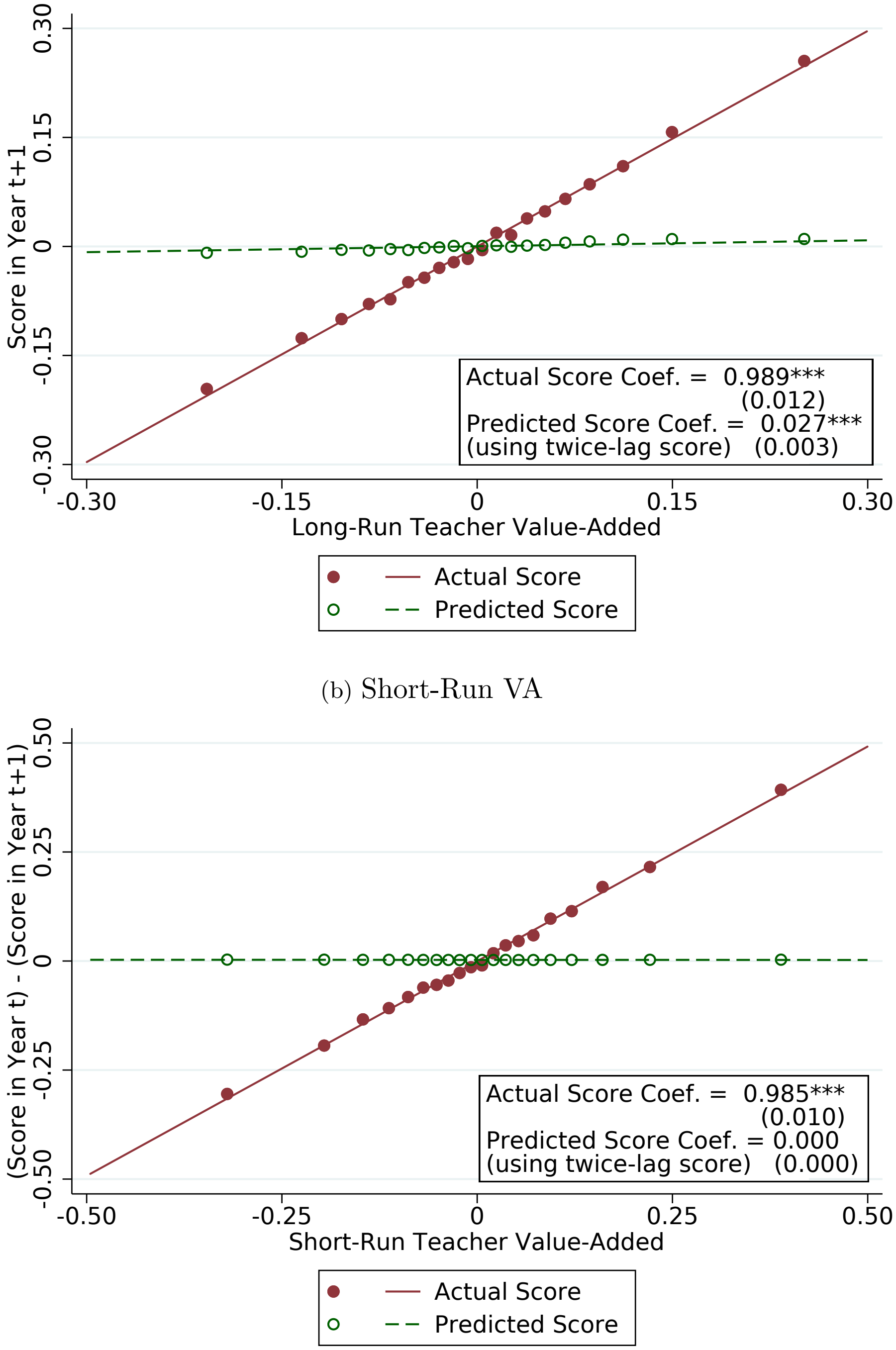
(j) Log Days Absent in HS

(k) Days Suspended in HS

(l) Held Back in HS

Notes: This figure shows the effect of teacher VA on high school outcomes for the long- and short-run value-added indices. Each figure is constructed in three steps: (i) residualize the high school outcome with respect to our control vector using within-teacher variation as described by equations (3.1) and (3.2), (ii) divide the long- or short-run VA indices, $\hat{m}^{\kappa}_{jt}$, into twenty equal-sized groups (vingtiles) and plot the mean of the high school outcome residuals in each bin against the mean of $\hat{m}^{\kappa}_{jt}$ in each bin, (iii) add back the mean of the high school 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^{κ} from equation (3.4), 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 high school outcomes, along with its standard errors in brackets below. Effects for long-run VA are the same as those reported in Figures 2 and 3. Standard errors are clustered at the student and classroom level. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure B.6: Effects of Long-Run and Short-Run VA on Actual and Predicted Scores (No Subsequent Teacher Fixed Effects)

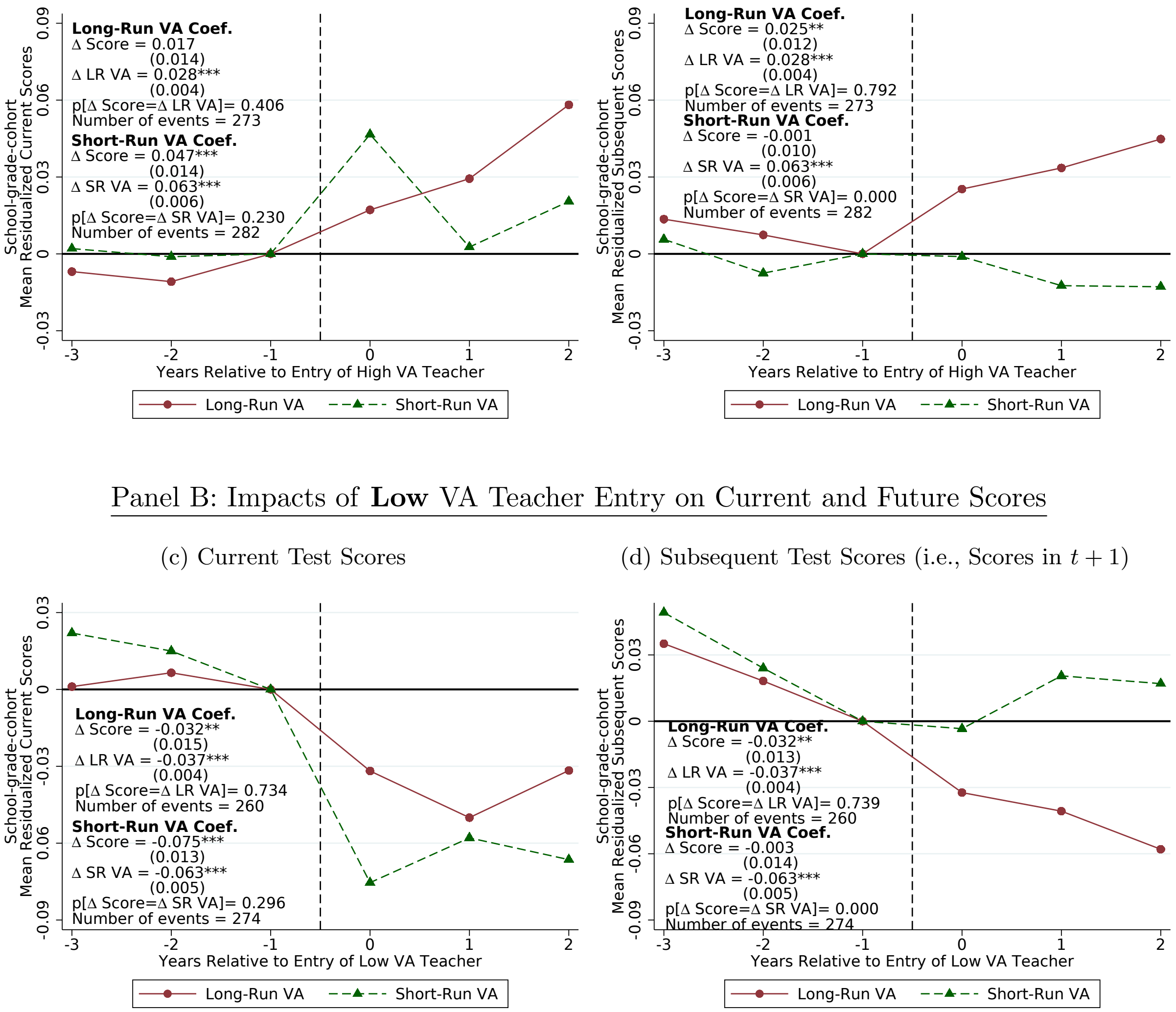


(a) Long-Run VA

(b) Short-Run VA

Notes: These figures replicate Figure 4 when subsequent teacher fixed effects are not included in the estimation of teacher value-added or the residualization of test scores. Note that the standard VA figure is excluded here as it would be identical to the one in Figure 4 since subsequent teacher fixed effects are not used in the estimation of standard VA. As in Figure 4, these figures assess whether students sort on variables that predict test score residuals but are omitted from the VA models. We predict scores based on twice-lagged test score outcomes separately by subject. 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 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 – $A_{ij,t+1}$, $A_{ijt} - A_{ij,t+1}$ for long-run and short-run VA, respectively – against the VA estimates. 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 student and classroom level reported in parentheses. ***,** and * denote significance at the 1%, 5% and 10% levels, respectively.

Figure B.7: Impacts of High and Low Long- and Short-Run VA Teacher Entry on Test Scores (No Subsequent Teacher Fixed Effects)



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

(a) Current Test Scores

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

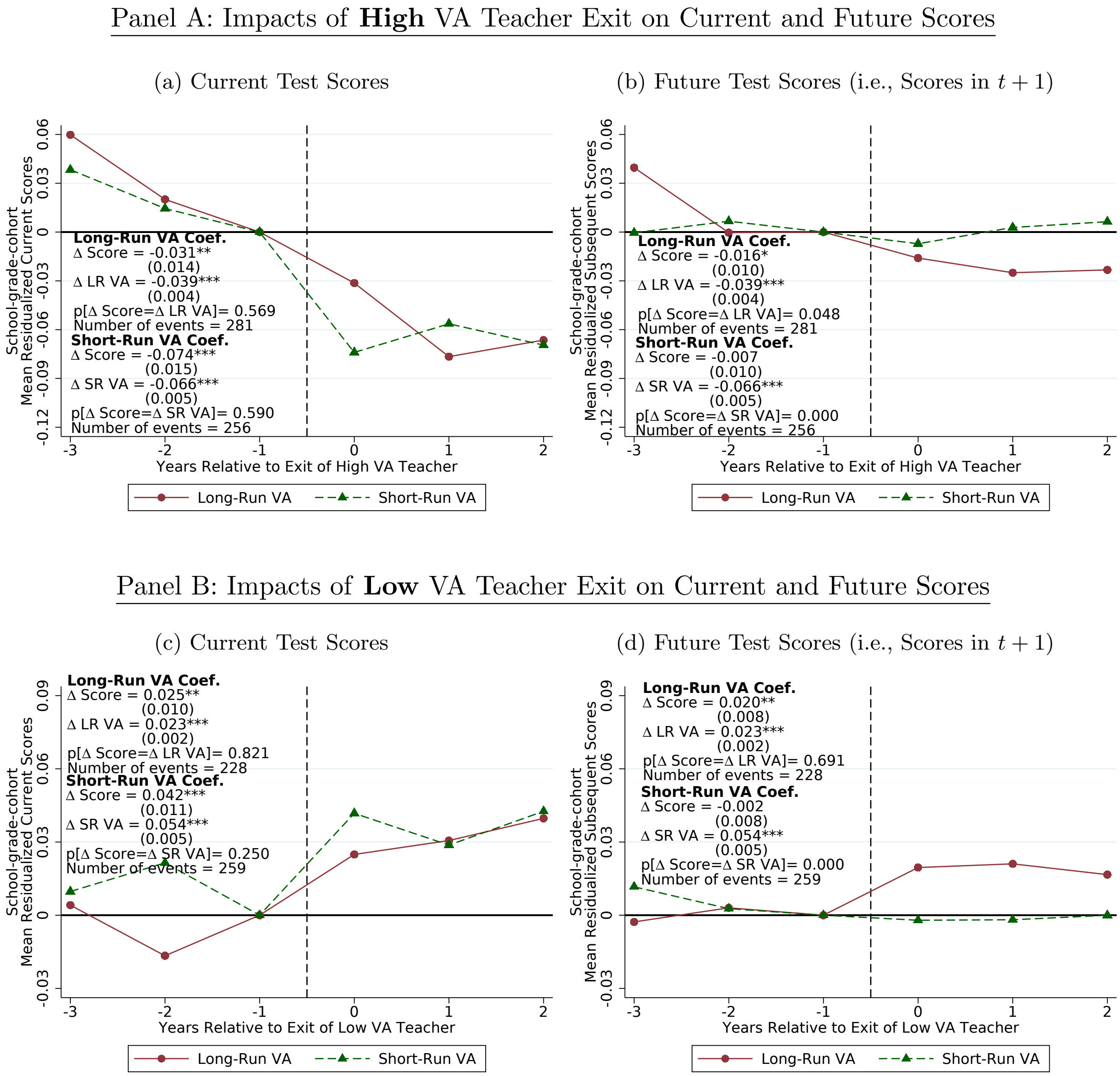
Panel B: Impacts of **Low** VA Teacher Entry on Current and Future Scores

(c) Current Test Scores

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

Notes: These figures replicate Figure 5 when subsequent teacher fixed effects are not included in the estimation of teacher value or the residualization of test scores. As in Figure 5, these figures plot event studies of (residualized) test scores by cohort as teachers enter a school-grade-subject cell at event-time 0. Panel A does so for high VA teachers, while Panel B does so for low VA teachers. Each figure consists of two series whereby VA is measured as long-run (solid series) and short-run VA (dashed series). The y-axis of the left-hand side figures is 'current scores,' the residualized mean contemporaneous test scores of the school-grade-cohort, while right-hand side figures y-axis is 'subsequent scores,' the residualized mean test scores of the school-grade-cohort in the *following* year. Test scores are residualized using the control vector defined in Section 2.2 *without* subsequent teacher fixed effects. To construct each panel we: (i) identify the set of teachers who entered a school-grade-subject cell and did not teach at that school for the preceding three years and define event time as the school year relative to the year of entry, (ii) estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$ from their VA calculation, (iii) classify high- and low-VA teachers as those with VA estimates in the top or bottom 5% of the distribution among teachers who entered schools in that year, (iv) 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. The test score changes at event year -1 are normalized to zero. 'Δ Score' reports the change in current or future test scores from 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 from 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 B.8 reports results from similar event studies that leverage teacher entry without subsequent teacher fixed effects.

Figure B.8: Impacts of High and Low Long- and Short-Run VA Teacher Exit on Test Scores (No Subsequent Teacher Fixed Effects)



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 $t+1$)

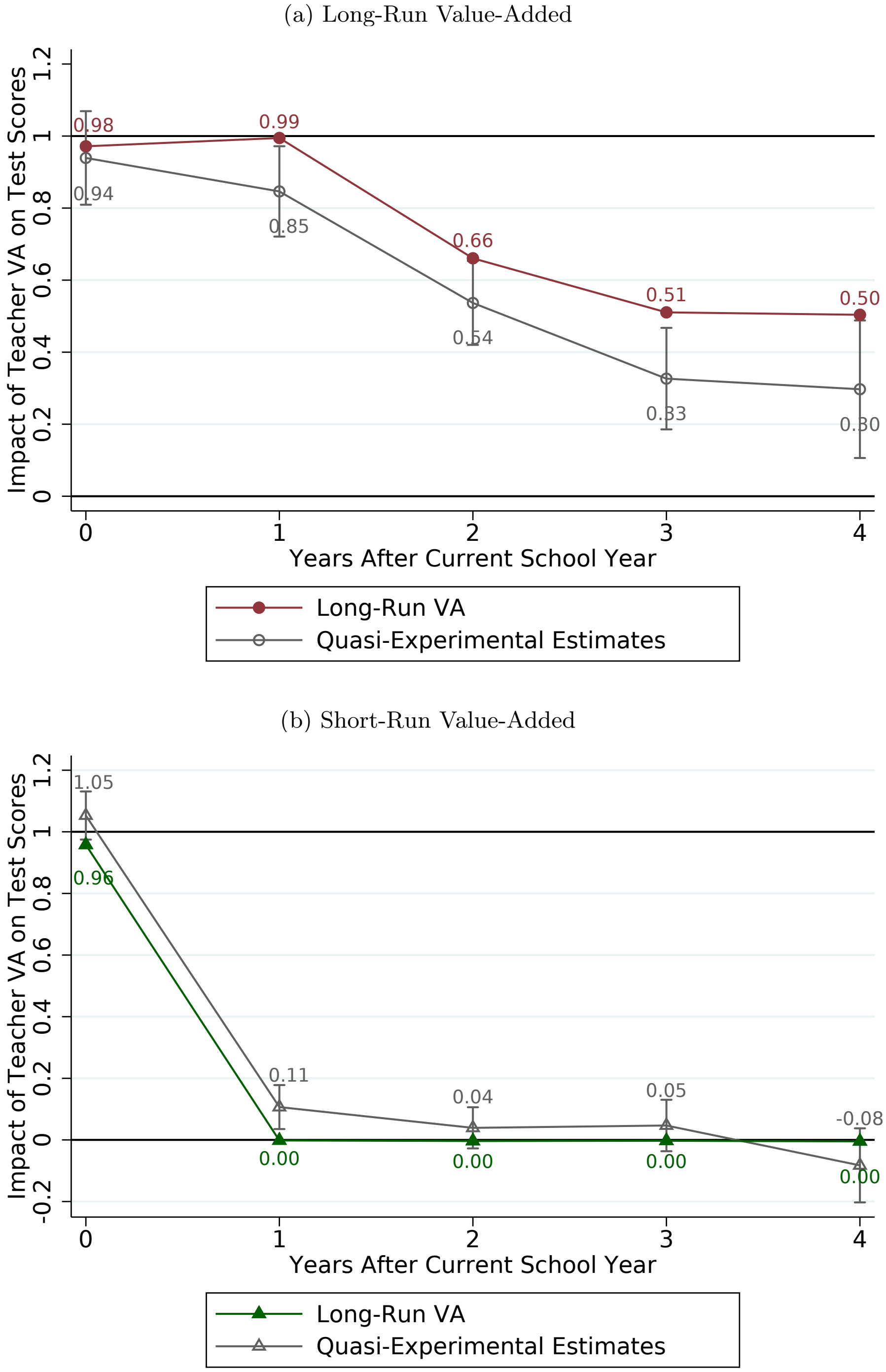
Panel B: Impacts of **Low** VA Teacher Exit on Current and Future Scores

(c) Current Test Scores

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

Notes: These figures replicate Figure A.6 when subsequent teacher fixed effects are not included in the estimation of teacher value or the residualization of test scores. As in Figure A.6, these figures plot event studies of (residualized) test scores by cohort as teachers exit a school-grade-subject cell at event-time 0. Panel A does so for high VA teachers, while Panel B does so for low VA teachers. Each figure consists of two series whereby VA is measured as long-run (solid series) and short-run VA (dashed series). The y-axis of the left-hand side figures is 'current scores,' the residualized mean contemporaneous test scores of the school-grade-cohort, while right-hand side figures y-axis is 'subsequent scores,' the residualized mean test scores of the school-grade-cohort in the *following* year. Test scores are residualized using the control vector defined in Section 2.2 *without* subsequent teacher fixed effects. To construct each panel we: (i) identify the set of teachers who exit a school-grade-subject cell and do not return to that school for at least three years and define event time as the school year relative to the year of exit, (ii) estimate each teacher's long- or short-run VA in event year $t = 0$ using data from classes taught excluding event years $t \in [-3, 2]$ from their VA calculation, (iii) classify high- and low-VA teachers as those with VA estimates in the top or bottom 5% of the distribution among teachers who exited schools in that year, (iv) 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. The test score changes at event year -1 are normalized to zero. 'Δ Score' reports the change in current or future test scores from 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 from the period after the teacher exited (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 B.7 reports results from similar event studies that leverage teacher entry without subsequent teacher fixed effects.

Figure B.9: Cross-Sectional vs. Quasi-Experimental Estimates of Long-Run and Short-Run Value-Added on Future Test Scores (No Subsequent Teacher Fixed Effects)


(a) Long-Run Value-Added

(b) Short-Run Value-Added

Notes: These figures replicate Figure 6 when subsequent teacher fixed effects are not included in the estimation of teacher value-added or the residualization of test scores. Note that the standard VA figure is excluded here as it would be identical to the one in Figure 6 since subsequent teacher fixed effects are not used in the estimation of standard VA. As in Figure 6, these figures compare the cross-sectional estimates from Figure **??** to quasi-experimental estimates that leverage teacher school-switchers. The cross-sectional estimates are identical to those reported in Figure **??**. The quasi-experimental estimates come from instrumental variable regressions that regress changes in school-grade mean residualized test scores across cohorts against changes in mean teacher VA, instrumenting for the change in mean teacher VA using the change in teacher VA coming from school-switchers as described in equation (5.1). Point estimates from our regressions are reported above or below each point. The whiskers represent 95% confidence intervals for our quasi-experimental estimates with standard errors clustered at the school-cohort level.