NBER WORKING PAPER SERIES

WHY DOES VALUE-ADDED WORK? IMPLICATIONS OF A DYNAMIC MODEL OF STUDENT ACHIEVEMENT

Douglas O. Staiger
Thomas J. Kane
Brian D. Johnson

Working Paper 34119
http://www.nber.org/papers/w34119

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2025

We thank the North Carolina Education Research Data Center (NCERDC) for providing access to the NC data and Eric Taylor and Mary Laski for sharing their code when constructing our analysis file. We also thank seminar participants in the NBER Summer Institute in the Economics of Education in 2024 and at the University of California at Los Angeles and University of California in Santa Barbara for their comments. We are grateful to the Walton Family Foundation for research support. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w34119

Why Does Value-Added Work? Implications of a Dynamic Model of Student Achievement
Douglas O. Staiger, Thomas J. Kane, and Brian D. Johnson
NBER Working Paper No. 34119
August 2025
JEL No. I21

ABSTRACT

Non-experimental value-added models have been shown to yield forecast-unbiased estimates of teacher and school effects. To investigate, we propose a dynamic state-space model of knowledge accumulation, in which test scores are imperfect measures of knowledge, and students receive temporary and persistent shocks to their stock of knowledge each period. The model identifies two primary sources of bias: transient factors in baseline scores (measurement error or transitory effects of prior teachers) and heterogeneous student growth rates. We propose diagnostic tests and corrections for each. Using eleven years of data from North Carolina, we find little evidence of heterogeneous student growth. Rather, the primary source of bias is attenuation of the baseline coefficient, which opens the door to selection on correlates of true baseline knowledge. Although conventional value-added estimates are biased for individual teachers due to attenuation, we find they are forecast unbiased when applied across a sample of teachers, due to the offsetting relationship between bias and teachers' true effectiveness. When achievement follows the state-space model and there is no heterogeneity in growth rates, the attenuation-corrected value-added model (ACVAM) should yield unbiased estimates of a wide range educational interventions, not just teachers and schools.

Douglas O. Staiger
Dartmouth College
Department of Economics
and NBER
douglas.staiger@dartmouth.edu

Thomas J. Kane
Harvard University
Harvard Graduate School of Education
and NBER
tom_kane@gse.harvard.edu

Brian D. Johnson
Harvard University
brian_johnson@gse.harvard.edu

Introduction

For more than five decades, economists have used non-experimental "value-added" models (VAM) to estimate impacts of schools, teachers and other interventions. (For early examples, see Hanushek [1971], Murnane [1975], Summers and Wolfe [1977], Hanushek [1979]). More recently, eleven independent studies have found that such value-added methods yield forecast-unbiased estimates of teachers and schools, validly predicting the results of experiments or quasi-experiments. Did authors happen to choose settings where there was no selection on unobservables? Or do these studies reveal something fundamental about the time series properties of student achievement? We present evidence for the latter.

To understand the conditions required for unbiased value-added estimation, we propose a simple state-space model of student achievement. In the model, test scores are a noisy measure of an underlying state variable (student knowledge). Teachers and other interventions have persistent effects on student knowledge, which persist at a common rate (the decay rate of knowledge), as well as transitory effects on measured achievement, which last for only a single period (for example, when a teacher "teaches to the test" rather than long-term conceptual understanding). In addition, we allow each student to have their own fixed rate of knowledge growth.

Using eleven years of longitudinal data for all students in grades 3-8 in North Carolina, we show that the state space model captures the time series properties of student achievement. The recursive Kalman filter—which summarizes a student's entire history of achievement – outperforms the single year of baseline achievement and performs nearly as well as a fully flexible model with all prior lags of test scores and teacher effects. As a summary of prior knowledge and teacher inputs, the Kalman filter fulfills one of the key requirements for unbiasedness of value-added models outlined by Todd and Wolpin (2003)—namely, that prior knowledge and the effect of any persistent contribution to knowledge decay each period at the same rate.[1] Moreover, the Kalman filter can explain a wide range of stylized facts commonly observed in value added specifications, including the significance of cubic terms in baseline test scores and of test scores from other subjects and earlier lags.

We use the model to identify and correct for four potential sources of bias in value-added models.

The first source of bias is attenuation of the coefficient on baseline achievement due to measurement error in baseline test scores, which opens the door to selection on any correlate of true baseline knowledge (including private information held by teachers within the school or unmeasured family background). To address attenuation bias, we propose an attenuation-corrected value-added model (ACVAM), using twice-lagged test scores as instruments for lagged test scores.

[1] Our model also allows for temporary impacts of teachers and other interventions which disappear after the initial period.

In our data we find considerable sorting of students to teachers related to true baseline knowledge which leads to bias in the presence of attenuation. We test whether correcting for attenuation eliminates the threat of selection using two measures not typically included in administrative data: prior teacher ratings of students' mastery of the subject and students' neighborhood income. We find that prior teacher ratings do have large and significant predictive power in conventional VAM models, but not in our attenuation-corrected value-added model (ACVAM). In fact, the predictive power of the prior teacher's rating is minimized at our preferred ACVAM estimate of the baseline knowledge persistence rate. We find similar results using students' neighborhood median household income. When the baseline coefficient is dis-attenuated, selection on such factors does not lead to bias.

The second potential source of bias in value added estimates arises from the transitory effect of a student's prior-year teacher on baseline scores. Many studies have found that a considerable portion of teacher effects "fades out" in the following year (Koedel et al. 2015). If teachers have transitory effects, their students will have lower than expected test score growth in the following year, potentially biasing value-added estimates of subsequent teachers if used for selection. To account for the transitory effect of the prior-year teacher, our ACVAM model controls for both current and prior teacher and directly estimates the variation in both persistent and transitory teacher effects. We find substantial variation across teachers in their transitory effect on test scores, similar in magnitude to teachers' persistent effects on student knowledge.[2]

A third potential source of bias in value added estimates occurs if teacher assignment is based on transitory student-level measurement error in baseline achievement (analogous to Ashenfelter's dip). If students were mechanically assigned to teachers or other interventions because their baseline measure was low (or high), we could be overstating (or understating) the impact. (Contrary to intuition, assignment based on an observed covariate leads to bias in the state space model, since it is true knowledge, not measured achievement, which determines expected end of year achievement.) Fortunately, many states do not report test scores from the prior spring until late in the summer or after school starts in the fall, making them difficult to use for this type of selection. In cases where baseline scores are used for assignment, we suggest using an alternative baseline test that is administered following assignment (or otherwise not observed at the time of assignment) and estimating ACVAM with the new baseline measure.

The final potential source of bias in value added estimates arises from unobserved heterogeneity in students' rates of growth. We find surprisingly little evidence of persistent differences in students' growth rates in North Carolina. The covariance in residual achievement growth over long lags—an estimate of the variance in persistent growth rates – is small relative

[2] Like Jacob et al. (2010), we find that the rate of fade-out varies from teacher to teacher. In fact, teachers' persistent and transitory effects on student achievement are negatively correlated, implying that those who "teach to the test" also have smaller longer term impacts on student achievement.

to the student or teacher-level variation in value added. Moreover, the coefficient on baseline achievement is largely insensitive to the inclusion of additional student characteristics which can proxy for unmeasured growth (race/ethnicity, gender, economic disadvantage, English language learners and special education status).[3] When we apply a bias correction inspired by Oster (2019) and Altonji et al. (2005), using the sensitivity of the baseline coefficient to inclusion of student characteristics, the estimated baseline coefficient is largely unchanged. Taken together, these findings imply that there is little predictable variation in unmeasured student growth available to be used for selection. Moreover, the existing variation is largely captured by student demographics and does not seem to bias the coefficient on baseline scores. While there is a large literature in psychology reporting heterogeneity in rate of student achievement growth *post hoc*, our results imply such heterogeneity is spurious and unpredictable *ex ante*.[4]

In the final part of the paper we return to the original question posed in the title: If conventional value-added estimates are biased (particularly through attenuation bias on baseline test scores), why have so many studies found them to be forecast unbiased? Conventional value-added estimates of teacher and school effects can be individually biased and yet pass tests of forecast bias across a sample of teachers or schools if the overstatement of the variance in teacher effects due to bias is offset by the covariance between bias and teacher's true effects. As noted above, the bias in conventional value-added is due to the combination of student sorting and the attenuated coefficient on baseline achievement. However, the correlation between such bias and actual teacher and school effects is small and negative; better teachers and schools are assigned slightly weaker students. The net result of overstating the signal variance in teacher/school effects and underestimating the difference between the best and worst teachers/schools is a forecast coefficient close to one. The forecast coefficient is even closer to one since most validation has typically been done within-schools and within-districts. Using our ACVAM estimates as noisy estimates of true teacher effects, we estimate that the within-school forecast coefficient for conventional value-added would be .95 in math and .85 in reading.

In sum, when achievement follows the state-space model, when there is little heterogeneity in growth rates to be used for selection and when selection is based on beliefs about students' baseline knowledge (not test scores themselves), the ACVAM estimator should yield unbiased estimates of program effects even in the absence of complete information on the factors used for treatment assignments. Although we use North Carolina data, the findings of forecast unbiasedness in other settings--inside and outside the U.S-- imply that achievement broadly follows a similar model. For those seeking to confirm whether the ACVAM estimator is likely unbiased in their own setting, we propose three tests which can be applied with panels covering 4 or more grades: first, testing that student achievement follows the state-space model (and the Kalman filter summarizes past history of achievement); second, testing that the

[3] The insensitivity of teacher value-added estimates to the inclusion of student characteristics has been widely noted in the literature, but the implications have not been recognized. See Koedel et al. (2015), Aaronson et al. (2007), Ehlert et al. (2014), and McCaffrey et al. (2004).

[4] For a widely cited text on latent growth models and their use in psychology, see Bollen and Curran (2006). For a more recent review of the literature on latent growth modelling, see Miller et al. (2024).

correlation in residual growth at two or more lags is zero (or small); and, third, testing that the baseline coefficient is insensitive to inclusion of additional student covariates. (If there is evidence of the latter, our model suggests a way to adjust the baseline coefficient for bias.) If there were additional measures of prior student knowledge not included in the value-added model—such as the prior teacher ratings we have in North Carolina or additional prior assessments—such data would enable a fourth test: testing that the additional information about baseline knowledge is not predictive of subsequent achievement after dis-attenuating the baseline coefficient.

Although much of the research on the validity of value-added models has been done with teacher and school effects, our results imply that the ACVAM estimator should yield unbiased estimates for a wide variety of educational interventions (such as tutoring programs or education software interventions). Given the widespread availability of longitudinal achievement data, our results have important implications for the field of program evaluation in K-12 education. Given the relationship between test scores and earnings, even a small improvement in the rate at which effective interventions are identified could yield large financial returns (Doty et al. 2025; Hanushek and Woessmann 2008). In future work, we intend to test the unbiasedness of the ACVAM model in replicating experimental impacts of tutoring programs, a particularly challenging setting given the strong potential for selection bias.

Literature Review

For decades, economists have been estimating the impacts of teachers or schools or other interventions non-experimentally using value-added models to adjust for baseline achievement. The earliest paper by Hanushek in 1971 used a parsimonious specification, with a single measure of baseline achievement, demographics and fixed teacher effects. However, since that initial paper, value-added researchers have included an expanding array of measures to capture baseline achievement and family background. For example, Murnane (1975) included two measures of prior achievement, from the fall and prior Spring, student demographics, neighborhood characteristics, student attendance and the mean baseline achievement of peers in the class. Shortly after, Summers and Wolfe (1977) included prior achievement, peer characteristics, student absences, school characteristics as well as a wide range of interactions between baseline achievement and peer characteristics.

Over time, the standard specification has expanded to include polynomial functions of baseline achievement in the same subject as well as in other subjects (including controls for reading when math was the outcome and vice versa.) The student characteristics available in administrative data were typically limited to the subgroups required in federal accountability law (race/ethnicity, gender, economic disadvantage, special education status and English language learning status). To account for "peer effects," researchers began adding controls for the classroom and school means for baseline achievement and each of the student covariates. (For a review of the recent literature, see Koedel et al. [2015] and Ehlert et al. [2014]. [5])

Since Hanushek (1971), the conventional specification has included one year lag of student achievement and student characteristics. However, students could presumably be assigned to treatment based on achievement in earlier years as well. For example, Rothstein (2010) and Ehlert et al. (2014) find that two and three-year lagged scores played a role, even after controlling for the single lag. Rothstein speculated that excluding multiple lags of achievement and teacher assignments could lead to bias. However, he also found that their inclusion did not lead to any meaningful differences in estimated teacher effects in North Carolina.

With the introduction of large student-level panel data sets in the mid-2000's, it became feasible for value-added researchers to include student fixed effects to control for selection on unmeasured student characteristics (e.g., McCaffrey et al. [2009]). Given the threat of selection bias, it made sense to estimate even more saturated models, including student fixed effects. However, in one of the first papers acknowledging the time series properties of achievement, Rothstein (2010) noted that inclusion of student fixed effects depended on a restrictive, static view of selection and likely violated strict-exogeneity requirements. Models with student fixed effects also performed poorly in validation studies (Kane and Staiger 2008; Kane et al. 2013).

Similar models have been used to estimate the impacts of other educational interventions, such as summer school (Callen et al. 2023), math textbooks (Blazar et al. 2020) and teacher training programs (Plecki et al. 2012; Henry et al. 2014).

In Table 1, we summarize eleven studies testing the validity of value-added estimates of teacher or school effects. In each, researchers estimated a teacher's or school's value-added with one group of students and then tested the prediction in another group of students using a plausibly exogenous design, such as random assignment, a randomized school lottery, annual fluctuations in the make-up of those teaching in a particular school/grade/subject or discontinuities in school assignment.

Table 1 reports estimates of forecast bias for each of the eleven studies, using the coefficient on the empirical Bayes estimate of teacher and school effects. When an estimator is said to be forecast unbiased, the coefficient on the empirical Bayes estimate of impacts across a

---

[5] As part of the Measures of Effective Teaching project, Kane et al. (2013) decomposed value-added estimates of teacher effects into four mutually orthogonal components: a core estimate based on all three categories of controls (lagged achievement, student characteristics and peer achievement), the component that was removed with peer controls, the component that was removed with controls for student characteristics, and the component that was removed with controls for baseline achievement. While the last two components (student achievement and student demographics) played no role in predicting students' achievement post random assignment to teachers, the component associated with peer controls and the core estimate were both predictive of achievement following random assignment. The authors conclude that controlling for peer effects may be "over-controlling" by removing true differences in teacher efficacy.

sample of teachers or schools in predicting student outcomes is indistinguishable from one, i.e. "right on average." However, as discussed in Angrist et al. (2017), a set of estimates can be forecast unbiased, yet still be biased for individual teachers or schools.

Three studies (Kane and Staiger 2008; Kane et al. 2013; Glazerman and Protik 2015) estimated teacher value-added with historical data, randomized teachers and subsequently observed differences in student achievement following random assignment. With one exception (middle school math in Glazerman and Protik [2015]), the authors could not reject that the value-added estimates based on the conventional value-added model and the less common "gain score" model (essentially constraining the coefficient on lagged achievement to be equal to one) are forecast unbiased.[6] The studies involved teachers across 7-14 large urban districts.[7]

As noted above, in models including student fixed effects, Kane and Staiger (2008) found forecast bias coefficients near two—implying that the value-added estimates were understating differences between teachers. As described by Meghir and Rivkin (2011), student fixed effects essentially borrow from a student's future outcomes to estimate the fixed effect. Thus, unless there is 100 percent fade-out of the effect in future periods (i.e., the intervention only affects current year achievement), the student fixed effect models should be biased downward. If there were no fade-out, the bias would be equivalent to dividing by 2, consistent with the findings of Kane and Staiger (2008).

Three more studies compared predictions based on VAM estimates of school effects to the results of randomized school lotteries (Deming 2014; Angrist et al. 2017; Angrist et al. 2024). With one exception, all three studies could not reject that value-added estimates of school effects were forecast unbiased estimates of school effects. The one exception where forecast unbiasedness failed was for high school impacts on SAT math scores in New York, where Angrist et al. (2024) estimate a coefficient (.78) discernibly different from 1, but still large.

As Deming (2014) notes, the finding of forecast unbiasedness in the case of school effects is even more surprising than with teacher effects: while a researcher might plausibly be able to condition on the same data used to make teacher assignments within schools (that is, baseline knowledge, race, gender, program participation, etc.), families presumably sort across schools for many other unobserved reasons, such as the parents' ability/willingness to pay for school quality.

Because their lottery-based estimates are over-identified, Angrist et al. (2017) go beyond "forecast unbiasedness" and test whether the value-added estimates are biased for individual schools. They conclude that while value-added may be forecast unbiased (i.e., "right

[6] Glazerman and Protik (2015) speculate that the middle school math results may have been due to unusually poor compliance with random assignment in those grades/subject.

[7] We say "up to" 14 districts, because while the first two studies name seven districts, Glazerman and Protik do not name the seven "large urban districts" where they conducted their study.

on average"), the estimates for individual schools are indeed biased, with differences larger than would have been expected due to sampling variation alone. As we will report later in the paper, they also find the bias to be negatively correlated with true school effects.

We identified three studies which use non-experimental variation in the make-up of teaching teams within school/grade/subject over time to test the validity of predictions based on teacher's value-added estimates from other groups of students. Using data from a "large northeastern" school district, Chetty et al. (2014a) generated an estimate of forecast bias between .91 and .99. The other two studies essentially replicate Chetty et al.'s methodology in two other, quite different settings: Los Angeles Unified School District (Bacher-Hicks et al. 2014) and the state of North Carolina (Rothstein 2017). Both find similar results when using the previous study's methods. Rothstein (2017), preferred an alternative specification, including changes in achievement in the prior school year, and found a forecast bias coefficient of .860, which was statistically distinct from one (standard error of .017).

Finally, we identified two additional studies which used other quasi-experimental designs to test for validity of school-level value-added estimates. Britton et al. (2023) tested the validity of value-added estimates of middle school impacts on student exam scores in England, using discontinuities in admission eligibility by travel distance from the parent's home to the school. Andrabi et al. (2022) measured the impact on students when schools closed in Pakistan (using pre-closure differences in value-added relative to schools in the same village). Neither study could reject that school-level value-added estimates were unbiased predictors of impacts on students.

Figure 1 portrays the confidence intervals for the forecast bias estimates in Table 1. Only three of the estimates are able to reject forecast unbiasedness: Glazerman and Protik (2015) for middle schools, Angrist et al. (2024) for NY high schools, and Rothstein (2017). Nevertheless, the remainder of the estimates are centered around 1.

In sum, value-added methods have been shown to generate forecast unbiased estimates of school and teacher impacts in a variety of settings, whether using random assignment or quasi-experimental methods to test their validity. However, without any theoretical understanding of the sources of forecast unbiasedness, it has been impossible to generalize from these cases to other cases where validation is not possible.

Model and Estimation

In the model we propose, achievement growth for individual students follows a simple state-space model in which test scores are a noisy measure of an underlying state variable (knowledge). Some innovations—from teachers and other sources—have persistent effects on student knowledge and are passed along to the next period at the same rate as baseline knowledge. In addition to transitory noise in test scores, teachers can have impacts on achievement which disappear after one period. We also allow students to have different pre-existing rates of achievement growth. Similar models have been used in prior work on value added (Jacob et al. 2010; Andrabi et al. 2011).

The simple state-space structure of the model imposes strong testable restrictions on how prior scores and teacher effects enter value added equations (Todd and Wolpin 2003): The recursive Kalman Filter summarizes a student's full history of teachers and test scores in each grade. The structure of the Kalman Filter also provides a unified interpretation for common specifications used in value-added models such as controlling for cubics in prior scores, test scores in other subjects, and test scores from multiple prior grades – each of these terms enter the value-added model because of measurement error in prior scores.

We then use the model to motivate two alternative approaches to estimation and to discuss the potential biases inherent in each approach. The first approach is a standard Value-Added Model (VAM) that estimates teacher effects after controlling for prior year test scores and student characteristics. The second approach, which we refer to as Attenuation Corrected VAM (ACVAM) is similar but corrects for attenuation bias due to measurement error in prior year test scores using an earlier test score to instrument for prior year test scores.

1. Statistical Model

Our model is based on the idea that knowledge is cumulative; each year a student adds to their existing stock of knowledge. We assume that each student's true state of knowledge evolves over time according to a simple structure:

$$(1) \quad \mu_{it} = \delta\mu_{i,t-1} + \theta_{jt} + \alpha_i + \upsilon_{it}$$

Equation (1) defines knowledge (μ_{it}) for student i at time t as the sum of their prior knowledge ($\mu_{i,t-1}$) that depreciates at rate δ and three terms representing new additions to knowledge in year t.[8] [9] The first term (θ_{jt}) is the effect of having teacher j in grade t (or more generally could represent any intervention j to which student i was assigned in grade t).[10] This is usually the key parameter of interest, as it captures the persistent impact of teacher j on student knowledge. The second term (α_i) allows for heterogeneity in knowledge growth across students. This term captures persistent differences across students in family inputs or capacity to learn that results in greater knowledge growth every year. The final term (υ_{it}) is an i.i.d. shock to knowledge for each student. This term captures idiosyncratic student learning each year.

[8] For simplicity, we consider the case of knowledge in a single subject (e.g., math), but it is straightforward to allow for knowledge in each subject to also depend on prior knowledge in other subjects. In our empirical work, we allow for this and find little evidence of such cross-subject effects of prior knowledge.

[9] The state-space model meets the conditions described in Todd and Wolpin (2003) for including contemporaneous inputs and excluding prior inputs while conditioning on prior achievement: with the exception of the prior year teacher's temporary effect, the coefficients on earlier inputs, including initial achievement, decline geometrically at the same rate, δ. The key differences are that our model allows for an individual specific growth term each period, α_i, measurement error in the outcome variable, η_{it}, and allows for a temporary (single year) teacher effect, ψ_{jt}, and permanent teacher effect, θ_{jt}.

[10] Throughout this section we will use grade and year interchangeably to refer to t.

Thus, α_i captures persistent differences across years in a student's knowledge growth, while v_{it} captures the remaining independent shocks.

Papers in education using longitudinal student-level data on test scores (y_{it}) often use hierarchical linear models (HLM) that incorporate student-level random intercepts and slopes, where $y_{it} = \beta_{0i} + \beta_{1i}t + u_{it}$. In these models the slope coefficient (β_{1i}) captures heterogeneity in student growth ($y_{it} - y_{i,t-1}$) and is analogous to α_i in equation (1). Note that if equation (1) is the correct specification, HLM trend estimates (β_{1i}) will be biased because the residual is highly persistent – analogous to spurious trends in random walks. VAM estimates of α_i with δ near 1 are analogous to first differencing, a standard solution for random walks. In ongoing work, we find that VAM models outperform HLM in one-year-ahead forecasts of test scores, as would be expected if VAM was the correct specification and HLM estimated spurious trends.

If knowledge was perfectly measured by test scores, equation (1) would represent a standard value-added regression specification regressing end of year test score on prior year test score and teacher fixed effects. There would be two potential sources of bias arising from the presence of α_i in the error term. First, if student assignment to teachers is non-random and correlated with α_i (e.g., fast learners or higher income students with tutors are assigned to particular teachers) then estimates of teacher effects would be biased. Second, we might expect prior knowledge to be correlated with α_i (e.g., fast learners or students from higher income families may have higher prior knowledge), which would bias the estimate of δ upward. This would result in mis-controlling for prior knowledge and would bias estimates of teacher effects if students were being assigned to teachers based on their prior knowledge (e.g. tracking). As is commonly done in VAM specifications, we will control for a short list of student characteristics (race, ethnicity, gender, economic disadvantage, learning disability, and English language learner) in part to account for variation across students in α_i. Below we propose an alternative method to bias-correct ACVAM using an approach similar to that proposed by Oster (2019).

Hypothetically, one could estimate models such as equation (1) using dynamic panel methods that allow for a student fixed effect in growth (Arellano and Bond 1991; Andrabi et al. 2011), but in practice these methods are poorly identified in short panels with δ near to 1.[11] Alternatively, some VAM models include student fixed effects but then do not control for prior score. If equation (1) is the correct specification, these student fixed-effect models will yield biased estimates of teacher effects. These models implicitly assume that the current teacher has transitory effects that do not accumulate, which will underestimate the effect of teachers if their impact on knowledge are persistent (Meghir and Rivkin 2011). Kane and Staiger (2008)

[11] Alternatively, Blundell and Bond (1998) propose that in steady state, lagged first differences in scores will be uncorrelated with α_i and therefore can be used as an instrument for prior year score. The steady state assumption is unlikely to hold here – i.e., we would expect lagged first differences in scores to be positively correlated with student knowledge growth α_i. Nevertheless, using lagged first differences in scores as an instrument for lagged scores yields estimates that are very similar to ACVAM.

found teacher effect estimates from conventional VAM models with lagged score controls were forecast unbiased while student fixed-effect models were downward biased forecasts when teachers were randomly assigned to classrooms, supporting the VAM model.

Of course, observed test scores (y_{it}) are imperfect measures of accumulated knowledge and contain substantial measurement error. We assume that this measurement error is purely transitory and follows a simple structure:

$$(2)\ y_{it} = \mu_{it} + \psi_{jt} + \eta_{it}\ .$$

The first noise component (ψ_{jt}) is a teacher-level transitory impact that teacher j has on all her students. This component captures teaching to the test or other non-persistent ways of boosting test achievement. We include this component in the model to account for the fact that teacher effects (and other short-term impacts of interventions) are regularly found to partially "fade out" quickly (Koedel et al. 2015). While we focus on teacher effects, this term could represent transitory impacts of any set of interventions indexed by j (e.g., schools, tutoring providers, etc.). The second noise component is at the student level (η_{it}) and represents the measurement error associated with any test, commonly referred to as the standard error of measurement. Typically, standardized tests have reported reliability ratios in the .8-.9 range, which implies that 10-20% of the variance in observed test scores will be due to test measurement error.

2. Relationship of Kalman Filter to VAM Models

One way to think about VAM models is to plug equation (1) directly into equation (2), which yields:

$$(3)\ y_{it} = \delta\mu_{i,t-1} + \left(\theta_{jt} + \psi_{jt}\right) + \ \alpha_i + \upsilon_{it} + \eta_{it}$$

We cannot estimate equation (3) directly because a student's prior knowledge ($\mu_{i,t-1}$) is unobserved. Instead, we form a prediction of $\mu_{i,t-1}$ using information available at time t-1. Call this prediction $\hat{\mu}_{i,t-1|t-1}$ and let $\hat{\mu}_{i,t-1|t-1} = Z_{i,t-1}\pi,$ where $Z_{i,t-1}$ typically includes prior year scores and student characteristics. Substituting $\hat{\mu}_{i,t-1|t-1}$ for $\mu_{i,t-1}$ in equation (3) yields:

$$(4)\ y_{it} = Z_{i,t-1}\pi\delta + \left(\theta_{jt} + \psi_{jt}\right) + \ \alpha_i + \upsilon_{it} + \eta_{it} + \delta\left(\mu_{i,t-1} - \hat{\mu}_{i,t-1|t-1}\right)$$

Equation (4) is analogous to a standard VAM model that controls for prior year scores and student characteristics. In this model, $Z_{i,t-1}\pi\delta$ is the student's expected score at time t, and current teacher effects are estimated based on the difference between actual and expected score at time t (end of year).

The key potential source of bias introduced in equation (4) arises from the fact that the prediction error in predicting true knowledge at baseline $\left(\mu_{i,t-1} - \hat{\mu}_{i,t-1|t-1}\right)$ now appears in the residual. If students are assigned to teachers based on private information about $\mu_{i,t-1}$ that is not captured by $\hat{\mu}_{i,t-1|t-1}$, then estimates of teacher effects derived from equation (4) will be biased. Thus, it is particularly important to incorporate as much information as possible into $\hat{\mu}_{i,t-1|t-1}$. For this reason, VAM models sometimes include a broad set of controls in $Z_{i,t-1}$, including prior year scores in both own and other subjects, cubics in prior scores, and scores from earlier years (t-2, t-3, etc.). However, even the best controls are unlikely to completely capture the private information available to school staff.

Because equations (1) and (2) define a simple state-space model, we can use the Kalman filter to efficiently construct optimal predictions based on the entire history of observable information on each student at baseline. For this calculation we assume $\alpha_i = 0$. Let $\hat{\mu}_{i,t|t-1} = \delta\hat{\mu}_{i,t-1|t-1}$ be the prediction of knowledge at time t given information available at time t-1, and let $\hat{u}_{i,t|t-1} = y_{i,t} - \hat{\mu}_{i,t|t-1}$ be the corresponding prediction error. The Kalman filter estimates $\hat{\mu}_{i,t|t}$ using the following recursive relationship:

$$(5)\ \hat{\mu}_{i,t|t} = \hat{\mu}_{i,t|t-1} + K_t\left(\hat{u}_{i,t|t-1}\right) + \theta_{j,t}$$

Equation (5) states that the optimal prediction in year t updates the prediction from t-1 based on the residual difference between the actual test score in year t and the prediction from t-1, and then adds on the persistent effect of the student's teacher in year t. If $y_{i,t}$ is above (below) the prediction from t-1, then the prediction at time t is revised upward (downward). The weight placed on the residual (K_t) is referred to as the Kalman gain and is less than 1 and falls over time as the existing prediction becomes more precise and less weight is placed on the noisy new information in y_{it}.

Plugging equation (5) into equation (4) yields the Kalman filter estimating equation:

$$\begin{aligned}(6)\ y_{it} = {} & \delta\left(\hat{\mu}_{i,t-1|t-2}\right) + \delta K_{t-1}\left(\hat{u}_{i,t-1|t-2}\right) + \delta\theta_{j',t-1} + \left(\theta_{jt} + \psi_{jt}\right) + \alpha_i + v_{it} + \eta_{it} \\ & + \delta\left(\mu_{i,t-1} - \hat{\mu}_{i,t-1|t-1}\right)\end{aligned}$$

Equation (6) is similar to the conventional VAM specification in equation (4) but replaces a potentially long list of control variables ($Z_{i,t-1}\pi\delta$) with three terms: The predicted test score in t-1, the prediction residual from t-1, and a teacher effect from t-1. The first two terms add up to the test score in t-1, so this is a generalization of the usual VAM specification that distinguishes between the prior prediction of knowledge $\left(\hat{\mu}_{i,t-1|t-2}\right)$ and the new information that comes from the test score in t-1 $\left(\hat{u}_{i,t-1|t-2}\right)$. Because the new information in t-1 includes measurement error, the new information is given lesser weight than the prior prediction in predicting knowledge in time t. Equation (6) is run recursively to obtain estimates of δ and K_t, and then these estimates and the Kalman filter equation (5) are used to estimate $\hat{\mu}_{i,t-1|t-2}$ and $\hat{u}_{i,t-1|t-2}$ for the next grade.[12] To form these predictions, empirical Bayes estimates (best linear

[12] To estimate $\hat{\mu}_{i,t|t}$ in the initial period, when no baseline score is available, we simply run a regression of y_{it} on student covariates.

unbiased predictions) of the prior year's teacher persistent effect ($\theta_{j',t-1}$) from ACVAM are used (e.g., applying shrinkage to the estimated fixed effects [Kane and Staiger 2008]).[13]

The Kalman filter has three useful empirical implications. First, the Kalman filter provides an empirical test of the simple state space model along the lines suggested by Todd and Wolpin (2003). If the model in equations (1) and (2) is correct, then the simplified VAM specification using the Kalman terms (equation 6) should predict as well as an unstructured specification that controls for all prior scores and teacher effects (equation 4). Second, the Kalman filter provides an interpretation for why additional lags of student scores and scores in other subjects enter VAM specifications. Even when knowledge in a subject depends only on prior year knowledge in the same subject (equation 1), scores from earlier years and other subjects help to predict baseline knowledge when prior test scores are noisy estimates of prior knowledge (equation 2). Finally, the Kalman filter provides an explanation for why VAM models commonly find a cubic relationship between current and prior year test scores, with a flatter relationship to prior year score in the tails. Standardized tests commonly have more measurement error in the tails, and the Kalman filter places less weight on prior year scores with more error (a smaller Kalman gain, K_t, for students in the tails). If the relationship between baseline test score and the Kalman gain is approximately quadratic, then the product of the Kalman residual ($\hat{u}_{i,t|t-1} = y_{i,t} - \hat{\mu}_{i,t|t-1}$) and the Kalman gain ($K_t$) in equation (6) introduces a cubic term in prior scores.

3. Conventional VAM Estimation

By relying on observed test scores rather than true student knowledge, value-added models introduce additional potential sources of bias. To see this, note that equation (2) implies that $\mu_{it} = y_{it} - \psi_{jt} - \eta_{it}$, and plugging this into equation (1) yields a familiar VAM-style estimating equation:

$$(7)\ \ y_{it} = \delta y_{i,t-1} + \left(\theta_{jt} + \psi_{jt}\right) - \delta\psi_{j',t-1} + u_{it},\ with\ u_{it} = \alpha_i + v_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right),$$

(where $\psi_{j',t-1}$ represents the transitory effect of teacher j' that student i had in year t-1).

Equation (7) differs from equation (1) in some important ways:

1. The current year teacher effect $\left(\theta_{jt} + \psi_{jt}\right)$ is now the sum of the teacher's persistent impact on knowledge and transitory impact on test scores. Typical VAM models do not distinguish between permanent and transitory teacher effects and simply estimate the teacher's total contemporaneous impact on test scores. It is important to separate out these two components because transitory impacts on test scores have no long-term value (although they may be valued by school administrators under accountability pressure). On average, the total teacher effect may either over or understate a teacher's persistent impact on knowledge depending on the covariance between the permanent

[13] In the empirical work testing the Kalman restrictions, we use estimates of persistent teacher effects from the ACVAM specification described below.

and transitory teacher impact:
$E[\theta_{jt}|(\theta_{jt}+\psi_{jt})] = \beta_{fade}(\theta_{jt}+\psi_{jt})$
where $\beta_{fade} = \frac{Cov(\theta_{jt},\theta_{jt}+\psi_{jt})}{Var(\theta_{jt}+\psi_{jt})} = \frac{Var(\theta)+Cov(\theta,\psi)}{Var(\theta)+Var(\psi)+2cov(\theta,\psi)}$
Moreover, this average fadeout may mask considerable variation across teachers.

2. Equation (7) now includes a prior year teacher effect ($-\delta\psi_{j',t-1}$) that reflects the fading out of the prior-year teacher's transitory effect on the student's test score. Typical VAM models do not control for a student's prior-year teacher. Failing to control for the prior-year teacher could further bias the estimate of the current-year teacher effect unless the prior-year teacher transitory effect is uncorrelated with current-year teacher assignments. However, it is straightforward to add prior year teacher effects to a standard VAM specification, and this estimates both $\theta_{jt}+\psi_{jt}$ and $-\delta\psi_{j',t-1}$, which along with an estimate of δ identifies both persistent and transitory teacher effects.

Finally, even after conditioning on current and prior year teacher, the error in equation (7) still depends on the measurement error in both the current and lagged test score $(\eta_{it} - \delta\eta_{i,t-1})$. This could introduce bias to traditional VAM estimates in two ways. First, the negative correlation between y_{it} and $-\delta\eta_{i,t-1}$ will generate standard attenuation bias on estimates of δ. Again, this would result in under-controlling for prior knowledge, leaving a portion of prior knowledge in the error term. More specifically, if $\hat{\delta}$ is the value of the attenuated coefficient, then the error term will contain $(\delta-\hat{\delta})y_{i,t-1} = (\delta-\hat{\delta})(\mu_{i,t-1}+\psi_{j',t-1}+\eta_{i,t-1})$ which introduces prior knowledge into the error term. If students were assigned to teachers based on better information about their prior knowledge (e.g., from the reports of prior teachers) then estimates of teacher effects will be biased. Similarly, if students are assigned to teachers in part based on knowledge of $\eta_{i,t-1}$ (or equivalently, knowledge of $\mu_{i,t-1}+\psi_{j,t-1}$ if $y_{i,t-1}$ is known), then teacher assignment would also be correlated with the error term. For example, if an administrator could identify which students simply had a bad day on the prior year test and assigned all those students to a particular teacher, that teacher would appear to have large impacts on test score growth through simple mean reversion.

The preceding discussion suggests that a VAM model that regresses end-of-year test score on prior score, current and lagged teacher fixed effects, and some student covariates (to account for α_i) faces three sources of potential bias: (1) sorting of students to teachers based on student-specific growth (α_i) that is not captured by covariates, (2) sorting of students to teachers based on the measurement error in the prior-year score ($\eta_{i,t-1}$), and (3) attenuation of the coefficient δ due to measurement error in prior test scores along with sorting of students to teachers based on prior student knowledge (tracking).

4. Attenuation Corrected Value Added Models (ACVAM)

Since it is likely that students are sorted to teachers based on information about the student's prior knowledge, attenuation of the lagged score coefficient is particularly worrisome in VAM models. A standard solution for attenuation bias due to measurement error is instrumental variables: IV estimates of equation (7) using a twice lagged test score ($y_{i,t-2}$) as an instrument for the lagged score will yield unbiased estimates of δ.[14] The twice lagged score will be correlated with the lagged score through equation (1), yet has measurement error that is independent of the error in equation (7), $u_{it} = \alpha_i + v_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right)$. We refer to this method as attenuation-corrected VAM (ACVAM), to emphasize that the key issue is correcting for the attenuation bias in the estimate of δ.

By eliminating attenuation bias (and controlling for current and lagged teacher effects), ACVAM is no longer biased by teacher assignment correlated with prior knowledge ($\mu_{i,t-1}$) since prior knowledge is uncorrelated with the error in equation (7).[15] To the extent that sorting on prior knowledge (tracking) is an important source of bias in OLS estimates of VAM, this is an important strength of ACVAM.

While conventional VAM is biased by sorting based on students' prior year knowledge, ACVAM is biased if teacher assignment is correlated with prior year measurement error in student scores ($\eta_{i,t-1}$). For example, this would be an important source of bias in ACVAM estimates if teacher assignment was determined solely by prior year scores (e.g., tracking on $y_{i,t-1}$), since baseline scores include the measurement error. However, if the baseline score $y_{i,t-1}$ was not known at the time of assignment, then this bias will not be present. For example, until recently many states did not report end-of-year test scores until the following fall.[16] In such states, teacher assignment could not rely on prior year scores. More generally, if value-added used a fall test from the beginning of the school year as the lagged score, this would ensure that assignment was not based on the baseline test. Similarly, evaluations of non-randomized interventions using ACVAM could avoid bias by using a baseline test administered to all students after assignment.

5. Accounting for heterogeneous student growth (α_i)

As discussed previously, we cannot directly account for heterogeneous student growth (α_i) using student fixed effects because methods that estimate dynamic models with student fixed effects are poorly identified in our setting. The presence of α_i in the error term of equation (7) will bias ACVAM estimates of teacher effects if teacher assignment is explicitly based on

[14] More generally, any score from time t-1 or before with independent measurement error will be a valid instrument. Alternatively, direct knowledge of the amount of measurement error in test scores could be used to correct for attenuation bias. While we do not have this information in our data, it could be a promising approach in other settings.

[15] Prior knowledge may be correlated with the error through α_i, but we consider the bias arising from α_i separately.

[16] Our data come from North Carolina. North Carolina was unusual in that schools and students received their scores often within days of taking the test.

student growth. Moreover, over time we might expect that students with high growth would accumulate more knowledge and have higher scores, generating a positive correlation between α_i and the instrument (twice lagged scores) used in ACVAM. This would bias the ACVAM estimate of δ and open the door for student baseline knowledge to appear in the error term, leading to further potential bias in estimates of teacher effects (as in conventional VAM).

One possible solution is to let $\alpha_i = W_i\gamma + \tilde{\alpha}_i$ and directly control for observable student characteristics (W_i) in estimating equation 7 by ACVAM. If the observed student characteristics explain all the variation in α_i, this approach would eliminate bias. Moreover, we can test for remaining heterogeneity ($Var(\tilde{\alpha}_i)$) using the covariance in the resulting ACVAM errors at 2 or more lags since $u_{it} = \tilde{\alpha}_i + v_{it} + (\eta_{it} - \delta\eta_{i,t-1})$ implies that $Cov(u_{it}, u_{it-k}) = Var(\tilde{\alpha}_i)$ for k≥2. However, if there is still heterogeneity in student growth $\tilde{\alpha}_i$, then any remaining correlation of student growth with the instrument or teacher assignment will result in both an underestimate of $Var(\tilde{\alpha}_i)$ and potential bias in $\hat{\delta}$.

To adjust for any remaining bias in our estimate of δ due to the presence of $\tilde{\alpha}_i$, we use a technique proposed by Oster (2019). Suppose that the degree of selection on the observed ($W_i\gamma$) and unobserved ($\tilde{\alpha}_i$) components is equivalent, i.e. assignment to teachers depends on the sum of the two (α_i). Let δ^W represent our ACVAM estimate of δ controlling for W_i, and let δ^0 represent our ACVAM estimate without controlling for W_i. Oster (2019) shows that an unbiased estimate (δ^*) of δ is given by:

$$(8)\ \delta^* = \delta^W + (\delta^W - \delta^0) \cdot \frac{MSE^W - MSE^{min}}{MSE^0 - MSE^W}$$

Where MSE^W and MSE^0 are the mean squared error from estimating equation (7) by ACVAM with and without controlling for W_i, and MSE^{min} is the minimum mean squared error we could achieve if we observed the remaining unobservables ($\tilde{\alpha}_i$).[17] In Oster's static setting, MSE^{min} is unknown and she suggests various bounding exercises. However, in our dynamic model, the minimum MSE is given by:

$$(9)\ MSE^{min} = Var\left(v_{it} + (\eta_{it} - \delta\eta_{i,t-1})\right) = Var(u_{it}) - Var(\tilde{\alpha}_i)$$

Since $Cov(u_{it}, u_{it-k}) = Var(\tilde{\alpha}_i)$ for k≥2, this provides a second equation which we can use to identify δ^* and MSE^{min}:

$$(10)\ MSE^{min} = Var(u_{it}) - Cov(u_{it}, u_{it-k}),\ \text{k≥2}$$

We use these two equations (8) and (10) to identify the two unknown parameters as follows. For any value of δ^* we re-estimate equation (7) by ACVAM imposing the restriction that $\delta = \delta^*$ and obtain the residuals (u_{it}). With those residuals, we estimate MSE^{min} using

[17] Oster states the last term (the bias correction) equivalently in terms of R-squared values.

equation (10).[18] We then use this estimate of MSE^{min} in equation 8 to obtain a new value of δ^* and iterate the process until it converges. In simulated data this yields consistent estimates of δ.

Even at the true value of δ, this method will underestimate $Var(\tilde{\alpha}_i)$ if assignment to teachers depends on student growth (since some of that variance will be absorbed into the teacher effects). In the appendix we derive an adjustment based on the within and between teacher variance in $W_i\gamma$ (the observable component of α_i) that yields a consistent estimate of $Var(\tilde{\alpha}_i)$ under the same assumption as that needed for the Oster bias correction (that the degree of selection on the observed $(W_i\gamma)$ and unobserved $(\tilde{\alpha}_i)$ components is equivalent). For our estimates, this adjustment makes very little difference to the estimate of $Var(\tilde{\alpha}_i)$.

Overall, an ACVAM model that instruments for prior score with an earlier score, includes current and lagged teacher fixed effects, and uses the Oster (2019) approach to control for student covariates and account for any remaining bias due to $\tilde{\alpha}_i$, faces only one remaining source of potential bias: Sorting of students to teachers based on the measurement error in the prior-year score ($\eta_{i,t-1}$). This bias is eliminated in situations where the baseline score was not known (or otherwise not used) at the time of assignment (either by chance or by design). Unlike conventional VAM estimates, assignment to teachers based on a student's prior knowledge does not bias ACVAM estimates.

6. Tests of model assumptions and potential bias

We perform a series of analyses to explore the plausibility of the model assumptions.

First, we test the statistical assumptions imposed by our model by comparing regressions predicting end-of-grade test scores using the Kalman filter to more and less flexible models. We first test the Kalman filter model against the more restrictive conventional VAM models using only the prior year score and student characteristics as covariates with no lagged teacher effects. We then test the Kalman filter model against less restrictive models that flexibly include student characteristics and a student's complete history of prior scores and teachers. The Kalman model imposes strong restrictions on how these prior variables enter the regression by assuming all prior contributions to knowledge depreciate at the same rate δ. Given our sample size, these tests have high power. As a result, we focus on whether the additional flexibility meaningfully improves the regression's adjusted R-squared.

Second, we consider three empirical tests for the presence of residual student heterogeneity in growth after controlling for student covariates $(\tilde{\alpha}_i)$ in each of our models (conventional VAM and ACVAM). First, we add to each model the student's census tract income (something not typically available in VAM models) to see if income can predict any of the remaining heterogeneity in student growth $(\tilde{\alpha}_i)$. Our second empirical test for the presence of

[18] We average the residual variance across all grades, and average the covariance of all residuals 2 or more years apart (e.g., cov(8,6), cov(8,5) and cov(7,5)).

$\tilde{\alpha}_i$ uses the covariance of each model's residuals two or more years apart to estimate the variance of $\tilde{\alpha}_i$: Since the remaining terms in the residual are expected to be uncorrelated beyond one lag, these covariances are estimates of the variance of student-specific growth. Finally, we implement the bias correction for δ (using the Oster approach described previously) to obtain unbiased estimates of the variance of $\tilde{\alpha}_i$ and the resulting bias (if any) in our estimates of δ.

Third, we explore whether private information about student prior knowledge ($\mu_{i,t-1}$) could contribute to bias in each model. We add to each model the prior teacher's judgement about each student's level of knowledge. For three of the six cohorts we study, teachers provided their subjective judgements of students' mastery of state standards before seeing the end of grade test results. Although typically not available to the value-added researcher, this is the kind of private information that could be used in teacher assignments.[19] As with income, we add this variable to our models and quantify the amount of variation this type of private information can explain.

Fourth, we explore whether sorting on baseline knowledge or measurement error occurs in practice using observable proxies for each component. We use the Kalman prediction of student knowledge at time t-1 $\left(\hat{\mu}_{i,t-1|t-2}\right)$ as an observable proxy for baseline knowledge, and the Kalman prediction residual $\left(\hat{u}_{i,t-1|t-2}\right)$ as an observable proxy for measurement error. While these are imperfect proxies because the Kalman residual also contains unpredicted baseline knowledge in addition to measurement error, they will give an indication of the relative importance of sorting on each source of information. To do this, we estimate teacher random effect models (using the mixed command in Stata) in which the dependent variable is one of the observable proxies measured at the student level. The variance of the random teacher effect indicates how much sorting of students occurred across teachers based on these observable proxies.

Fifth, we use the ACVAM estimates to do a full decomposition of the variance components in the ACVAM model $\left(\theta_{jt}, \varphi_{jt}, \tilde{\alpha}_i, \upsilon_{it}, \eta_{it}, and\ \mu_{i,t}\right)$. We use these variance component estimates to investigate the importance of teacher versus other innovations contributing to learning. We estimate the variance and covariance of the persistent and transitory teacher effects $\left(\theta_{jt}, \psi_{jt}\right)$ using standard empirical Bayes methods that correct for (correlated) estimation error in the fixed effect estimates (Kane and Staiger 2008). This requires estimates of the estimation error in the current and lagged teacher fixed effects, which we estimate using a Bayesian bootstrap (Rubin 1981). These estimates provide direct evidence on whether it is important to account for fadeout in value added models, and we also use them to estimate the average rate of fadeout of teacher effects $\left(\beta_{fade}\right)$. We estimate the variances of $\alpha_i, \upsilon_{it}, \eta_{it}$ using the ACVAM model's residuals $\left(u_{it} = \tilde{\alpha}_i + \upsilon_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right)\right)$. We estimate a Hierarichical Linear Model (using

[19] We also have information on each student's expected grade, which yields very similar results.

the mixed command in Stata) for the student residuals from grades 5-8, allowing for random student intercept ($\tilde{\alpha}_i$) and a Toeplitz error structure with one lag to account for the additional variance and covariance at one lag due to $\upsilon_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right)$. Using our estimate of δ from the ACVAM model, these estimates can be transformed to yield estimates of the variance of υ_{it} and η_{it}. Finally, the variance of μ_{it} is estimated based on equation (2), which implies $Var(\mu_{it}) = Var(y_{it}) - Var\left(\psi_{jt}\right) - Var(\eta_{it})$.

Finally, we explore whether our model yields estimates that are consistent with conventional VAM teacher effects being approximately forecast unbiased as has been found in prior work. Let $\left(\theta_j^{VAM}, \psi_j^{VAM}\right)$ and $\left(\theta_j^{ACVAM}, \psi_j^{ACVAM}\right)$ represent the persistent and transitory teacher effects derived from conventional VAM and ACVAM. For these estimates, we estimated conventional VAM without lagged teacher effects but including teacher means of all control variables (peer effects) as is standard in VAM models, whereas ACVAM excluded the peer measures and included lagged teacher effects. As above, we estimate the variance and covariance of these four parameters using standard empirical Bayes methods that correct for (correlated) estimation error in the fixed effect estimates (derived from a Bayesian bootstrap). Assuming that there is no sorting of students to teachers based on measurement error in the baseline test, the ACVAM parameters are the true teacher effects that would be the estimated in a randomized or quasi-randomized design. A test for forecast bias is analogous to running a regression of the ACVAM teacher effects on the conventional VAM teacher effects. All existing papers test for forecast bias in the total teacher effect $\left(\theta_j^{VAM} + \psi_j^{VAM}\right)$, which is equivalent to testing whether $\pi = 1$ in the following regression:

$$(11) \quad \left(\theta_j^{ACVAM} + \psi_j^{ACVAM}\right) = \pi \cdot \left(\theta_j^{VAM} + \psi_j^{VAM}\right) + e_j$$

We use the variances and covariances of $\left(\theta_j^{VAM}, \psi_j^{VAM}, \theta_j^{ACVAM}, \psi_j^{ACVAM}\right)$ to estimate the slope coefficient π in equation (11) and other statistics of interest such as:

1. The variance of the bias (e.g. the difference between VAM and ACVAM teacher effects). Even when VAM is forecast unbiased on average ($\pi = 1$) VAM may still be biased for individual teachers $\left(Var\left(e_j\right) \neq 0\right)$.
2. The correlation of the bias with the true teacher effect – does the bias accentuate or mute the true differences across teachers?
3. The forecast bias of the permanent and transitory components when estimated using conventional VAM. Even if the total teacher effect $\left(\theta_j^{VAM} + \psi_j^{VAM}\right)$ is forecast unbiased, the separate components may not be.

We further construct all these estimates using only within-school-grade variation across teachers, which aligns with the variation used in all of studies testing whether teacher value added is forecast unbiased. To do this, we construct all the moments above only using within-school-grade variances and covariances of the teacher effects. Finally, we construct analogous estimates of school effects (and within-district estimates) based on ACVAM and VAM estimates of persistent and transitory school effects (simply replacing current and lagged teacher effects with current and lagged school effects).

Data

We use student-level panel data from the North Carolina Education Research Data Center (NCERDC). Our primary analysis sample consists of students in third through eighth grade between 2007 and 2017. We exclude data prior to 2007 due to a lack of student-teacher linkages. We exclude data after 2017 because students could choose to take an end of course Algebra I test rather than the 8th grade end of grade assessment taken by other students. Because of the change in 2018, we would have been missing 8th grade end of grade assessments for a non-random sample of students.

Further, due to the data requirements of the state-space model, we exclude from our sample any students missing an observation in grades 3-8 and those who repeat or skip a grade. We also exclude students missing any of the following data in any year: math and reading scores, race/ethnicity, indicators for economic disadvantage, limited English proficiency, gender, or learning disability, or the id codes for primary math and reading instructor.[20]

For a subset of our analyses, we use data from 2007 through 2013 in which teachers were asked to provide their subjective assessments of students' mastery of state content standards in math and reading. Therefore, our analyses involving prior grade teacher ratings are limited to the three cohorts of students who were in 8th grade between 2012 and 2014. When using prior year teacher judgements, we exclude from our regressions the small percentage of students for whom teacher judgements were missing.

The NCERDC data also include block group identifiers from the U.S. Bureau of the Census. Between 2007 and 2009, the data provided by NCERDC corresponded to the 2000 block group boundaries; between 2010 and 2017, the data provided correspond to 2010 Census block group boundaries. Accordingly, for 2007 through 2009, we use the 2000 Census median household income measure by block group. For 2010 to 2017, we the 2013 ACS 5-year estimate of median household income by block group. Across years, separately for 2007-2009 and 2010-2017, we sort the neighborhood income measures into deciles.[21] Because the block group identifiers are provided for individual students—not school—we observe variation in neighborhood income within teachers and schools and not just between. In the analyses involving neighborhood characteristics, we exclude anyone missing the neighborhood identifiers. This primarily excludes charter school students from the analyses using neighborhood income because NCERDC did not provide block groups for these students.

---

[20] All the methods we discuss can be modified to accommodate students with missing years or data, but we focus on this complete data sample to simplify the presentation and analysis.

[21] While we report results for income deciles, we obtain very similar results using deciles of the block group's Area Deprivation Index.

Results

In **Table 2**, we evaluate how the state space model fits the time series of student achievement in North Carolina. We focus on 8th grade math and reading because students in those grades have 5 years of previous test scores and teachers (providing a strong test of the Kalman filter's ability to summarize the student's prior data). In column (1) for math and column (5) for reading, we report the conventional value-added model as a linear function of prior year math and reading scores, with fixed effects for current year teachers, but not for prior year teachers. As is common in conventional value-added models, the baseline scores in both subjects seem to contribute, with the coefficient on baseline in same subject .669 in math and .596 in reading and smaller coefficients on baseline scores in other subjects: .141 for reading predicting math and .204 for math predicting reading. In columns (2) and (6), we add fixed effects for prior year teachers. The results are largely unchanged: in math, the adjusted R^2 increases slightly, from .753 to .760 and from .679 to .682 in reading.

In columns (3) and (7), we estimate the Kalman filter specification, breaking the lagged score in both subjects into two parts: The Kalman filter prediction of t-1 achievement (which is based on a students' history of achievement and teacher assignments) and its residual (actual baseline score minus the Kalman prediction). The coefficient on the Kalman filter prediction of own-subject lagged score increases to .91 in math and .94 in reading, while the coefficient on the residual of the lagged same subject score is .394 in math and .304 in reading. The difference indicates that the conventional value-added model is mis-specified, since it imposes the constraint that both pieces of baseline achievement carry the same weight in predicting end of year achievement.

Note also that the coefficients on the Kalman prediction for other subject score declines sharply, to about .02 for both reading and math. This finding implies that baseline knowledge in the other subject has little predictive value after conditioning on own subject. The Kalman residuals for other subjects have modest coefficients as would be expected if baseline knowledge was correlated across subjects (.087 when using Kalman reading residual to predict math, and .122 when using Kalman math residual to predict reading). Such findings imply that the only reason that other subject scores play such an important role in conventional value-added is attenuation bias in own subject.

In columns (4) and (8), we add lags 3-5 of prior scores and the Empirical Bayes ACVAM estimates of each prior teacher permanent effect (available in grades 5-7) as additional covariates.[22] The state-space model dictates a specific functional form—reflected in the Kalman filter-- for how prior test scores and prior teacher effects should contribute to baseline achievement. Specifically, the model implies that the influence of prior scores and prior teachers should decline geometrically at each additional lag—one of the key requirements for

[22] Because the model includes fixed effects for prior teacher in the same subject, the same subject teacher permanent effect for grade 7 is not included.

unbiasedness of value-added models identified by Todd and Wolpin (2003). The specifications in columns (4) and (8) relax these restrictions. The change in adjusted R^2 is quite small, .0009 in math and .0003 in reading, implying that the state-space model (as reflected in the Kalman filter) captures the contributions of prior achievement measures and teacher effects. The p-value for the F test on the additional terms in columns (4) and (8) in the bottom of Table 2 formally rejects the hypothesis that the Kalman captures all the information from prior scores and teacher effects, but the large sample size makes that nearly inevitable.

Linear vs. Cubic in Baseline Achievement

In each of the columns of **Table 3**, we compare the average marginal effect of baseline achievement for conventional VAM vs. ACVAM in math and reading. In the top panel, we report results for the linear model. In columns (1) vs. (3) and (6) vs. (8), we compare the coefficients on baseline math and reading for the conventional and ACVAM models in math and reading respectively. As noted above, other subject scores play an important role in conventional value-added, with coefficients of .137 when reading scores predict math and .199 when math scores are used to predict reading. This implies that in conventional value-added—with no adjustment for attenuation bias—other subject scores remain a potential source of selection. We have controlled for math and English scores, but if someone in the school has access to another test measure, such as science or social studies, such scores could also be used for selection.

However, after adjusting for attenuation bias by instrumenting with twice-lagged achievement in (3) and (8), the same subject baseline score predominates (with a coefficient of .952 in math and .959 in reading), while the coefficient on other subject scores is *de minimus* (.006 in reading predicting math and .015 in math predicting reading.) In other words, in the ACVAM model, selection on other subject knowledge, even if they were not included in the model, would generate very little bias. We provide a stronger test of this proposition using the subject ratings of prior teachers later in the paper.

In the bottom panel of **Table 3**, we report the average marginal effect of baseline achievement from a cubic specification, for both the conventional and attenuation-corrected models. In columns (1) to (4) and (6) through (9), the marginal effects are very similar to those reported for the linear model.

Given our large sample size, the specification test of linearity in the bottom of **Table 3** implies that the quadratic and cubic terms are non-zero in all specifications. However, much of the nonlinearity in baseline achievement found in conventional value-added estimation appears to be an artifact of attenuation bias. In **Figure 2**, we report the predicted end of year achievement at different values of the baseline score implied by the cubic specifications for both conventional VAM and ACVAM (i.e., columns (1), (3), (6) and (8)). For the conventional models, predicted score is clearly a cubic function of baseline score, with flatter marginal effects of baseline achievement in both tails, for both math and reading. However, when we estimate the cubic specification with an attenuation correction (instrumenting with the cubic of twice lagged scores), predicted score is nearly linear in baseline score in both math and reading. In the remaining analyses comparing VAM and ACVAM, we use the cubic specification to avoid disadvantaging the conventional value-added model. However, our results for ACVAM are generally similar when using the linear or cubic specifications.

Persistent, Heterogeneous Growth

In columns (5) and (10) of Table 3, we report the results of a bias correction for unmeasured student growth on the baseline achievement, inspired by Oster (2019). If there were substantial heterogeneity in unmeasured growth, one would expect the coefficient on baseline achievement to be biased upward (assuming a positive correlation between such growth and baseline achievement). Although the direction of bias differs, any bias in the baseline coefficient opens the door to selection on private information about a student's true baseline knowledge. Similar to Oster (2019), we use the sensitivity of the baseline coefficient to the inclusion of additional covariates (indicators for race/ethnicity, gender, economic disadvantage, English Language learner and special education student) as an indicator of bias. However, the state space model allows us to do more than that. As an extension of Oster's model, we can directly estimate the total variance in unmeasured growth using the covariance in residual growth at two or more lags. Thus, we iterate until the direct estimate of heterogeneity in growth corresponds with the variance in growth used in the bias adjustment. In math, the bias adjustment lowers the coefficient on baseline achievement slightly from .957 to .954. In reading, the bias adjusted coefficient is the same as the unadjusted ACVAM coefficient to three digits. Thus, for the remainder of the paper, we will use the ACVAM coefficients, not adjusted for correlation with heterogeneous growth.

In **Table 4**, we investigate possible differences in student growth rates by neighborhood, using the decile of each student's block group median household income (which varies within school). Neighborhood income does predict end of year achievement in the conventional value-added specification but has nearly zero predictive power after correcting the baseline achievement measure for attenuation bias. In other words, to the extent that neighborhood income is a potential source of bias, it is largely a result of attenuation bias on the coefficient in baseline achievement in conventional value added. Neighborhood income is correlated with baseline knowledge. But we see little evidence that students in higher income neighborhoods have faster growth.

Evaluating Prior Teacher Ratings as a Potential Source of Selection Bias

Until 2013, North Carolina teachers were asked to provide their own subjective ratings of each student's mastery of the math and English standards when students were sitting down to take the state test.[23] Teachers could choose from five categories: "insufficient mastery" (3.74 percent of valid responses in math and 4.37 percent in English for the three cohorts in our data where judgements are fully available), "inconsistent mastery" (19.19 percent; 20.54 percent), "consistent mastery" (50.71 percent; 50.31 percent), "consistently superior mastery" (26.32 percent; 24.74 percent), or "none of the above" (.04 percent; .04 percent). We use prior teacher responses as proxies for private information about a student's true knowledge not reflected in baseline achievement. (In fact, it is hard to imagine anyone having more accurate information about students' true baseline knowledge than their prior teacher.) In **Table 5**, we report the coefficients on prior teacher ratings after including them as covariates in the conventional and ACVAM models. In the conventional model, they have strong independent predictive power. For instance, relative to those who were rated as having "consistent" mastery, the students who were rated as having "insufficient mastery" at the end of the prior year scored .319 and .304 standard deviations *below* other students with similar baseline achievement in end of year math and reading tests respectively. The students who were rated "consistently superior" scored .209 standard deviations *above* in math and .185 standard deviations *above* in reading than their counterparts with similar baseline scores. Clearly, with an attenuated coefficient on baseline achievement, there is substantial room for selection on private information about a student's true knowledge at baseline.

In columns (2) and (4), we report the coefficients on prior teacher ratings following attenuation correction of the baseline coefficient. Although the F-test would still reject that they are zero, the absolute magnitudes are near zero, .026 and .036 for insufficient mastery in math and English respectively, and .002 and -.009 for consistently superior mastery. In most cases they have switched sign. This is as we would expect. While the conventional VAM baseline coefficient is understated due to attenuation, the unadjusted ACVAM coefficient on baseline achievement would be overstated due to positive correlation with unmeasured growth. (In columns (4) and (9) of Table 3, the adjusted coefficient on baseline achievement was slightly smaller.) Thus, under the unadjusted ACVAM, unobserved factors which are positively related to true knowledge will be negatively related to end of period achievement. The estimated effects of prior teacher ratings would switch sign.

In **Figure 3**, we investigate the predictive power of prior teacher ratings given different values of the baseline coefficient. To do so, we constrain the baseline coefficient (δ^*) to take on values within a .1 band on either side of the ACVAM estimate (while holding the quadratic and cubic coefficients fixed). Using $Y_{it} - \delta^* Y_{it-1}$ as the dependent variable, we re-estimate the value-added specification, pulling out the F-statistic on the teacher rating terms. As seen in Figure 3, the F-statistic on prior teacher ratings is minimized at the ACVAM estimator for math and at a point slightly below the ACVAM estimator in reading, at an adjustment of -.01. The

[23] The North Carolina Department of Public Instruction used such data to evaluate whether the proficiency cut scores were placed appropriately. If there were similar data collected in other states, they should allow for a similar specification test.

horizontal red line is drawn at 2.37, the 5 percent critical value of the F-statistic. Despite the large sample size, we cannot reject the hypothesis that prior teacher ratings play no role once the ACVAM coefficient has been adjusted for correlation with unmeasured growth.

The State-Space Model Parameters

In **Table 6**, we report our estimates of the model parameters in North Carolina. Our estimate of the baseline coefficient (averaged across grades) is .95 in math and .96 in reading. As we reported in Table 3, such estimates are larger than typically found in conventional value-added models (.20 higher in math and .26 higher in English). Such attenuation bias in conventional VAM opens the door to selection on any correlate of true knowledge.

Of the variance in measured achievement (.88 in math and English), we infer that .098 and .139 is due to measurement error (η_{it}) in math and English respectively, implying a score reliability of .89 in math and .84 in reading. Our estimates are derived from observed fluctuations in residual achievement, not from test item parameters. And, yet, they are similar to the psychometric reliability estimates reported by the North Carolina Department of Public Instruction: of .93 in math and .89 in reading (Howard et al. 2020; Mbella et al. 2016).

Even after combining students' histories of scores and prior teacher assignments, the variance in the Kalman filter predictions (.63 in math and .61 in reading) understates the true variance in baseline achievement (.76 in math and .73 in reading). This underscores the futility of adding lags and prior teacher assignments in order to estimate a student's expected end of year achievement. The only solution to selection on private information about a student's true knowledge is ACVAM, combining an unbiased estimate of the baseline coefficient with any unbiased measure of baseline knowledge (even if imperfect.)

We estimate that the variance in unobserved (but persistent) growth rates ($\tilde{\alpha}_i$) is .0008 in both math and reading, implying a standard deviation in persistent growth of .028. Although not zero, the estimate can be used in bounding the potential bias due to selection on persistent growth. We estimate an additional .0007 and .0010 variance in unobserved growth in math and reading respectively is absorbed by observed student characteristics (indicators for race/ethnicity, gender, economic disadvantage, English language learners and special education). In other words, about half of the variance in persistent growth rates is accounted for by including the student covariates commonly used in value-added models.

We estimate the variance in persistent impacts of teachers (θ_j) of .0135 in math and .0075 in reading, implying a standard deviation of .12 and .09 respectively. However, the variance in teacher effects is small relative to our estimates of the variance of other unobserved, but persistent, shocks to achievement which occur each period: .040 in math and .036 in reading (a .20 and .19 standard deviation shock respectively.) These i.i.d. shocks are either due to in-school interventions, environmental changes or developmental changes which are incorporated into baseline knowledge and passed on to the next period, but do not seem to be predictable.

In addition, we find evidence of substantial variance in transitory teacher effects on measured achievement (.024 in math and .009 in reading) which disappear after one period. Moreover, we find that teachers with larger transitory effects on student achievement tend to have smaller persistent effects (a correlation of -.09 in math and -.29 in reading). Others have reported evidence of fade-out of teacher effects (Jacob et al. 2010; Kane and Staiger 2008; Rothstein 2010). For example, Chetty et al. (2014b) find that teacher effects fade out by approximately 75 percent after 4 years. However, like Jacob et al. (2010), our findings suggest that the degree of fade-out varies from teacher to teacher. We estimate the average first-year fade-out of teacher effects to be .34 in math and .45 in reading.

All prior work estimating and validating teacher effects has focused on the total teacher (persistent plus transitory). Our estimates of the variance of the total teacher effect of .0346 (SD=.19) in math and .0115 (SD=.11) in reading are consistent with the literature, but these are only correlated about 0.55 with the teacher's persistent effect on knowledge.

In **Figure 4**, we report the covariance in ACVAM residuals over multiple lags. The ACVAM residual consists of three components: measurement error ($\eta_{it} - \delta\eta_{it-1}$), a persistent innovation (v_{it}), and the student's persistent growth rate ($\tilde{\alpha}_i$). In the current period (lag 0), the variance in residuals would represent the sum of all three components and is estimated to be approximately .22 in math and .31 in reading. At one lag, the shared components are $-\delta\eta_{it-1}+ \tilde{\alpha}_i$ from the current period and $\eta_{it-1}+ \tilde{\alpha}_i$ from the first lag. Because of the opposite signs on the shared measurement error component, the covariance is estimated to be negative. However, at two or more lags, the only shared component in the residuals is $\tilde{\alpha}_i$, meaning that the covariance in residuals is a direct estimate of the variance in residual growth. Yet, as reported in Figure 4, we find close to zero covariance in the ACVAM residuals at two or more lags, implying close to zero variance in residual growth.

In **Figure 5**, we report an out-of-sample test for teachers' temporary and persistent effects on students. In the top panel, we sort students by the persistent effect of their 5th grade teacher (θ_{jt}). Similar to Chetty et al. (2014b), we "leave out" the student's cohort and estimate the 5th grade teacher's effectiveness only with the data from the other cohorts. Those with a top quartile teacher in 5th grade not only have higher achievement at the end of 5th grade, but the gap in mean achievement largely persists into 6th, 7th and 8th grade. This is what we would expect in a state-space model with the decay rate close to zero.

In the bottom panel, we sort students by the transitory effectiveness of their 5th grade teachers (ψ_{jt}). As above, we "leave out" the student's cohort and estimate the teacher's effectiveness only with the data from other cohorts. Those who were assigned 5th grade teachers estimated to have only transitory effects on students, indeed, witnessed large swings in achievement in 5th grade, which essentially disappeared in 6th, 7th and 8th grade.

The Nature of Teacher Sorting

In **Table 7**, we investigate the amount of sorting on student baseline achievement, using various measures of baseline achievement as the outcome and estimating the teacher effect as a random effect. There is a high degree of sorting by teacher on prior year scores, with the standard deviation in the teacher component being .468 in math and .456 in reading. This means that the variance in mean baseline scores by teacher is about a quarter of the total variance in student scores.

We also report the amount of sorting on the Kalman prediction of baseline achievement and its residual. Most of the sorting is based on the predictable component of baseline achievement using achievement and teacher effects from t-2 and earlier, with a standard deviation in the teacher effect of .410 and .412. There is less sorting on the Kalman residual, which combines measurement error and new additions to student knowledge. This pattern suggests that sorting on baseline knowledge (reflected in the Kalman prediction) is the most important source of potential bias in value added models, while sorting on measurement error (reflected in the Kalman residual) is of less concern.

Such a high degree of sorting highlights the importance of correcting for attenuation bias in the baseline coefficient. Even a small amount of bias in the baseline coefficient can open the door to bias from student sorting on baseline knowledge.

Evaluating Bias and Forecast Bias in Conventional Value-Added

If conventional value-added models are biased for individual teachers, why have prior studies found them to be forecast unbiased? Under the state-space model, ACVAM should yield noisy estimates of true teacher effects. Thus, we use these to calculate the implied forecast bias in the conventional VAM models in North Carolina.

Prior studies of forecast bias have focused on the total teacher effect, combining teachers' persistent effects on student achievement (θ_{jt}) with their transitory effects (ψ_{jt}). Thus, we focus on the sum of a teacher's persistent and transitory effects (θ_{jt}+ψ_{jt}). Moreover, as reported in Table 1, the six papers testing for forecast bias in conventional value-added estimates of teacher effects all included covariates for peer characteristics. We do the same, including a covariate for mean baseline achievement of peers in the conventional model of teacher effects, but using our ACVAM model without peer effects as estimating the "true" teacher effect. Because the five papers in Table 1 estimating school effects do not include peer effects, we exclude them from both the ACVAM and conventional VAM estimates of school effects in Table 8.

As we report in columns 1 and 3 of **Table 8**, we estimate the math forecast coefficient on conventional value-added estimates of teacher and school effects to be .91 for teachers and .83 for schools. These are both somewhat smaller than the prior literature. However, due to limitations on their ability to randomly assign students or teachers across schools or assign students to schools outside their current district, random assignment studies have been done within-school for teachers and within-district for schools. When we limit comparisons within-school and within-district for teachers and schools, we find higher forecast coefficients of .95 for teachers and .84 for schools.

In reading, there is more evidence of forecast bias remaining, with a within-school forecast coefficient of .89 for teachers and a within-district forecast coefficient of .65 for schools.

We estimate the standard deviation in the bias in conventional value added for teachers to be .088 in math and .077 in reading.[24] Such bias appears sizable relative to the standard deviation in ACVAM estimates of .186 in math and .107 in reading. How is it, then, that the forecast coefficient in conventional value-added is so close to one, .91? A key reason is that the covariance between the bias and true teacher (and school) effects is negative—that is, higher value-added teachers and schools receive students with slightly weaker baseline achievement. Even with substantial bias for individual teachers, the forecast coefficient will equal 1 if the variance in the bias is of equal magnitude but opposite sign as the covariance between the bias and the true teacher (or school) effect.[25] As reported in Table 8, the variance in the bias for teacher effects in math is .0078, while the covariance between the bias and the true teacher effect is of opposite sign and slightly smaller magnitude, -.0049. The net result of overstating the signal variance in teacher/school effects by adding bias and underestimating the difference between the best and worst teachers/schools is a forecast coefficient close to one.

Conclusion

We conclude that the answer to the puzzle of forecast unbiasedness in conventional value-added lies in two time-series properties of student achievement: student knowledge accumulation follows a state-space model and there is little latent heterogeneity in students' growth rates with which to predict future growth (especially after conditioning on student demographics). Moreover, we argue that the existing bias in conventional value-added is primarily due to attenuation of the baseline coefficient, which opens the door to selection on correlates of students' baseline knowledge. To close that door, we propose an attenuation-corrected value-added model (ACVAM) and provide diagnostic tests for determining whether the conditions required for its unbiasedness hold.

In some ways, our findings are analogous to Robert Hall's work on aggregate consumption. Before Hall, macroeconomic modelers attempted to predict future consumption

[24] Angrist et al. (2017) estimate the bias in school level value-added to be .182 in math, but also report a small negative correlation between the bias and true school effects.

[25] If the conventional value-added estimator of the current year effect for teacher j is $VAM_j=(\theta_j^{VAM}+\psi_j^{VAM})$, and $ACVAM_j=(\theta_j^{ACVAM}+\psi_j^{ACVAM})$ and the bias in VAM_j is b_j, then the forecast bias coefficient will equal $\frac{Cov(ACVAM_j,VAM_j)}{Var(VAM_j)}=\frac{Cov(ACVAM_j,ACVAM_j+b_j)}{Var(ACVAM_j+b_j)}=\frac{Var(ACVAM_j)+Cov(ACVAM_j,b_j)}{Var(ACVAM_j)+Var(b_j)+2Cov(ACVAM_j,b_j)}$. The ratio will equal one if $Var(b_j)=-Cov(ACVAM_j,b_j)$.

by assembling as many relevant predictors of consumption as could be found, including hard-to-measure concepts such as wealth. Like conventional value-added modellers, they were forced to assume access to the same information that actors in the economy had. That is a tenuous position to defend. However, after presenting evidence that consumption followed a random walk, Hall (1978) concluded that "no variable apart from current consumption should be of any value in predicting future consumption".[26]

The state-space model of student knowledge accumulation provides a similar implication: no private information can improve upon the prediction that a student's knowledge next year will equal the product of baseline knowledge and the knowledge persistence rate, $\delta\mu_{i,t-1}$. As long as one has an unbiased estimate of the knowledge persistence rate, δ, then selection on any private information about true baseline knowledge (such as prior teacher ratings) or family background traits (such as neighborhood income) cannot cause bias—because such information has no incremental predictive power.[27]

Value-added models, as with any non-experimental method, have been treated with skepticism as a program evaluation tool, given the widespread presumption that students are being assigned to treatment based on unobserved factors. However, in the state-space model, selection on private information about baseline knowledge does not lead to bias, once attenuation has been addressed. Moreover, in our data, we find surprisingly little evidence of heterogeneity in students' growth rates on which to select. The absence of such heterogeneity may explain why conventional value-added models have yielded forecast unbiased estimates of schools —not just teachers (Deming 2014; Angrist et al. 2017; Angrist et al. 2024). Because they have access to administrative data, researchers might believe they are controlling for the factors used to assign students to teachers within schools. However, one would expect families choosing charter schools or higher achieving public schools to differ in many ways not available to the researcher (e.g. family homework routines, student motivation and parental help with homework.) And, yet, controlling for baseline achievement and a few student characteristics seems to yield school impact estimates consistent with school admission lotteries. The implication is that whatever unobserved factors lead families to choose charter schools do not seem to predict achievement growth once baseline achievement has been taken into account.

Assuming that the same time series properties of achievement that we observe in North Carolina prevail elsewhere, our findings would justify an expanded effort to evaluate educational products and interventions using ACVAM. Given the economic value of student achievement for earnings and productivity growth, even a small increase in the rate of discovery

[26] As noted by John Cochrane, Hall's key insight was that "any estimate or test must still work if agents in the economy have more information than economists put in our models…" (Cochrane 2024). Value-added modellers will never have the same information as school personnel or parents. And yet, value-added models can still be unbiased.

[27] The primary remaining threat of bias is selection on the measurement error in baseline achievement, $\eta_{i,t-1}$. If one measure of baseline achievement is used in treatment assignment, a solution would be to collect another independent measure of baseline achievement and instrument for that.

and spread of effective interventions would produce large social returns (Hanushek and Woessman 2008; Doty et al. 2025). To that end, we will be testing the validity of the ACVAM model in comparison to randomized trials of tutoring interventions, where one would expect substantial selection.

Our results also have important implications for the sources of inequality in student achievement in later grades. If there is little latent heterogeneity in student knowledge growth (at least after 3rd grade, when we begin observing students), then much of the inequality in later achievement is due to the quality of teachers, schools, summer camps and other interventions that students receive, not pre-existing differences in rates of growth.

Moreover, our results have implications for other literatures using panel data on achievement. For example, our estimates imply strong persistence of additions to student knowledge, with a knowledge persistence rate (δ) near to one. Apparently, student test scores contain a large unit root component (plus some transitory measurement error). It is well known in the time series literature that unit roots can generate spurious student-level trends and spurious associations with other trending variables. They also introduce bias into models with student fixed effects. Other aggregations of student achievement, such as school and district averages, may also be affected – with unit roots in the aggregate measures as well. We will be exploring those implications in future research.

References:

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. ”Teachers and Student Achievement in the Chicago Public High Schools.” *Journal of Labor Economics* 25 (1): 95–135. https://doi.org/10.1086/508733.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools.” *Journal of Political Economy* 113 (1): 151–184. https://doi.org/10.1086/426036.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." *American Economic Journal: Applied Economics* 3 (3): 29–54. http://www.aeaweb.org/articles.php?doi=10.1257/app.3.3.29.

Andrabi, Tahir, Natalie Bau, Jishnu Das, and Asim Ijaz Khwaja. 2022. “Heterogeneity in School Value-Added and the Private Premium.” Working Paper No. 30627. National Bureau of Economic Research. http://www.nber.org/papers/w30627.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. “Leveraging Lotteries for School Value-Added: Testing and Estimation.” *The Quarterly Journal of Economics* 132 (2): 871–919. doi:10.1093/qje/qjx001.

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, Christopher R. Walters. 2024. “Credible School Value-Added with Undersubscribed School Lotteries.” *The Review of Economics and Statistics* 106 (1): 1–19. https://doi.org/10.1162/rest_a_01149.

Arellano, Manuel, and Stephen Bond. 1991. “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations.” *The Review of Economic Studies* 58 (2): 277–297. https://doi.org/10.2307/2297968.

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 No. 20657. National Bureau of Economic Research. https://www.nber.org/papers/w20657.

Blazar, David, Blake Heller, Thomas J. Kane, Morgan Polikoff, Douglas O. Staiger, Scott Carrell, Dan Goldhaber, Douglas N. Harris, Rachel Hitch, Kristian L. Holden, and Michal Kurlaender. 2020. “Curriculum Reform in the Common Core Era: Evaluating Elementary Math Textbooks Across Six U.S. States.” *Journal of Policy Analysis and Management* 39 (4): 966–1019. https://doi.org/10.1002/pam.22257.

Blundell, Richard, and Stephen Bond. 1998. “Initial Conditions and Moment Restrictions in Dynamic Panel Data Models.” *Journal of Econometrics* 87 (1): 115–143. https://doi.org/10.1016/S0304-4076(98)00009-8.

Bollen, Kenneth A., and Patrik J. Curran. 2006. *Latent Curve Models: A Structural Equation Perspective*. John Wiley & Sons.

Britton, Jack, Damon Clark, and Ines Lee. 2023. “Exploiting Discontinuities in Secondary School Attendance to Evaluate Value Added.” Working Papers 23/24. Institute for Fiscal Studies, London, UK. https://doi.org/10.1920/wp/ifs.2023.2523.

Callen, Ian, Maria V. Carbonari, Michael DeArmond, Daniel Dewey, Elise Dizon-Ross, Dan Goldhaber, Jazmin Isaacs, Thomas J. Kane, Megan Kuhfeld, Anna McDonald, Andrew McEachin, Emily Morton, Atsuko Muroga, and Douglas O. Staiger. 2023. “Summer School as a Learning Loss Recovery Strategy After COVID-19: Evidence From Summer 2022.” Working Paper No. 291-0823. Center for Analysis of Longitudinal Data in Education Research, American Institutes for Research.

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 (9): 2593–2632. http://dx.doi.org/10.1257/aer.104.9.2593.

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 (9): 2633–2679. http://dx.doi.org/10.1257/aer.104.9.263.

Cochrane, John H. 2024. “Bob Hall and Consumption.” Paper presented at Hoover Institution Celebration of the Service of Robert E. (Bob) Hall, November 22. https://static1.squarespace.com/static/5e6033a4ea02d801f37e15bb/t/6764eaf27b14934e6350d992/1734666997291/Hall.pdf.

Deming, David J. 2014. “Using School Choice Lotteries to Test Measures of School Effectiveness.” *American Economic Review* 104 (5): 406–11. DOI: 10.1257/aer.104.5.406.

Doty, Elena, Thomas J. Kane, Tyler Patterson, and Douglas O. Staiger. “What Do Changes in State NAEP Scores Imply for Birth Cohorts’ Later Life Outcomes*?”* Preprint, *Journal of Policy Analysis and Management*, June 6, 2025, https://doi.org/10.1002/pam.70018.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael J. Podgursky. 2014. “The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri.” *Statistics and Public Policy* 1 (1): 19-27. https://doi.org/10.1080/2330443X.2013.856152.

Glazerman, Steve, and Ali Protik. “Validating Value-Added Measures of Teacher Performance.” Working Paper. Mathematica Policy Research, February 20, 2015.

Grilliches, Zvi. 1977. “Estimating the Returns to Schooling: Some Econometric Problems.” *Econometrica* 45 (1): 1–22. https://doi.org/10.2307/1913285.

Hall, Robert E. 1978. “Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence.” *Journal of Political Economy* 86 (6): 971–987. https://www.jstor.org/stable/1840393.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review Papers and Proceedings* 61 (2): 280–288. https://www.jstor.org/stable/1817003.

Hanushek, Eric A. 1979. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *The Journal of Human Resources* 14 (3): 351–388. https://doi.org/10.2307/145575.

Hanushek, Eric A., and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46 (3): 607–668. https://doi.org/10.1257/jel.46.3.607.

Henry, Gary, Kelly M. Purtell, Kevin C. Bastian, C. Kevin Fortner, Charles L. Thompson, Shanyce L. Campbell, and Kristina M. Patterson. 2014. "The Effects of Teacher Entry Portals on Student Achievement." *Journal of Teacher Education* 65 (1): 7–23. https://doi.org/10.1177/0022487113503871.

Howard, Tammy, Maxey Moore, Kinge Mbella, and Thakur Karkee. 2020. *Mathematics 3-8 End-of-Grade (EOG) NC Math 1 and NC Math 3 End-of-Course (EOC) Technical Report 2018-2019*. 5th ed. North Carolina Department of Public Instruction. https://www.dpi.nc.gov/technical-report-eog-mathematics-and-eoc-math-i-2018-19-edition-5/open.

Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources* 45 (4): 915–943. https://www.jstor.org/stable/25764794.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*. Bill & Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper No. 14607. National Bureau of Economic Research. https://www.nber.org/papers/w14607.

Koedel, Cory, Kata Mihaly, and Jonah Rockoff. 2015. "Value-Added Modelling: A Review." *Economics of Education Review* 47: 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006.

Mbella, Kinge, Min Zhu, Thakur Karkee, and Hope Lung. 2016. *The North Carolina Testing Program Technical Report 2012-2015: English Language Arts/Reading Assessments (ELA) End-of-Grade 3-8 and End-of-Course English II*. North Carolina Department of Public Instruction. https://www.dpi.nc.gov/documents/files/technical-report-2012-2015-ela-eog-3-8-and-eoc-english-ii/open.

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 (1): 67–101.

McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4 (4): 572–606. https://doi.org/10.1162/edfp.2009.4.4.572.

Meghir, Costas, and Steven G. Rivkin. "Econometric Methods for Research in Education." In *Handbook of the Economics of Education Vol. III*, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier, 2011. https://doi.org/10.1016/B978-0-444-53429-3.00001-6.

Miller, Matt L., Emilio Ferrer, and Paolo Ghisletta. 2024. "The Current Practice of Latent Growth Curve Modeling in the Social and Behavioral Sciences: Observations and Recommendations." *International Journal of Behavioral Development* 49 (4): 389–397. https://doi.org/10.1177/01650254241269723.

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

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* 37 (2): 187–204. https://doi.org/10.1080/07350015.2016.1227711.

Plecki, Margaret L., Ana M. Elfers, and Yugo Nakamura. 2012. "Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures." *Journal of Teacher Education* 63 (5): 318–334. https://doi.org/10.1177/0022487112447110.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1): 175-214. https://doi.org/10.1162/qjec.2010.125.1.175.

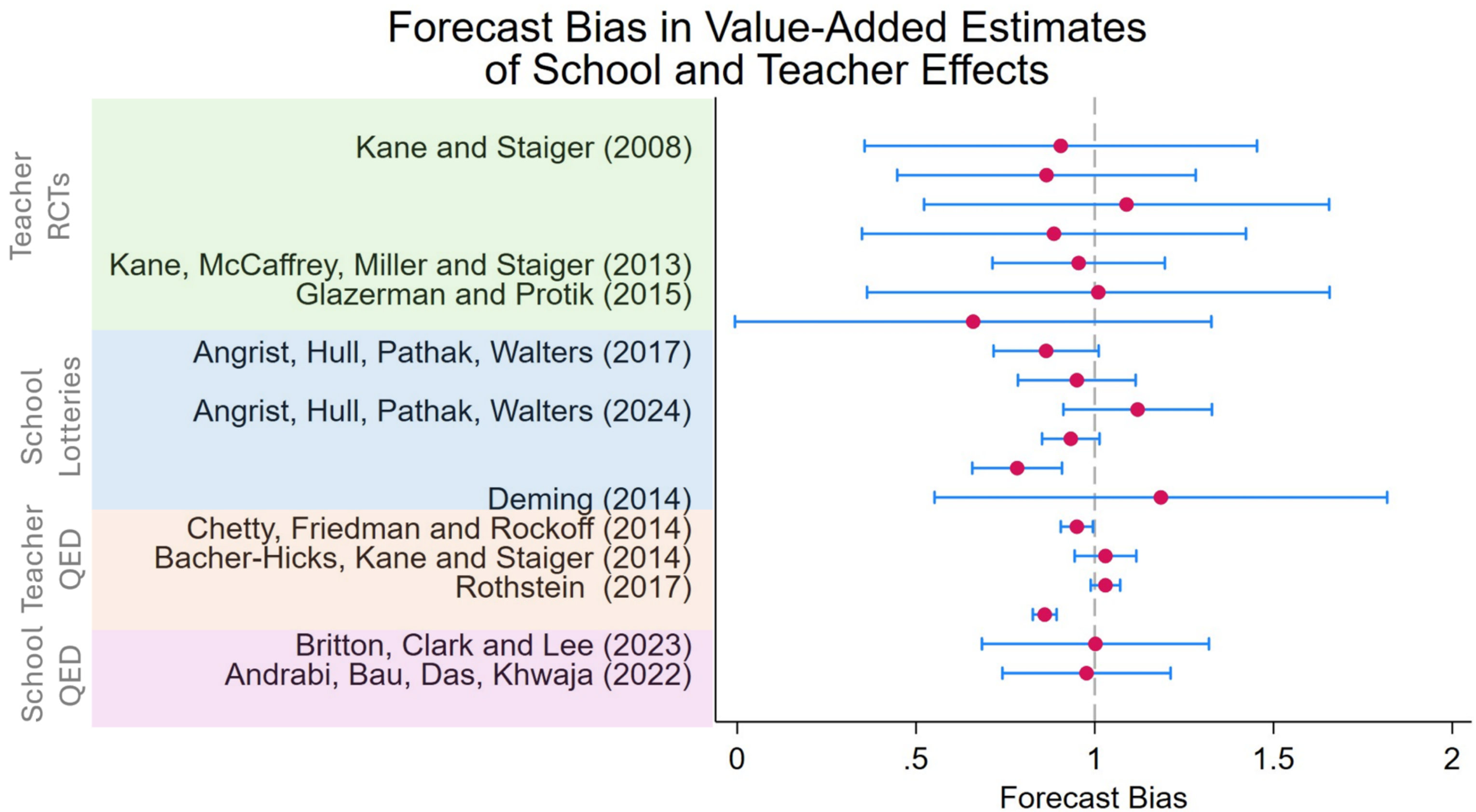
Rothstein, Jesse. 2017. "Revisiting the Impacts of Teachers." Working Paper No. 101-17. Institute for Research on Labor and Employment, UC Berkeley. http://irle.berkeley.edu/files/2017/Revisiting-the-Impacts-of-Teachers.pdf.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *The Annals of Statistics* 9 (1): 130-134. https://www.jstor.org/stable/2240875.

Summers, Anita A., and Barbara L. Wolfe. 1977. "Do Schools Make a Difference*?" American Economic Review* 67 (4): 639- 652. https://www.jstor.org/stable/1813396.

Todd, Petra, and Kenneth Wolpin. 2003. “On the Specification and Estimation of the Production Function for Cognitive Achievement.” *The Economic Journal* 113 (485): F3–F33. https://www.jstor.org/stable/3590137.

Figure 1.



Note: A coefficient of one implies forecast unbiasedness. See Table 1 for details on the specifications, grade levels, subject, and type of validation.

Table 1. Forecast Bias in Value-Added Estimates of School and Teacher Effects

	Study:	Teacher/ School	Grade Level	Subject	Setting	Value-Added Model	Type of Validation	Forecast Bias
1	Kane and Staiger (2008)	Teacher	Elem.	Math	Los Angeles	Lagged Score (w demogs & peer controls)	Random assignment	.905 (.180)
		Teacher	Elem.	Math	Los Angeles	Gain Score (w demogs & peer controls)	Random assignment	.865 (.213)
		Teacher	Elem.	Reading	Los Angeles	Lagged Score (w demogs & peer controls)	Random assignment	1.089 (.289)
		Teacher	Elem.	Reading	Los Angeles	Gain Score (w demogs & peer controls)	Random assignment	.886 (.274)
		Teacher	Elem.	Math	Los Angeles	Student f.e.	Random assignment	1.859 (.470)
		Teacher	Elem.	Reading	Los Angeles	Student f.e.	Random assignment	2.144 (.635)
2	Kane, McCaffrey, Miller and Staiger	Teacher	Elem./Middle	Math/Reading	New York, Denver, Dallas Memphis, Hillsborough FL, Charlotte	Lagged Score (w demogs & peer controls)	Random assignment	.955 (.123)
3	Glazerman and Protik (2015)	Teacher	Elem.	Math	7 Large Districts	Lagged Score (w demog controls)	Random assignment	1.01 (.33)
		Teacher	Elem.	Reading	7 Large Districts	Lagged Score (w demog controls)	Random assignment	.66 (.34)
		Teacher	Middle	Math	7 Large Districts	Lagged Score (w demog controls)	Random assignment	-.05 (.36)
		Teacher	Middle	Reading	7 Large Districts	Lagged Score (w demog controls)	Random assignment	0.75 (.99)
4	Angrist, Hull, Pathak, Walters (2017)	School	Middle	Math	Boston	Lagged Score (w demog controls)	Random assignment	.864 (.075)
		School	Middle	Math	Boston	Gain Score (w demog controls)	Random assignment	.950 (.084)
5	Angrist, Hull, Pathak, Walters (2024)	School	Middle	Math	Denver	Lagged Score (w demog controls)	Random assignment	1.12 (.106)
		School	Middle	Math	New York	Lagged Score (w demog controls)	Random assignment	.933 (.041)
		School	H.S.	SAT Math	New York	Lagged Score (w demog controls)	Random assignment	.783 (.064)
6	Deming (2014)	School	Elem./Middle	Math/Reading	Charlotte	Lagged Score (w demog controls & drift adjustment)	Random assignment	1.185 (.323)
7	Chetty, Friedman and Rockoff (2014)	Teacher	Elem./Middle	Math/Reading	Large district	Lagged Score (w demogs, peers & drift adjustment)	Quasi Exp (shifting teacher assignments)	.950 (.023)
8	Bacher-Hicks, Kane and Staiger (2014)	Teacher	Elem./Middle	Math/Reading	Los Angeles	Lagged Score (w demogs, peers & drift adjustment)	Quasi Exp (shifting teacher assignments)	1.030 (.044)
9	Rothstein (2017)	Teacher	Elem.	Math/Reading	North Carolina	Lagged Score (w demogs, peers & drift adjustment)	Quasi Exp (shifting teacher assignments)	1.030 (.021)
		Teacher	Elem.	Math/Reading	North Carolina	Lagged Score (w demogs, peers & drift adjustment)	Quasi Exp (shifting teacher assignments) + change in prior year mean score	.860 (.017)
10	Britton, Clark and Lee (2023)	School	Middle/H.S.	Math/Reading	England	Lagged Score (w demog controls) (No Shrinkage)	Quasi Exp (RD in distance to school)	1.002 (.162)
11	Andrabi, Bau, Das and Khwaja (2022)	School	Elem./Middle	Math, English, Urdu	Pakistan	Lagged Score	Quasi Exp (School Closures)	.977 (.120)

Table 2. Testing the Kalman Filter: 8th Grade Math and Reading (Linear Specifications)

	Math				Reading			
	Conventional VAM (1)	Conventional VAM + Lagged Teacher (2)	Kalman Filter (3)	Augmented Kalman (4)	Conventional VAM (5)	Conventional VAM + Lagged Teacher (6)	Kalman Filter (7)	Augmented Kalman (8)
G7 Math	0.669***	0.678***	-	-	0.204***	0.206***	-	-
	(0.002)	(0.002)	-	-	(0.002)	(0.002)	-	-
G7 Kalman Math Prediction	-	-	0.905***	0.899***	-	-	0.019***	0.030***
	-	-	(0.002)	(0.006)	-	-	(0.002)	(0.006)
G7 Kalman Math Residual	-	-	0.394***	0.393***	-	-	0.122***	0.124***
	-	-	(0.002)	(0.002)	-	-	(0.002)	(0.002)
G7 Reading	0.141***	0.134***	-	-	0.596***	0.593***	-	-
	(0.001)	(0.001)	-	-	(0.002)	(0.002)	-	-
G7 Kalman Reading Prediction	-	-	0.021***	0.041***	-	-	0.943***	0.928***
	-	-	(0.002)	(0.006)	-	-	(0.003)	(0.007)
G7 Kalman Reading Residual	-	-	0.087***	0.089***	-	-	0.304***	0.304***
	-	-	(0.002)	(0.002)	-	-	(0.002)	(0.002)
G8 Subject Teacher	X	X	X	X	X	X	X	X
G7 Subject Teacher		X	X	X		X	X	X
G3-5 Math				X				X
G3-5 Reading				X				X
G5-7 Subject Theta				X				X
G5-7 Other Subject Theta				X				X
Adj. R^2	0.7533	0.7597	0.7882	0.7891	0.6787	0.6819	0.7298	0.7301
Difference in Adj. R^2	-0.0349 (rel Col 3)	-0.0285 (rel Col 3)	0 (rel Col 3)	0.0009 (rel Col 3)	-0.0511 (rel Col 7)	-0.0479 (rel Col 7)	0 (rel Col 7)	0.0003 (rel Col 7)
P-value of F-test of Constraint	-	0.000 (rel Col 3)	-	0.000 (rel Col 3)	-	0.000 (rel Col 7)	-	0.000 (rel Col 7)
N	272639	272637	272637	272637	272640	272638	272638	272638

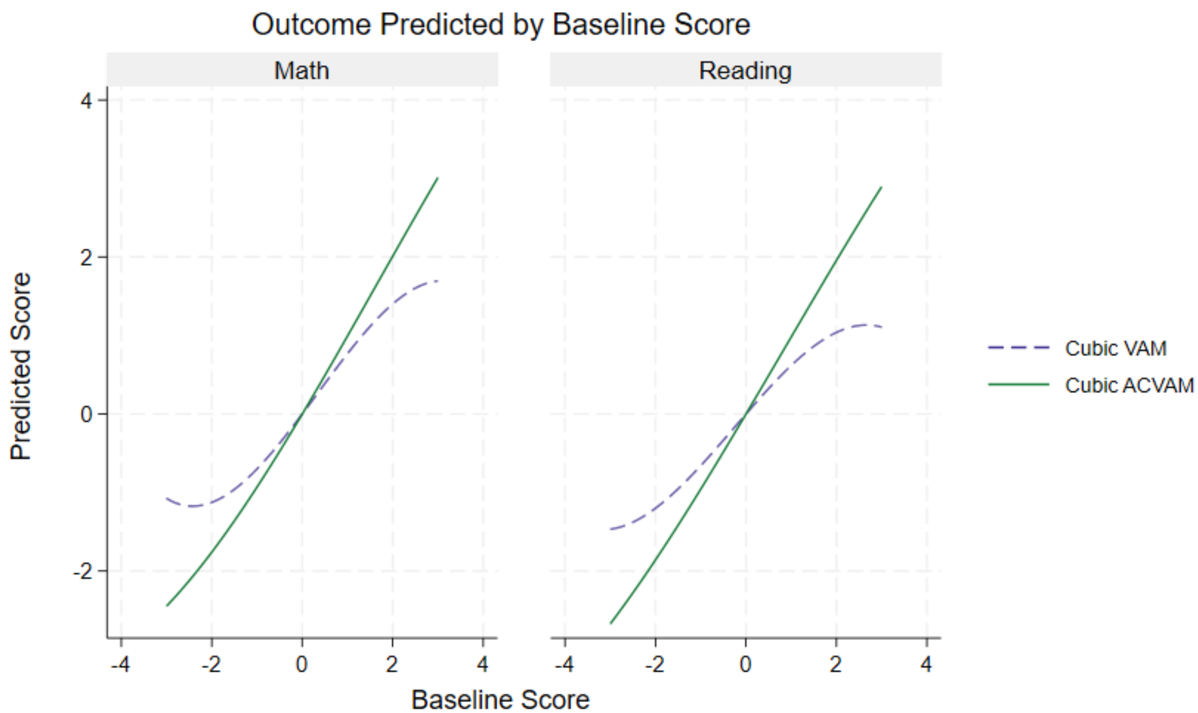
Note: All models are linear in the covariates and include student-level demographic controls (gender, race/ethnicity, economic disadvantage, special education status, and English language learner status) as well as year fixed effects. Kalman predictions incorporate shrunken permanent teacher effects from a linear ACVAM model in grades 5-7. Sample fluctuates slightly because singletons are dropped with multiple fixed effects. Standard errors are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3. The Average Marginal Effect of Baseline Achievement

	Math					Reading				
	Conventional VAM		ACVAM			Conventional VAM		ACVAM		
	Both Subjects	Own Subject	Both Subjects	Own Subject	Own Subject Bias Corrected	Both Subjects	Own Subject	Both Subjects	Own Subject	Own Subject Bias Corrected
Grades 5-8 Pooled Estimates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Average Marginal Effect of Baseline Achievement										
1. Linear model										
Baseline math:	0.685***	0.765***	0.952***	0.957***	0.954	0.199***	-	0.015***	-	-
	(0.001)	(0.001)	(0.002)	(0.001)	-	(0.001)	-	(0.002)	-	-
Baseline reading:	0.137***	-	0.006***	-	-	0.601***	0.712***	0.959***	0.970***	0.970
	(0.001)	-	(0.002)	-	-	(0.001)	(0.001)	(0.002)	(0.001)	-
2. Cubic model										
Baseline math:	0.679***	0.757***	0.952***	0.956***	-	0.196***	-	0.013***	-	-
	(0.001)	(0.001)	(0.002)	(0.001)	-	(0.001)	-	(0.002)	-	-
Baseline reading:	0.134***	-	0.005**	-	-	0.595***	0.705***	0.961***	0.969***	-
	(0.001)	-	(0.002)	-	-	(0.001)	(0.001)	(0.002)	(0.001)	-
P-value for F-test of all Baseline Quadratic, Cubic Terms, for All Subjects	0.000	0.000	0.000	0.000	-	0.000	0.000	0.000	0.000	-
P-value for F-test of Other Subject Baseline Score Terms	0.000	-	0.002	-	-	0.000	-	0.000	-	-
N (Grades 5-8)	1103000	1103000	1103000	1103000	1103000	1102699	1102699	1102699	1102699	1102699

Note: All models also include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. All controls, except for lagged scores, are interacted with grade level. As described in the text, the bias correction method in Columns (5) and (10) use the sensitivity of the baseline coefficient to inclusion of student race/ethnicity, gender, and indicators for economic disadvantage, English Language learner status, and special education status to correct for correlation with unmeasured student growth. The specifications in Columns (5) and (10) use only the lag of outcome subject score. Standard errors are in parentheses: * p < 0.05, ** p < 0.01, *** p < 0.001.

Figure 2. Model Predictions



Note: Here the cubic VAM and cubic ACVAM include the cubic of the lag score of both subjects, as well as controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. The lagged subject effects are pooled across grades 5-8, but otherwise covariates are fully interacted with grade level. Model predictions are based solely on the lagged outcome subject terms, with all other covariates fixed at zero.

Table 4. Direct Effect of Neighborhood Income

Grades 5-8 Pooled Estimates	Math		Reading	
	Conventional VAM	ACVAM	Conventional VAM	ACVAM
Neighborhood Median Income Decile				
1 (ref.)	-	-	-	-
	-	-	-	-
2	0.009**	0.004	0.006	0.003
	(0.003)	(0.003)	(0.003)	(0.004)
3	0.003	0.001	0.009*	0.000
	(0.003)	(0.003)	(0.003)	(0.004)
4	0.012***	0.007*	0.011**	0.003
	(0.003)	(0.003)	(0.003)	(0.004)
5	0.008**	0.002	0.015***	0.004
	(0.003)	(0.003)	(0.003)	(0.004)
6	0.012***	0.005	0.014***	0.004
	(0.003)	(0.003)	(0.003)	(0.004)
7	0.017***	0.008*	0.018***	0.004
	(0.003)	(0.003)	(0.003)	(0.004)
8	0.020***	0.008*	0.024***	0.006
	(0.003)	(0.003)	(0.004)	(0.004)
9	0.029***	0.011**	0.028***	0.006
	(0.003)	(0.003)	(0.004)	(0.004)
10	0.044***	0.013***	0.031***	0.005
	(0.004)	(0.004)	(0.004)	(0.004)
P-value for F-test	0.000	0.009	0.000	0.821
N (Grades 5-8)	662496	662496	662086	662086

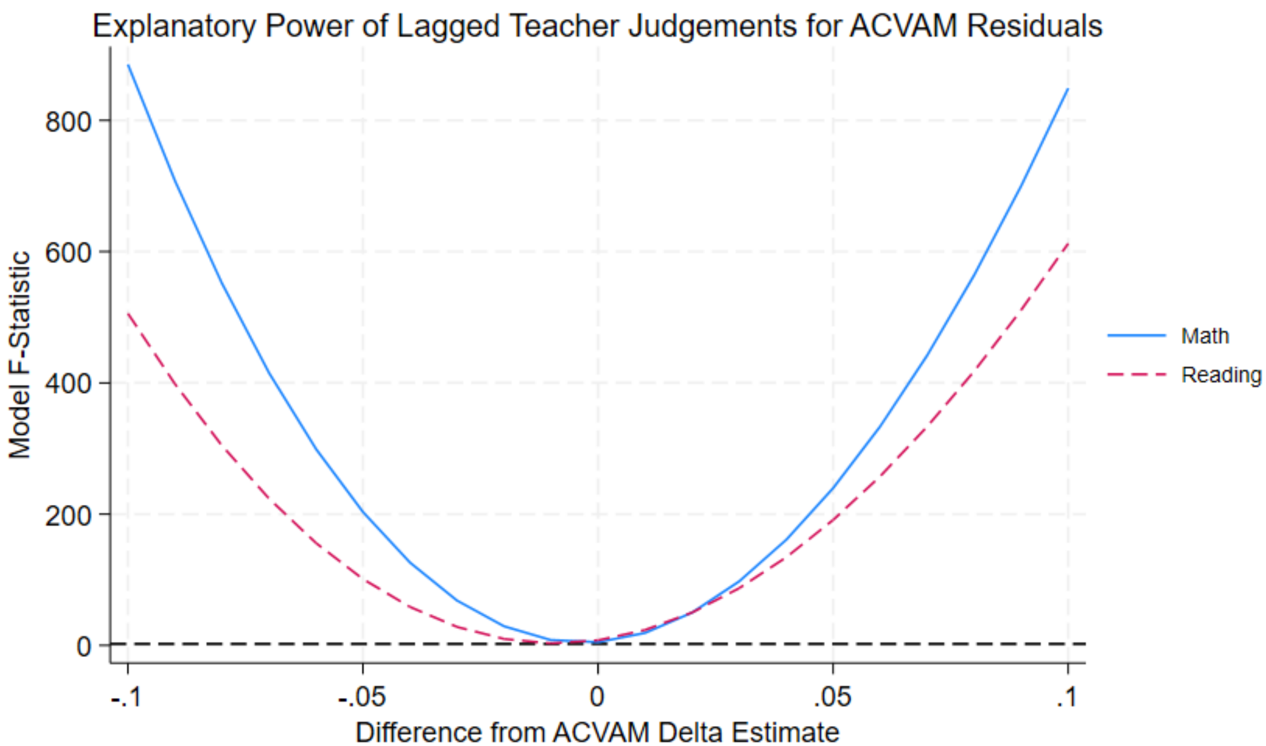
Note: The VAM and ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. All covariates are interacted with grade level except for the current income deciles. The p-value is reported for the joint F-test that all income decile coefficients are zero. Standard errors are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5. Role of Prior Teacher Judgement of Student Mastery

	Math		Reading	
Grades 5-8 Pooled Estimates	Conventional VAM	ACVAM	Conventional VAM	ACVAM
Prior Teacher Judgement				
Insufficient Mastery	-0.319***	0.026***	-0.304***	0.036***
	(0.005)	(0.006)	(0.005)	(0.006)
Inconsistent Mastery	-0.182***	0.011***	-0.171***	0.008**
	(0.002)	(0.003)	(0.002)	(0.003)
Consistent Mastery (ref)	-	-	-	-
	-	-	-	-
Consistently Superior	0.209***	0.002	0.185***	-0.009**
	(0.002)	(0.003)	(0.002)	(0.003)
None of the Above	-0.173***	0.007	-0.198***	0.045
	(0.036)	(0.041)	(0.044)	(0.049)
P-value for F-test	0.000	0.000	0.000	0.000
N (Grades 5-8)	490255	490255	490015	490015

Note: The state stopped collecting teacher judgements of student mastery after 2013, so these data are limited to the 2012-2014 8^{th} grade cohorts that have a full history of lagged judgements. The VAM and ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. All covariates are interacted with grade level except for the lagged judgements. The p-value is reported for the joint F-test that all lagged judgement coefficients are zero. Standard errors are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 3. Assessing Potential Bias in ACVAM Estimate of Delta Using Teacher Judgements



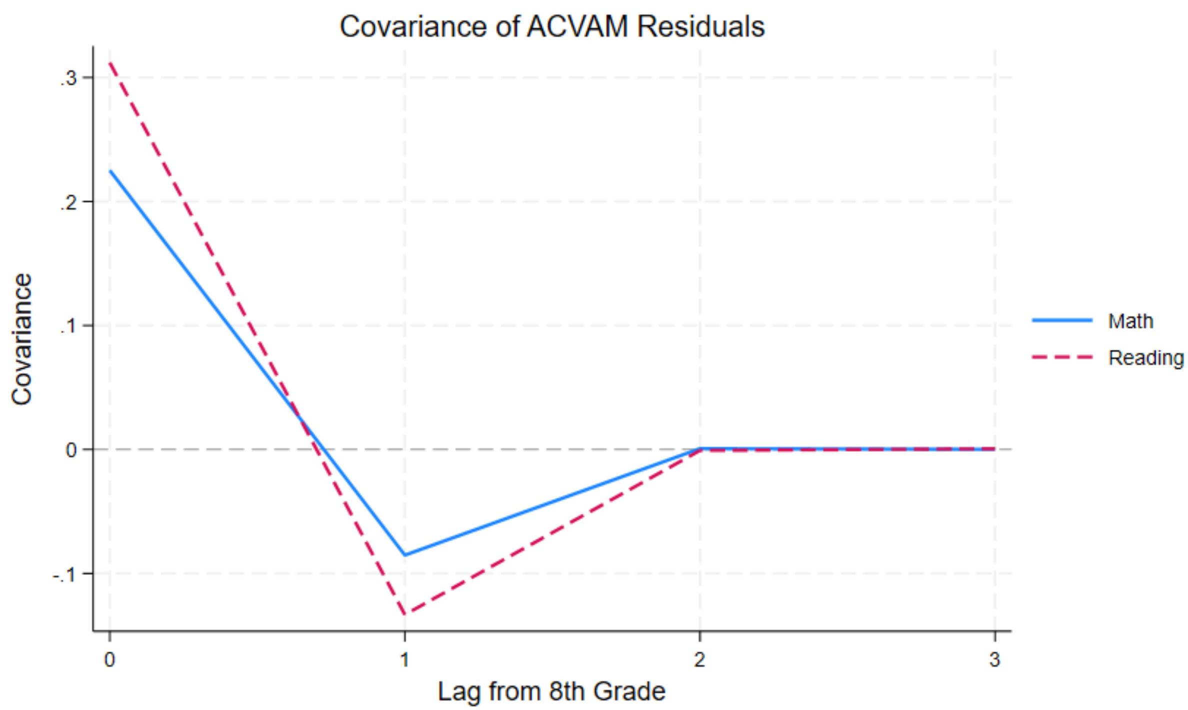
Note: The state stopped collecting teacher judgements of student mastery after 2013, so these data are limited to the 2012-2014 8th grade cohorts that have a full history of lagged judgements. ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. ACVAM residuals are generated grade by grade, for grades 5-8, as follows: 1) run the ACVAM model on observed scores 2) generate quasi-differences that adjust the lagged outcome subject linear coefficient and then fix all lagged score coefficients 3) regress the quasi-differences on full controls to generate residuals. Then ACVAM residuals are regressed on full controls interacted with grade, along with lagged judgements. Finally, we run a joint test that all lagged judgement coefficients are zero and plot the resulting F statistic. The horizontal line at 2.37 represents the F statistic critical value for significance at 5% for each subject.

Table 6. State Space Model Parameter Estimates Using ACVAM

Component:		Math	Reading
Delta (AME of lagged score by grade, avg. across grades)	$\hat{\delta}$	0.9543	0.9617
Measuring Achievement:			
Measured score variance	$Var(Y_{it})$	0.8832	0.8794
Measured error variance	$Var(\eta_{it})$	0.0983	0.1392
The gap between true and predictable baseline achievement:			
True score variance	$Var(\mu_{it})$	0.7605	0.7315
Kalman prediction variance	$Var(\hat{Y}_{it})$	0.6335	0.6142
Sources of persistent innovations:			
Student-level heterogeneity (unobserved)	$Var(\tilde{\alpha}_i)$	0.0008	0.0008
Student-level heterogeneity (assoc. with demographics/program use)	$E[Var(W\hat{\gamma}\|grade)]$	0.0007	0.0010
Persistent teacher effects	$Var(\theta_j)$	0.0135	0.0075
Other persistent innovations	$Var(v_{it})$	0.0403	0.0363
Transitory teacher effects:			
Transitory teacher effects	$Var(\psi_j)$	0.0244	0.0087
Correlation between persistent, transitory teacher effects	$Corr(\theta_j, \psi_j)$	-0.0915	-0.2899
Overall teacher effect	$Var(\theta_j + \psi_j)$	0.0346	0.0115
Implied teacher fadeout	$Cov(\theta_j, \theta_j + \psi_j)/\,Var(\theta_j + \psi_j)$	0.3416	0.4474
Correlation between total and persistent teacher effects	$Corr(\theta_j + \psi_j, \theta_j)$	0.5472	0.5546

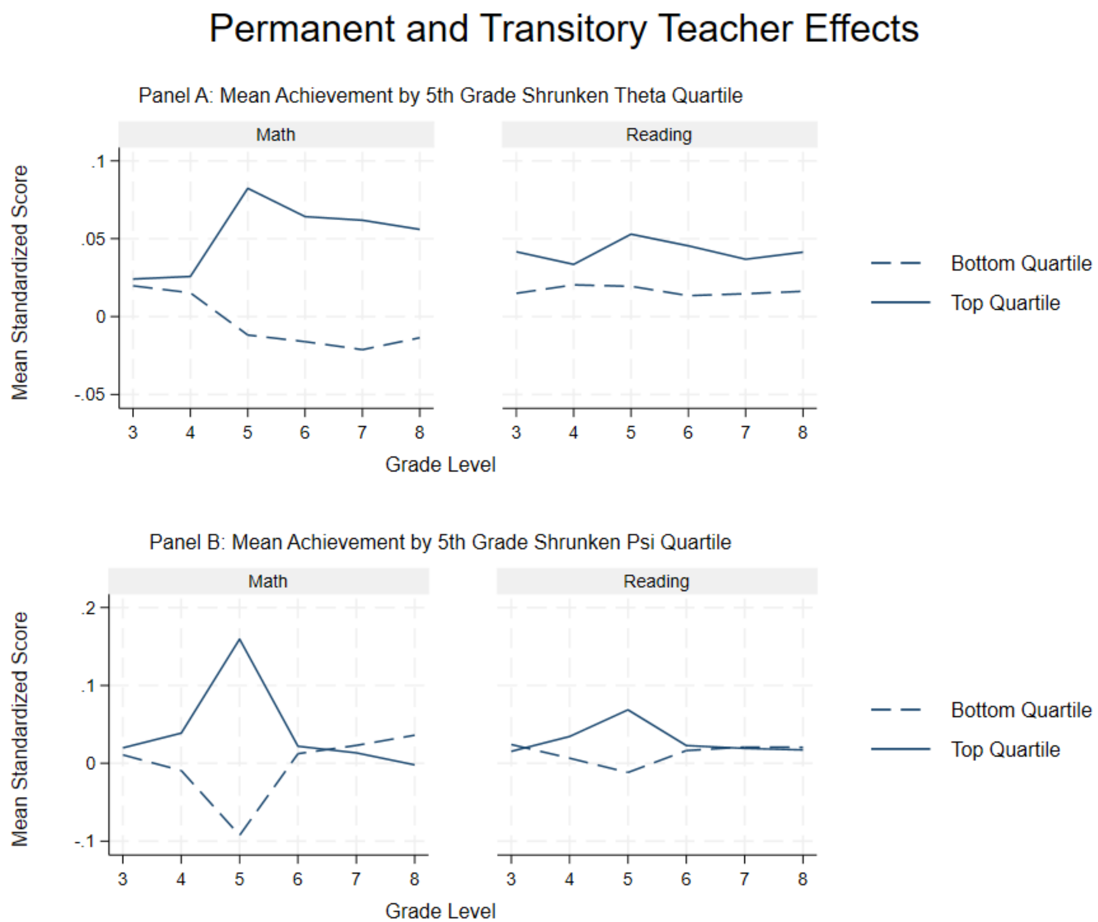
Note: ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. Kalman predictions are from models that are quadratic in the prior prediction for each subject and include the same demographic and teacher controls as ACVAM. The current and lagged teacher effects in the Kalman regressions are not included in the predictions; instead, we add to the predictions contemporaneous teacher permanent effects estimated with ACVAM. We run ACVAM grade by grade, for grades 5-8, to produce residuals and $W\hat{\gamma}$. All parameter estimates reported in the table are estimated across grades 5-8 except for the variance in the permanent and transitory teacher effects, for which variance is computed across grades 5-7, the grades in which both parameters are available. The correlation in the teacher permanent and transitory effects is likewise computed over grades 5-7. Estimates of $Var(\tilde{\alpha}_i)$, $Var(\eta_{it})$, and $Var(v_{it})$ come from an MA(1) model run on the ACVAM residuals. The other parameter estimates come from Bayesian bootstrap results from ACVAM.

Figure 4. Persistence of Unobserved Student Effects



Note: ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. ACVAM is run grade by grade, for grades 5-8, to generate residuals used to compute the covariances shown here.

Figure 5. Further Testing the State-Space Model



Note: ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. ACVAM is run grade by grade, for grades 5-6, to generate estimates of theta and psi in grade 5. Estimates of these teacher effects for any given student are estimated while excluding that student's cohort.

Table 7. Teacher Sorting on Baseline Achievement

	Math	Reading
SD of Teacher Effects		
Baseline Achievement	0.4678	0.4560
	(0.0030)	(0.0029)
Kalman Prediction for Baseline Achievement	0.4100	0.4117
	(0.0026)	(0.0026)
Kalman Residual for Baseline Achievement	0.1314	0.0802
	(0.0011)	(0.0009)
N (Grades 5-8)	1090684	1090768

Note: The results are estimated with hierarchical linear models that include current teacher random effects only. The models are run on grades 5-8, the grades for which baseline Kalman predictions and residuals are available. Standard errors are in parentheses. Kalman predictions and residuals are from models that are quadratic in the prior prediction for each subject, and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. The current and lagged teacher effects in the Kalman regressions are not included in the predictions; instead, we add to the predictions contemporaneous teacher permanent effects estimated with cubic ACVAM.

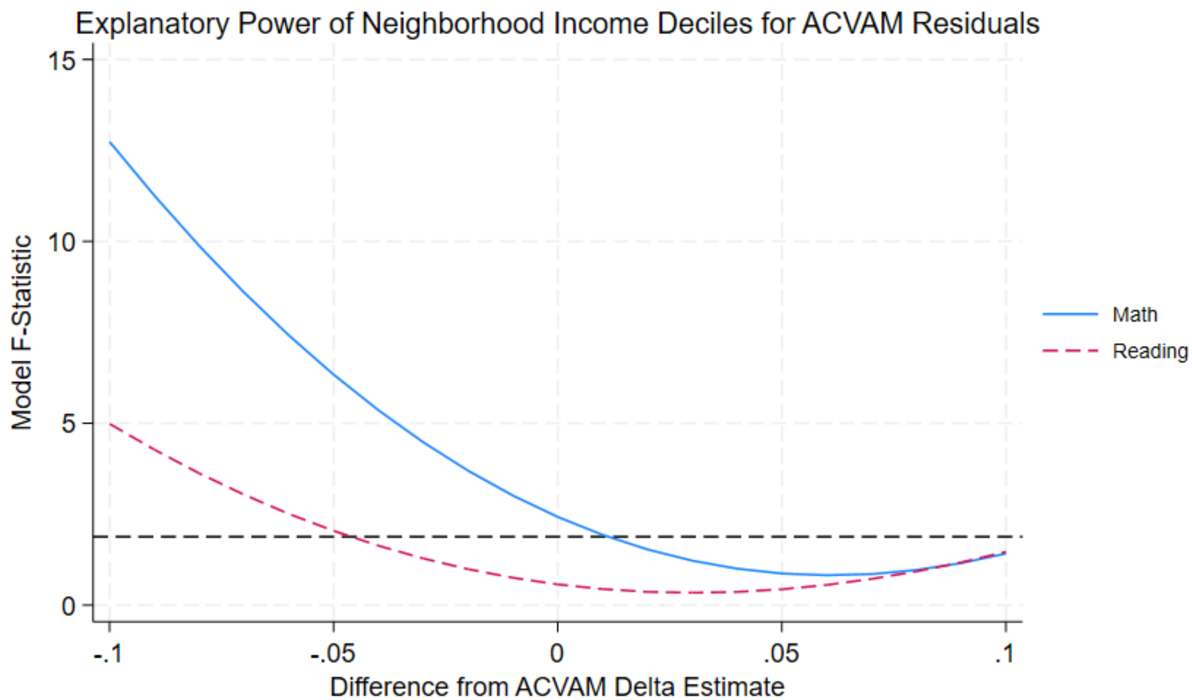
Table 8. Decomposing Forecast Bias in Conventional VAM

	Math				Reading			
	Teacher Theta + Psi		School Theta + Psi		Teacher Theta + Psi		School Theta + Psi	
	Total (1)	Within School (2)	Total (3)	Within District (4)	Total (5)	Within School (6)	Total (7)	Within District (8)
Var(ACVAM)	0.0345	0.0215	0.0181	0.0155	0.0115	0.0051	0.0068	0.0046
Var(VAM)	0.0326	0.0219	0.0168	0.0150	0.0095	0.0059	0.0054	0.0045
Cov(ACVAM, VAM)	0.0297	0.0207	0.0139	0.0126	0.0076	0.0053	0.0031	0.0029
Implied Corr(ACVAM, VAM)	0.8842	0.9550	0.7996	0.8259	0.7242	0.9560	0.5114	0.6324
Implied moments:								
Forecast bias coef	0.9106	0.9459	0.8282	0.8371	0.7965	0.8908	0.5741	0.6455
Variance of bias (VAM – ACVAM)	0.0078	0.0020	0.0070	0.0053	0.0059	0.0005	0.0060	0.0033
Corr(ACVAM, bias)	-0.2976	-0.1189	-0.3661	-0.3159	-0.4783	0.0818	-0.5796	-0.4481
SD(ACVAM)	0.1858	0.1467	0.1344	0.1243	0.1072	0.0717	0.0822	0.0681
SD of bias	0.0883	0.0442	0.0838	0.0729	0.0765	0.0227	0.0772	0.0578

Note: Each model specification includes the cubic of lagged scores from both subjects as well as controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, and current subject teacher or school. The conventional VAM model for teacher effects adds peer means of baseline scores for both subjects, demographics, and program participation. The ACVAM model, for both teacher effects and school effects, adds lagged subject teacher or school. Variance and covariance estimates are based on the Bayesian bootstrap. All results are computed over grades 5-7.

Appendix

Figure A1. Assessing Potential Bias in ACVAM Estimate of Delta Using Neighborhood Income



Note: ACVAM specifications are cubic in prior achievement of both subjects and include controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. ACVAM residuals are generated grade by grade, for grades 5-8, as follows: 1) run the ACVAM model on observed scores 2) generate quasi-differences that adjust the lagged outcome subject linear coefficient and then fix all lagged score coefficients 3) regress the quasi-differences on full controls to generate residuals. Then ACVAM residuals are regressed on full controls interacted with grade, along with neighborhood median income deciles. Finally, we run a joint test that all income decile coefficients are zero and plot the resulting F statistic. The horizontal line at 1.88 represents the F statistic critical value for significance at 5% for each subject.

Table A1. Decomposing Forecast Bias in Conventional VAM with Current and Lagged Teacher

	Math			Reading		
	Theta + Psi	Theta	Psi	Theta + Psi	Theta	Psi
Var(ACVAM)	0.0346	0.0135	0.0244	0.0115	0.0075	0.0087
Var(VAM)	0.0350	0.0263	0.0306	0.0113	0.0200	0.0210
Cov(ACVAM, VAM)	0.0338	0.0155	0.0256	0.0100	0.0073	0.0117
Implied Corr(ACVAM, VAM)	0.9728	0.8239	0.9367	0.8800	0.5956	0.8659
Implied moments:						
Forecast bias coef	0.9663	0.5897	0.8361	0.8889	0.3643	0.5573
Variance of bias (VAM – ACVAM)	0.0019	0.0087	0.0038	0.0027	0.0129	0.0063
Corr(ACVAM, bias)	-0.0881	0.1874	0.1250	-0.2643	-0.0200	0.4062
SD(ACVAM)	0.1859	0.1160	0.1563	0.1073	0.0865	0.0933
SD of bias	0.0435	0.0935	0.0618	0.0523	0.1137	0.0793

Note: Each model specification includes the cubic of lagged scores from both subjects as well as controls for gender, economic disadvantage, disability, English language learner status, race/ethnicity, year, current subject teacher, and lagged subject teacher. Variance and covariance estimates are based on the Bayesian bootstrap. All results are computed over grades 5-7, the grades in which all parameter estimates are available for each model.

Appendix A. Estimating $Var(\tilde{\alpha}_i)$ if assignment to teachers depends on student growth.

Even at the true value of δ, we may underestimate $Var(\tilde{\alpha}_i)$ if assignment to teachers depends on student growth (since some of that variance will be absorbed into the teacher effects). In this appendix we derive an adjustment based on the within and between teacher variance in $W_i\gamma$ (the observable component of α_i) that yields a consistent estimate of $Var(\tilde{\alpha}_i)$ under the same assumption as that needed for the Oster bias correction (that the degree of selection on the observed ($W_i\gamma$) and unobserved ($\tilde{\alpha}_i$) components is equivalent).

Suppose the following two assumptions hold (same as those needed for Oster bias correction):

1. Let $\alpha_i = W_i\gamma + \tilde{\alpha}_i$ with $W_i\gamma$ independent of $\tilde{\alpha}_i$, $Var(W_i\gamma) = \sigma^2_{W_i\gamma}$ and $Var(\tilde{\alpha}_i) = \sigma^2_{\tilde{\alpha}}$.
2. Students are sorted to teachers based on $z_i = \alpha_i + u_i$, where u_i is independent of α_i and $Var(u_i) = \sigma^2_u$.

Note that assumption 2 implies that sorting on the unobservable component of student growth ($\tilde{\alpha}_i$) is as strong as sorting on the observable component ($W_i\gamma$). As Oster (2019) notes, this is likely to provide an upper bound on the amount of bias since sorting is, if anything, likely to be stronger on easily observable correlates of student growth in $W_i\gamma$.

Our goal is to estimate the unconditional variance of $\tilde{\alpha}_i$, $\sigma^2_{\tilde{\alpha}}$, based on our estimate of the variance in $\tilde{\alpha}_i$ that is conditional on the teacher assignment. If students are sorted to teachers based on z_i, then conditioning on teacher assignment is equivalent to conditioning on z_i, so that:

Equation A1: $Var(\tilde{\alpha}_i|teacher) = Var(\tilde{\alpha}_i|z_i) = \sigma^2_{\tilde{\alpha}} \cdot \left(1 - R^2_{\tilde{\alpha}_i|z_i}\right)$

Where $R^2_{\tilde{\alpha}_i|z_i}$ is the R-squared from a regression of $\tilde{\alpha}_i$ on z_i, which under our assumptions is:

Equation A2: $R^2_{\tilde{\alpha}_i|z_i} = \frac{\sigma^2_{\tilde{\alpha}}}{\sigma^2_{W_i\gamma} + \sigma^2_{\tilde{\alpha}} + \sigma^2_u}$

Similarly, the R-squared from a regression of $W_i\gamma$ on teacher effects is equivalent to the R-squared from a regression of $W_i\gamma$ on z_i, which under our assumptions is:

Equation A3: $R^2_{W_i\gamma|z_i} = \frac{\sigma^2_{W_i\gamma}}{\sigma^2_{W_i\gamma} + \sigma^2_{\tilde{\alpha}} + \sigma^2_u}$

Taking the ratio of the R-squared values in Equations A2 and A3 yields:

Equation A4: $\frac{R^2_{\tilde{\alpha}_i|z_i}}{R^2_{W_i\gamma|z_i}} = \frac{\sigma^2_{\tilde{\alpha}}}{\sigma^2_{W_i\gamma}}$, or equivalently $R^2_{\tilde{\alpha}_i|z_i} = \frac{\sigma^2_{\tilde{\alpha}} \cdot R^2_{W_i\gamma|z_i}}{\sigma^2_{W_i\gamma}}$

Plugging Equation A4 into Equation A1 yields:

Equation A5: $Var(\tilde{\alpha}_i|z_i) = \sigma^2_{\tilde{\alpha}} \cdot \left(1 - \frac{\sigma^2_{\tilde{\alpha}} \cdot R^2_{W_i\gamma|z_i}}{\sigma^2_{W_i\gamma}}\right)$

Or equivalently:

Equation A6: $Var(\tilde{\alpha}_i|z_i) - \sigma_{\tilde{\alpha}}^2 + \left(\sigma_{\tilde{\alpha}}^2\right)^2 \left(\frac{R_{W_i\gamma|z_i}^2}{\sigma_{W_i\gamma}^2}\right) = 0$

We use the quadratic formula to solve Equation A6 for $\sigma_{\tilde{\alpha}}^2$ as a function of our estimates of $Var(\tilde{\alpha}_i|z_i)$, $R_{W_i\gamma|z_i}^2$, and $\sigma_{W_i\gamma}^2$ (selecting the smaller root).

Appendix Table 1 reports the results of applying this correction. We use estimates of $Var(\tilde{\alpha}_i|z_i)$ and $W_i\gamma$ from the bias-corrected AC-VAM model (Table 3, columns 5 and 10), and an estimate of $R_{W_i\gamma|z_i}^2$ using the adjusted R-squared from a regression of $W_i\gamma$ on teacher fixed effects. For both math and reading, the estimates of $\sigma_{\tilde{\alpha}}^2$ are only slightly larger than the estimates of $Var(\tilde{\alpha}_i|z_i)$ that we report in the main tables.

Appendix Table 1. Estimates of $Var(\alpha_i)$ that correct for sorting to teachers on α_i

	Math	Reading
Estimates from bias-corrected ACVAM:		
$Var(\tilde{\alpha}_i|z_i)$	0.000744	0.000257
$\sigma_{W_i\gamma}^2$	0.000722	0.001456
$R_{W_i\gamma|z_i}^2$	0.1844	0.3303
Implied estimate of $\sigma_{\tilde{\alpha}}^2$	0.000999	0.000274