Why Does Value-Added Work?
Implications of a Dynamic Model of Student Achievement

Douglas O. Staiger, Dartmouth College & NBER
Thomas J. Kane, Harvard University & NBER
Brian D. Johnson, Harvard University

July 20, 2024

Paper to be presented at NBER Summer Institute

Economics of Education Meeting

Preliminary Draft

Please do not cite without permission

Acknowledgements: We thank the North Carolina Education Data Research Center (NCERDC) for providing access to the NC data. We thank Eric Taylor and Mary Laski for sharing their code in building our analysis file.

Disclosures: Kane served as an expert witness in *Vergara v. State of California* and two other cases involving the use of value-added measures in teacher evaluations.

Abstract

To guide value-added modelling choices, we propose a state space model of student knowledge accumulation, in which test scores are an imperfect measure of student knowledge, and students receive temporary and persistent shocks to their stock of knowledge. The model clarifies that there are four sources of potential selection in value-added estimation: heterogeneity in student growth, measurement error in baseline achievement, transitory teacher effects, and private information about students' current knowledge. Following six cohorts of students in North Carolina between 3rd and 8th grade, we investigate all four sources of bias. We find little evidence of heterogeneity in student growth. Rather, the primary challenge in value-added is finding a valid measure of baseline knowledge against which to measure differences. The model points to two alternatives to conventional VAM models: the Kalman filter (which efficiently summarizes students' prior data) and instrumenting for baseline achievement with twice lagged scores. The IV estimator is ideal when achievement scores are not observed at the time of assignment (as is the case in most states) and when teachers have private information about students' current state of knowledge not reflected in test scores. The state-space model has implications for the sources of achievement inequality and for other analyses of student-level panel achievement data (for instance, calling into question the use of student fixed effects and student trends.)

Introduction

For more than five decades, economists have been using "value-added" models (VAM) to estimate the efficacy of schools, teachers and other interventions (Hanushek (1971), Murnane (1975), Hanushek (1979)). Conventional VAM models typically include a single year of baseline achievement and student covariates to control for prior inputs. Despite the strong assumptions and the potential for selection, at least eleven studies over the past sixteen years have shown that conventional VAM models yield forecast unbiased estimates of teacher and school effects. Did researchers happen to choose settings with little selection on unobservables? Or do the results reveal something fundamental about how student achievement evolves over time?

To investigate, we propose a simple state-space model, in which test scores are a noisy measure of an underlying state variable (knowledge). Teachers and other interventions have persistent effects on student knowledge, which are passed on to the next period, as well as temporary effects on measured achievement, which are not. In addition, we allow each student to have their own fixed rate of growth.

The model identifies four sources of potential bias in value-added models: heterogeneity in student growth, measurement error in baseline achievement, transitory teacher effects on achievement measures (which could be used for selection) and private information about students' current true level of achievement. The latter three are all special cases of mismeasurement of students' baseline knowledge.

Following six cohorts of NC students from third through eighth grade, we find little evidence of student-level heterogeneity in growth across different grades. Similarly, we see little evidence that students from high and low-income neighborhoods have differing levels of growth (after conditioning on teacher effects.) Both results imply that that there is little unobserved heterogeneity in student growth which teachers, schools or parents could use for selection.

Rather, the primary challenge in value-added estimation is having a valid measure of baseline achievement against which to measure differences. We evaluate two alternatives to conventional value-added: the Kalman filter, which summarizes historical data, and an IV estimator (using twice-lagged achievement to instrument for baseline achievement).

First, we show that the recursive Kalman filter outperforms the single year of baseline achievement and performs nearly as well as a fully flexible model in summarizing prior test scores and teacher effects. The results imply that the state space model is a good approximation of the time series properties of student achievement.

Second, the most vexing source of selection in the state space model is private knowledge of a students' current state not reflected in test scores. Using unique data on teachers' subjective assessments of students' mastery of the North Carolina state standards, we

find that prior grade teachers' assessments are indeed predictive of growth even after conditioning on the Kalman filter. Fortunately, the state space model also points to a solution: instrumenting for baseline achievement using twice-lagged achievement. We provide evidence that the IV estimates of teacher effects (using twice lagged achievement) are not sensitive to prior grade teachers' ratings.

Third, consistent with the prior literature, we find that teachers have both transitory and permanent effects on student achievement. Transitory effects have been underappreciated as a potential source of selection, yet there are substantial differences in transitory effects between schools and teachers. The model also points to a straightforward solution: including fixed effects for current as well as prior teachers.

Fourth, while the conventional VAM estimate is biased and the IV estimator and Kalman estimators are each unbiased (albeit under different conditions), the bias in each appears to be small relative to the variance in teacher effects. The correlations in teacher effects from the three estimation methods (conventional, IV and Kalman) are all greater than .81. Thus, the bias in conventional value-added models may be sufficiently small to be undetectable under the standard of forecast unbiasedness, which is a more lenient standard.

Finally, our findings imply that, aside from differing starting points, the most important source of achievement inequality is the quality of educational interventions that students receive. Using our IV models, students from higher income neighborhoods tend to have teachers with larger persistent impacts on achievement, while students from lower income neighborhoods tend to have teachers with larger transitory impacts on achievement. Although neighborhood income has little direct relationship to achievement growth conditional on teacher effects, there are sizable differences in school and teacher quality for students in high and low income neighborhoods.

Value-added modelling has the potential for much broader use than measuring the effects of teachers and schools. The state space model implies diagnostics which could be used to gauge the presence of student-level heterogeneity in growth and identifies the conditions under which the Kalman filter and IV strategies could be relied upon to generate valid program impacts. Given the ubiquity of student-level panel data on student achievement in state and local education agencies, value-added methods could allow for much more widespread testing of a variety of interventions in education, such as curricula. The state-space model has strong implications, not just for value-added modelling, but for other analyses of student-level panel data (for example, calling into question models using student fixed effects and student achievement trends.)

Literature Review

In reviewing the prior literature, we focus on models which combine three elements common to many VAM applications: additive separability, a single year lagged measure of

achievement in the same subject (either as a linear or cubic function of a standardized score) and indicators for student demographics or program participation (race/ethnicity, free lunch status).[1] While VAM models have been used most often to estimate teacher or school impacts, 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) and Henry et al. (2014)).

Before arriving at the current convention, researchers experimented with other variants of VAM models. For instance, with the introduction of large multi-year student-level panel data sets in the mid-2000's, researchers began including student fixed effects to control for selection on unmeasured student characteristics, while estimating teacher or school effects (e.g. McCaffrey et al. (2009)). Such models became rare after Rothstein (2010) noted that they likely violated strict-exogeneity requirements. They also perform poorly in validation studies (Kane and Staiger (2008), Kane, McCaffrey, Miller and Staiger (2013)).

Others began using the mean baseline characteristics of students to control for classroom peer effects (see Ehlert et al. (2014)). As part of the Measures of Effective Teaching project, Kane, McCaffrey, Miller and Staiger (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 (without multiple years and teacher fixed effects) is "over-controlling" by removing true differences in teacher efficacy.

Although the conventional value-added model only includes one lag, Rothstein (2010) and Ehlert et al. (2014) find that they could not validly exclude two and three-year lagged scores. Rothstein suggested that excluding multiple lags of achievement and teacher assignments could lead to bias. However, Rothstein also found that their inclusion did not lead to any meaningful differences in estimated teacher effects in North Carolina.

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 some 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 the coefficient on the estimate in

[1] Often analysts also include lagged values for student test scores in other subjects.

predicting the outcome of the experiment (empirical Bayes adjusted to account for estimation error). Forecast unbiasedness (a coefficient of one on the VAM estimate) implies that the estimates are "right on average." However, as discussed in Angrist et al. (2017), a set of estimates can be forecast unbiased, yet still contain estimates of individual school or teacher effects which are biased—i.e. differences which are larger than would plausibly be driven by sampling error.

Three studies (Kane and Staiger (2008), Kane et al. (2013) and Glazerman and Protik (2015)) estimated teacher value-added with historical data, randomized teachers and 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 lagged score model and the less common "gain score" model (essentially constraining the coefficient on lagged achievement to be equal to one) are forecast unbiased.[2] The studies involved teachers across 7-14 large urban districts.[3]

In both models including student fixed effects, Kane and Staiger (2008) found forecast bias coefficients near two. As described by Meghir and Rivkin (2011), student fixed effects essentially borrow from a student's future outcomes to estimate the fixed effect. 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 are 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, Hull, Pathak and Walters (2017), Angrist, Hull, Pathak and Walters (2024)). With one exception, all three studies could not reject that value-added estimates of school effects were forecast unbiased estimates of school effects. 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 very same data—test scores, race, gender, program participation, etc.—used to make teacher assignments within schools, families presumably sort across schools for many other reasons, such as the parents' ability/willingness to pay for school quality. The one exception was for high school impacts on SAT math scores in New York, where the coefficient (.78) discernibly different from 1, but still large.

Because their lottery-based estimates are over-identified, Angrist et al. can go beyond "forecast unbiasedness" and test whether the value-added estimates are consistent with multiple instruments. They conclude that while value-added may be forecast unbiased, the estimates for individual schools are subject to bias, with differences larger than would have

[2] 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.

[3] 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.

been expected due to sampling variation alone. Thus, conventional value-added models may be biased, but still meet the lower standard of forecast unbiasedness (i.e. being "right on average.")

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, Friedman and Rockoff (2014) generated an estimate of forecast bias between .91 and .99. The other two studies essentially replicate Chetty, Friedman and Rockoff's methodology in two other, quite different settings: Los Angeles Unified School District (Bacher-Hicks, Kane and Staiger (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, Clark and Lee (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, Bau, Das and Khwaja (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) estimate for middle schools, Angrist et al. (2024) estimate 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, using both random assignment and quasi-experimental methods to test their validity. Despite these findings, there is still no consensus on the conditions under which value-added methods should be expected to yield unbiased estimates.

There are three primary reasons for the lack of consensus: First, opportunities for model validation have been rare. The eleven studies represent a small share of studies using value-added estimation. Second, the existing validation tests based on forecast unbiasedness are often underpowered to detect specific model misspecifications which could lead to small amounts of bias— even with forecast unbiasedness. But the third and most important reason is the absence of any agreed-upon statistical model describing the sources of variation in student achievement over time. In the absence of such a model, any claims of validity are therefore

contingent on a particular use case (e.g. estimating teacher or school effects on math achievement) and in a particular setting (such as Charlotte or Boston or New York.) In the next section, we present a framework for evaluating different sources of bias.

Model and Estimation

In this section we describe a statistical model of test scores and student learning based on a standard value-added structure (and similar to that used in previous work such as Jacob, Lefgren & Sims, 2007). 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) that accumulates persistent innovations over time. We allow for student and teacher components in both the transitory noise in test scores and the persistent innovations to knowledge. We then use the model to motivate three 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 is similar but uses an earlier test score to instrument for prior year test scores. The final approach uses the recursive Kalman Filter to predict each student's expected baseline score based on the student's history and estimates teacher effects after conditioning on the Kalman filter prediction.

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

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.[4] [5] The first term (θ_{jt}) is the effect of having teacher j in year t (or more generally could represent any intervention j to which student i was assigned in year t). 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

[4] 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.

[5] 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, $\psi_{jt}$and permanent teacher effect, θ_{jt}.

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. Thus, α_i captures persistent differences across years in a student's knowledge growth, while υ_{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. This would result in under-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, free lunch eligibility) in part to account for variation across students in α_i.

Hypothetically, one could estimate models such as equation (1) using dynamic panel methods that allow for a student fixed effect in growth (e.g. Arellano and Bond, 1991), but in practice these methods are poorly identified in short panels with δ near to 1. 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, so tend to underestimate effects of teachers if their impacts on knowledge are persistent (Meghir and Rivkin, 2011). Kane and Staiger (2008) found teacher effect estimates from VAM models 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 students' accumulated knowledge and contain substantial measurement error. We assume that this measurement error is purely transitory and follows a simple structure:

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

The first noise component (φ_{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 learning. 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. 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. 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 $\mu_{it} = y_{it} - \varphi_{jt} - \eta_{it}$, and plugging this into equation (1) yields a familiar VAM-style estimating equation:

(3) $y_{it} = \delta y_{i,t-1} + \left(\theta_{jt} + \varphi_{jt}\right) - \delta\varphi_{j',t-1} + \alpha_i + \upsilon_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right)$,

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

Equation (3) differs from equation (1) in three important ways:

1. First, the current year teacher effect $\left(\theta_{jt} + \varphi_{jt}\right)$ 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 and transitory teacher impact:
   $E\left[\theta_{jt}|\left(\theta_{jt} + \varphi_{jt}\right)\right] = \beta_{fade}\left(\theta_{jt} + \varphi_{jt}\right)$ where $\beta_{fade} = \frac{Cov\left(\theta_{jt},\theta_{jt}+\varphi_{jt}\right)}{Var\left(\theta_{jt}+\varphi_{jt}\right)} = \frac{Var(\theta)+Cov(\theta,\varphi)}{Var(\theta)+Var(\varphi)+2cov(\theta,\varphi)}$
   Moreover, this average fadeout may mask considerable variation across teachers.

2. Second, equation (3) now includes a prior year teacher effect ($-\delta\varphi_{j',t-1}$) 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 $\theta_{jt} + \varphi_{jt}$ and $-\delta\varphi_{j',t-1}$, which along with an estimate of δ identifies both persistent and transitory teacher effects.

3. Finally, even after conditioning on current and prior year teacher, the error in equation (3) still depends on the measurement error in both the 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 expected value of the attenuated coefficient, then the error term will contain $(\delta - \hat{\delta})y_{i,t-1} = (\delta - \hat{\delta})(\mu_{i,t-1} + \varphi_{j',t-1} + \eta_{i,t-1})$ which introduces prior knowledge into the error term. If students were assigned to teachers based on better information about their prior knowledge (e.g. tracking) 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 $\mu_{i,t-1} + \varphi_{j,t-1}$ 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.

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).

3. IV Estimation

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. Therefore, IV estimates of equation (3) using a twice lagged test score

($y_{i,t-2}$) as an instrument for the lagged score will yield unbiased estimates of δ.[6] 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 (3), $\alpha_i + \upsilon_{it} + (\eta_{it} - \delta\eta_{i,t-1})$.

By eliminating attenuation bias, IV estimates of VAM models (controlling for current and lagged teacher effects) are not biased by teacher assignment correlated with prior knowledge ($\mu_{i,t-1}$) since prior knowledge is uncorrelated with the error in equation (3).[7] 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 IV estimates of VAM.

But we are still left with the two remaining potential sources of bias that were also present in OLS estimates of VAM models:

1. Teacher assignment correlated with student growth (α_i) will bias IV VAM estimates.
2. Teacher assignment correlated with prior year measurement error in student scores ($\eta_{i,t-1}$) will bias IV VAM estimates.

This second source of bias is likely to be the primary source of bias in IV VAM estimates. If teacher assignment is determined in part by prior year scores ($y_{i,t-1}$) then it will be correlated with the measurement error in prior year scores. 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.[8] 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 IV VAM would avoid bias by using a baseline test administered to all students after assignment.

Overall, a VAM model that instruments for prior score with an earlier score, and includes current and lagged teacher fixed effects, and perhaps some student covariates (to account for α_i) faces only two sources of potential bias: (1) sorting of students to teachers based on student-specific growth (α_i) that is not captured by covariates, and (2) sorting of students to teachers based on the measurement error in the prior-year score ($\eta_{i,t-1}$). This second 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 OLS VAM estimates, assignment to teachers based on a student's prior knowledge does not bias IV VAM estimates.

---

[6] more generally, any score from time t-1 or before with independent measurement error will be a valid instrument.

[7] Prior knowledge is correlated with the error through α_i, but we consider the bias arising from α_i separately.

[8] 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.

4. Kalman Filter Estimates

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

$$(4)\ y_{it} = \delta\mu_{i,t-1} + \left(\theta_{jt} + \varphi_{jt}\right) + \ \alpha_i + v_{it} + \eta_{it}$$

We cannot estimate equation (4) directly because a student's prior knowledge ($\mu_{i,t-1}$) is unknown.

However, suppose we formed an unbiased prediction of $\mu_{i,t-1}$ using all of the information available on the student up through and including the information in t-1. Call this prediction $\hat{\mu}_{i,t-1|t-1}$. Then consider the following equation substituting $\hat{\mu}_{i,t-1|t-1}$ for $\mu_{i,t-1}$ in equation (4):

$$(5)\ y_{it} = \delta\hat{\mu}_{i,t-1|t-1} + \left(\theta_{jt} + \varphi_{jt}\right) + \ \alpha_i + v_{it} + \eta_{it} + \delta\left(\mu_{i,t-1} - \hat{\mu}_{i,t-1|t-1}\right)$$

Equation (5) is analogous to a standard OLS VAM model except instead of conditioning on prior score, we condition on an optimal prediction of prior knowledge given all the information available. In this model, $\delta\hat{\mu}_{i,t-1|t-1}$ 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 (5) arises from the fact that our 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 derived from equation (5) will be biased. Thus, it is particularly important to incorporate as much information as possible into $\hat{\mu}_{i,t-1|t-1}$.

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 information available 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:

$$(6)\ \hat{\mu}_{i,t|t} = \hat{\mu}_{i,t|t-1} + K_t\left(\hat{u}_{i,t|t-1}\right) + \delta\theta_{j,t}$$

Equation (6) 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 $y_{i,t}$.

Plugging equation (6) into equation (5) yields the final Kalman filter estimating equation:

$$(7)\ y_{it} = \delta\left(\hat{\mu}_{i,t-1|t-2}\right) + \delta K_t\left(\hat{u}_{i,t-1|t-2}\right) + \delta\theta_{j',t-1} + \left(\theta_{jt} + \varphi_{jt}\right) + \alpha_i + \upsilon_{it} + \eta_{it} + \delta\left(\mu_{i,t-1} - \hat{\mu}_{i,t-1|t-1}\right)$$

Equation (7) is a vam specification that replaces the lagged score with three terms: The predicted test score in t-1, the prediction residual from t-1, and the persistent 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 (7) also identifies effects of the current and lagged teacher (as in the VAM and IV models). However, the lagged teacher effect ($\delta\theta_{j',t-1}$) now has a different interpretation: it captures the effect of the prior year teacher that persists into the current year.

We estimate equation (7) sequentially in each grade to obtain estimates of δ, K_t, and current and lagged teacher fixed effects. We then use these estimates and the Kalman filter equation (6) to estimate $\hat{\mu}_{i,t-1|t-2}$ and $\hat{u}_{i,t-1|t-2}$ for the next grade.[9] In forming predictions using equation (6), we use empirical Bayes to form best linear unbiased predictions of the prior year's teacher persistent effect ($\theta_{j',t-1}$), e.g. we apply shrinkage to the estimated fixed effects (Kane and Staiger, 2008).

Overall, the Kalman filter approach is an alternative to OLS VAM that attempts to minimize the bias arising from assignment to teachers based on better information about student's prior knowledge. It does so by more carefully controlling for the full history of available information. However, whether this eliminates bias in the estimation of teacher effects will depend on how accurate the Kalman predictions are, whether parents and administrators have access to better information, and whether this information is used for student assignment to teachers.

5. Allowing for non-linearity

VAM models commonly find non-linearity in the relationship between end-of-year scores and prior year scores, with a flatter relationship in the tails. Therefore, we estimate versions of the OLS and IV VAM specifications including a cubic in prior score (and in IV, instrumenting with a cubic in the two-year lag score). We normalize the test scores to be mean zero in each grade, so that the coefficient on the linear term (δ) can be interpreted as the

[9] 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.

depreciation rate of knowledge for a student with an average score. We use this estimate of δ for the average student in all the formulas above.

In the Kalman filter model, there is a natural way to interpret the cubic relationship between current and prior test scores. Recall that the coefficient on the prediction residual from t-1 $\left(\hat{u}_{i,t-1|t-2}\right)$ in the Kalman model (equation 7) is δK_t, where K_t is the Kalman gain. If all test scores have similar measurement error then K_t only varies by time, declining over time as the prediction in t-1 becomes more precise and the prediction residual in t-1 becomes mostly measurement error. However, most standardized tests are designed to minimize error for students with achievement near the mean, and therefore have less measurement error in the center of the distribution and more measurement error for scores in the tails.[10] If this is true, then the Kalman gain should be smaller for students whose prior score is in the tails of the distribution. A typical plot of the standard error of measurement against percentiles of the test score looks roughly quadratic – suggesting that we should interact the prediction residual from t-1 $\left(\hat{u}_{i,t-1|t-2}\right)$ with a quadratic a student's baseline score. Since the prediction residual includes the baseline score, this is analogous to including a cubic in baseline score in the OLS and IV VAM models. Therefore, we estimate versions of the Kalman filter model including interactions between the prediction residual and a quadratic in the baseline score.[11]

6. Tests of model assumptions and potential bias

For each of our models we perform a series of specification tests 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 OLS VAM models using only the prior year score as a covariate with no lagged teacher effects. We then test the Kalman filter model against less restrictive models that flexibly include a student's 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 two empirical tests for the presence of student heterogeneity in growth (α_i) in each of our three models (OLS VAM, IV VAM, and Kalman). First, using each student's census tract income (something not typically available in VAM models), we parameterize $\alpha_i = Income * \gamma$ and directly estimate the impact of income on growth and the

[10] The technical documentation on NC EOG tests includes details on the standard error of measurement suggesting much more error in the tails. This is likely to be less true with adaptive tests which adjust to ask more appropriate questions to students in the tails of the distribution.

[11] We also include the direct effect for the quadratic term in baseline score, although this has little impact.

variance of $Income * \gamma$. Our second empirical test for the presence of α_i uses the covariance of each model's residuals two or more years apart to estimate the variance of α_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. To estimate this covariance, we estimate a Hierarichical Linear Model (using the mixed command in Stata) for the student residuals, allowing for random student intercept (α_i) and a Toeplitz error structure with one lag to account for the additional error terms in each model that may add variance and covariance at one lag.

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.[12] As with income, we add this variable to our models and quantify the amount of variation this type of private information can explain.

Fourth, to quantify whether any of these potential biases matter in practice, we estimate teacher persistent and transitory effects from each of the models, with and without additional controls for income and prior teacher's judgement and report the correlation in these estimates across models.

Fifth, we use observable proxies for each type of potential sorting and estimate how these proxies vary systematically across teachers. To do this, we estimate HLM models (using the mixed command in Stata) in which the dependent variable is one of the observable proxies, and we estimate random intercepts at the school and teacher level. These models how much actual sorting actually occurred based on these observable proxies. We consider the following proxies, available at the student-level:

1. $Income * \gamma$, which proxies for sorting on student growth.
2. TeacherJudgement*b, which proxies for sorting on private information about student knowledge.
3. The Kalman prediction of student knowledge at time t-1 ($\left(\hat{\mu}_{i,t-1|t-2}\right)$), which proxies for the extent of sorting on observable information about student knowledge.
4. The Kalman prediction residual at time t-1 ($\left(\hat{u}_{i,t-1|t-2}\right)$), which proxies for the extent of sorting on the measurement error in test scores in t-1. This is an imperfect proxy because the Kalman prediction residual is a mix of true student knowledge and measurement error.

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

5. The prior year teacher's transitory effect ($\varphi_{j',t-1}$), which proxies for sorting on prior teacher fadeout.

Finally, we use the IV VAM estimates to do a full decomposition of the variance components in our model $\left(\theta_{jt}, \varphi_{jt}, \alpha_i, \upsilon_{it}, \eta_{it}, and\ \mu_{i,t}\right)$. 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 $\left(\theta_{jt}, \varphi_{jt}\right)$ using standard empirical Bayes methods that correct for (correlated) estimation error in the fixed effect estimates (Kane and Staiger, 2008). 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 IV VAM model's residuals $\left(\alpha_i + \upsilon_{it} + \left(\eta_{it} - \delta\eta_{i,t-1}\right)\right)$. We estimate a Hierarichical Linear Model (using the mixed command in Stata) for the student residuals from grades 5-8, allowing for random student intercept (α_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 IV VAM 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 $Var(\mu_{it}) = Var(y_{it}) - Var\left(\varphi_{jt}\right) - Var(\eta_{it})$.

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 excluded 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.

Due to the data requirements of the state-space model, we exclude from our sample any students missing 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 indicator, limited English proficiency or learning disability, or the id codes for primary math and reading instructor.[13]

For a subset of our analyses, we use data from 2007 through 2013 in which teachers were asked to provide their subject 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

[13] 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.

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 2010 and 2017, the data provided correspond to 2010 Census block group boundaries. Between 2007 and 2009, the data provided by NCERD correspond to the 2000 block group boundaries. As a measure of neighborhood income, we the 2013 ACS 5-year estimate of median household income by block group between 2010 and 2017. Between 2007 and 2009, we use the 2000 Census median household income measure by block group. We sort the neighborhood income measures each year into deciles, after converting to 2022 dollars.[14] 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 NCERD did not provide block groups for these students.

Results

In Table 2, we evaluate the Kalman filter estimates' ability to predict end-of-year test scores compared to conventional VAM specification and against more flexible specifications. Since the Kalman filter estimates are optimal if our statistical model is correct, these results provide a test of the assumptions in our statistical model. To provide the strongest test of the Kalman, we focus on 8th grade math achievement, given that we have achievement measures in five prior years (grades 3, 4, 5, 6 and 7) and teacher assignments in four prior grades (4th, 5th, 6th and 7th).

One key difference between our model and the conventional VAM model is the inclusion of fixed effects for both prior year and current teacher. Column (1) reports the results of the conventional VAM model with fixed effects for current grade 8 teachers, while column (2) includes fixed effects for both 7th and 8th grade teachers. Moving from column (1) to (2), the adjusted R^2 increases from .7448 to .7520 and the F-test rejects exclusion of the prior teacher effects. That is what we would expect if there were any transitory effect of last year's teacher not captured by students' the baseline score.

Column (3) reports results for the Kalman filter specification. The Kalman filter breaks baseline achievement into two parts: the expected achievement (based on prior scores and teacher assignments) and the new information which emerged in the baseline year, t-1. While the conventional VAM forces the same coefficient on both parts—estimating a coefficient of .758 on baseline score-- the Kalman filter allows us to estimate different coefficients on each, .925 on the Kalman prediction and .407 on the Kalman residual. One of the reasons that conventional value-added is biased is that it constrains these coefficients to be equal. The

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

Kalman is clearly a better fit, with an adjusted R^2 of .7877 vs. .7520. Given the large sample size, the p-value of the F-statistic comparing the specifications in columns (2) and (3) is less than .001. Note also that while the coefficient on the Kalman prediction of achievement in t-1 is near one, .925, the coefficient on the Kalman residual (which will include measurement error in t-1 achievement, $\eta_{it-1,}$as well as persistent effects of t-1 interventions, ν_{it}, and new information about a student's true state) is much smaller (.407).

In the remaining columns of Table 1, we test how well the Kalman filter summarizes the information from prior performance and teacher assignments by gradually allowing for more flexible specifications. For instance, in column 4, we add a control for 6th grade math achievement. In column (5) we estimate independent effects for lagged math achievement in grades 3 through 6. In both instances, the F-test rejects the Kalman specification in Column (2), but the adjusted R^2 remains the same through four digits, .7877. In column 6, we add lags of reading achievement in grades 3 through 6. The adjusted R^2 increases slightly to .7898, an increase of .0022. In the last column, we add fixed effects for each students' history of teacher assignments in grades 3-6. The increase in adjusted R^2 is somewhat larger, but still is less than a full percentage point larger (.0072) than the Kalman filter specification in column (2). We take this as evidence that the Kalman filter faithfully summarizes the history of students' achievement and teacher assignments quite well—certainly better than a single year of achievement as with conventional value-added and nearly as well as a completely flexible specification.

Table 3 reports similar results for reading. The Kalman specification fits better than the conventional VAM (adjusted R^2 of .7263 vs. .6653). As with math, the inclusion of all lags and the full history of teacher dummies in column (7) improves the adjusted R^2 by less than 1 percent, by .0068, relative to the two-part Kalman specification in column (3).

As discussed earlier in the methods section, VAM models commonly find non-linearity in the relationship between end-of-year scores and prior year scores, with a flatter relationship in the tails. To account for this non-linearity, conventional VAM models often include a cubic in prior score. In appendix Table A1 we compare conventional VAM models with linear controls for prior scores to those with cubic controls for prior scores and find that the additional cubic terms are highly significant and improve the adjusted R-squared. Similarly, when we estimate VAM by IV in Appendix Table A1, we find the additional cubic terms are highly significant (adjusted R-squared is not relevant for IV). Finally, as discussed in the methods section, an analogous specification in the Kalman specification would include interactions between a quadratic in baseline score and the Kalman residual to account for greater measurement error in the tails of the baseline score distribution. Again, in Table A1 we find that the additional interactions and quadratic terms are highly significant and improve the adjusted R-squared of the Kalman specification. Therefore, in the remainder of the paper we focus on specifications that include these non-linear terms, although none of the results are qualitatively changed if we reported results from more simple linear models.

In Table 4, we incorporate the subjective assessments of prior grade teachers regarding their students' mastery of math and reading standards. The data were collected as students were sitting down to take the state tests and before test results were available. Teachers chose from five categories from insufficient mastery to consistently superior.[15] Although such measures are typically not available to the value-added researcher, we use it here to test whether the prior grade teachers have additional information (beyond a students' achievement history and prior teachers) which could be used for sorting and bias. In the state-space model, year t achievement depends solely on period t inputs ($\nu_{it}, \alpha_i, \theta_{jt}, \psi_{jt}$) and true baseline achievement, μ_{it-1}. However, any additional information teachers have regarding a student's true baseline knowledge is an additional source of selection, which would lead to bias in conventional VAM and the Kalman filter specifications if used to sort students.

We included indicators for whether the prior teacher rated the students' mastery of state standards as "insufficient," "inconsistent", "consistently superior", or "none of the above" with the most common response, "consistent mastery," as the excluded category. Clearly, the prior grade teachers do have information not captured in the t-1 achievement scores: in the conventional VAM for math (column 1), the students who were rated as having insufficient mastery of math by their 7th grade teacher scored .323 SD lower on the 8th grade test than those who had the same 7th grade scores, but whose teachers reported they had "consistent" mastery. Meanwhile, those judged to have "consistently superior" achievement scored .205 SD higher.

We created an index of teacher judgements, using the coefficients in Table 3, and regressed the index on the remaining variables in Table 3, using the residual to calculate the potential bias from excluding such private judgements. The estimated variance was .011, implying an SD of .105. Because it is similar in size to the signal variance in teacher effect estimates, it implies that perfect sorting on teacher judgements could produce 100 percent bias in teacher effects. (However, we will present evidence in Table 5 that there is only modest sorting on teacher judgements in our data.)

In the next column, we report a similar test for the Kalman filter specification. Apparently, grade 7 teachers do have information not captured by the Kalman prediction or Kalman residual, as the coefficients on teacher judgements remain significant. However, the variance of the potential bias due to teacher judgements is about 70 percent smaller, .003, implying an SD of .059, about half the size of the SD of conventional VAM estimates of teacher effects. Thus, while the Kalman fails to fully capture the information that the prior year teacher has about student knowledge, it captures substantially more information than conventional VAM.

[15] In addition to the mastery questions, teachers were asked to report the grade they were expecting to give to each student. We have done a parallel analysis using teacher's planned grades and found very similar results.

In column (3), we report the same test for the IV specification. In the IV, the coefficients on the teacher judgement variables are near zero and individually insignificant. We cannot reject the joint hypothesis that teacher judgements carry no additional information (p-value=.801.) and the variance in potential bias is zero to three digits with an implied SD of .006.

As reported in the right-hand panel of Table 3, the results are generally similar in reading. Prior teacher judgement has effects that are large and significant in conventional VAM, about half the size but still significant in the Kalman specification, and near zero in the IV specification (although the p-value of the F-test is .008.) The implied SD in the potential bias due to teachers' private information is .119 in the conventional VAM, .050 with the Kalman specification and .008 for the IV.

The results in Table 4 highlight the main advantage of the IV specification. Unlike conventional VAM and to a lesser extent the Kalman specification, the prior-year teacher's judgement cannot predict student gains in the IV specification. The reason for this is twofold. First, by instrumenting with twice-lagged scores, the IV specification corrects for attenuation bias due to measurement error in the lagged score and gets an unbiased estimate of δ, the true depreciation rate of baseline knowledge. Second, because the IV specification correctly accounts for the depreciation of the baseline knowledge that is part of the lagged score, the IV residual no longer has any association with baseline knowledge – and, hence, private information about baseline knowledge plays no role in the IV specification.

In Table 5, we perform a similar analysis, incorporating information on the median household income of the Census block group (sorted into deciles) where a student resides. Again, such data are typically not available to the value-added researcher. We include them to test if there is any differential growth for students from high- and low-income neighborhoods, conditional on baseline achievement and teacher assignments. In the conventional VAM specification, students living in census tracts in the top three income deciles scored .03 to .04 standard deviations above those with similar baseline achievement residing in the lowest income decile neighborhoods. Columns (2) and (3) repeat the exercise using the Kalman Filter and IV specifications. Although there were small differences in achievement growth not captured by the Kalman filter (a potential bias variance of .00006), neighborhood income played little role in the IV model. In fact, the p-value of the constraint that all the differences by neighborhood income were zero, was .213 for math and .992 in reading. The potential bias variance due to exclusion of the neighborhood income declines was .00003 in math and .00002 in reading for the IV specification.

The results in Table 5 suggest that neighborhood income does not identify students who consistently have higher test score growth (α_i). This is surprising but may reflect that census block income is a weak proxy for the factors that contribute to α_i. In our model, if there is substantial variance in α_i across students, it will generate a positive correlation in their residuals across grades two or more years apart: students with unexpectedly high test-score growth in

early grades will continue to have unexpectedly high test-score growth in later grades.[16] For residuals two or more years apart, the correlation is .07 for conventional VAM in math and .09 in reading.[17] In the IV and Kalman, the correlation in residuals two or more years apart is below .01 for both math and reading. Thus, there is little evidence of "fast learners and slow learners", e.g. students who consistently have higher (or lower) than expected test score growth.

Under the IV specification, neither teacher judgements nor neighborhood income were predictive of student outcomes, implying that even if there were perfect sorting on those traits, it would not lead to bias for the IV. In contrast, when using conventional VAM or the Kalman filter, any sorting on neighborhood income or private teacher judgements would lead to bias. However, the amount of potential bias for the Kalman due to sorting on neighborhood income was considerably smaller than the potential bias from sorting on teacher's private information.

Table 6 gauges the amount of bias due to sorting on each of five factors: neighborhood income, teacher judgements of students' mastery, the transitory impact of prior teachers, the Kalman prediction of the prior grade score and the Kalman residual for the prior grade score. The estimates in Table 6 are derived from hierarchical linear models (HLM) regressing each of the listed variables on random school effects and random teacher effects nested within schools. We include the school-level random effects as an indicator of sorting of students between schools. The teacher random effects, because they are nested within schools, indicate the amount of sorting which occurred across teachers within schools. Sorting within schools will bias estimates of teacher effects but has no impact on estimates of school-level value added, while sorting between schools affects estimates of both teacher and school value added.

The top panel reports the amount of sorting on the index of predicted achievement growth by neighborhood income deciles (estimated in Table 5.) Because we used the direct effects of neighborhood income on achievement growth to create the index, the estimates in Table 6 are interpretable as estimates of the variance of actual bias which would result from excluding neighborhood income. As a benchmark, typical estimates of the variance of school and teacher effects range from .01 to .04 (SD .1 to .2). The ratio of our estimates of bias variance to these benchmark estimates of total variance indicates the degree of bias in teacher effects due to sorting on neighborhood income.

In the conventional VAM model, we estimate that the standard deviation of the bias due to excluding students' neighborhood income is .013 at the school level (a variance of .0002) and .005 for teachers within-schools (a variance of .000025). Relative to a total variance in school and teacher effects of .01-.04, the implied bias from excluding neighborhood income is expected to be less than 2 percent for schools and less than 1 percent for teachers. There are

[16] At the first lag, there is a substantial negative correlation in the conventional VAM and IV VAM residuals due to estimation error from the prior year score (η_{it-1}).

[17] This is because in the conventional VAM, the coefficient on baseline score is attenuated, leaving some of true baseline achievement in the residual.

similarly small amounts of bias from the exclusion of neighborhood income for the Kalman filter and IV estimators. The results for reading are similar. Thus, there is no evidence that the actual sorting of students to schools and teachers in NC contributed to bias.

The next panel in Table 6 reports estimates of bias due to the exclusion of teacher's judgement of students' mastery of standards. In the conventional VAM, for math, the standard deviation in school and teacher effects is .051 and .054 respectively, implying bias variance is .003 for schools and .003 for teachers within schools—as much as 30% percent of typical school and teacher variance estimates. In the Kalman filter, the standard deviations are .043 and .033 respectively, implying a bias variance of .002 and .001 for teachers and schools respectively, implying as much as 20% bias for teachers and 10% bias for schools. By contrast, the bias variance in the IV is very small, just .00012 for schools and .00012 for teachers, so excluding teacher judgement would lead to minimal (<2%) bias in the IV. As shown in the remaining columns, the results are similar for reading.

The third panel reports sorting based on the prior teacher's transitory effect on achievement. In order to interpret as bias, we used the empirical Bayes estimates of the transitory teacher effects of last year's teacher ($-\delta\ \psi_{jt-1}$) to account for the portion that is passed along from one year to the next. We have included fixed effects for prior teachers, so any sorting on this variable should not lead to bias. However, the results for both math and reading suggest that there would be considerable bias if not conditioning on prior year teacher. Much of the sorting happens at the school level – prior teachers at some schools tend to have highly transitory effects on test scores while prior teachers at other schools do not.

The fourth panel reports the amount of sorting based on the Kalman prediction of the baseline score. Since the Kalman model includes controls for the Kalman prediction, such sorting does not lead to bias. Nonetheless, the result suggest that there is a considerable amount of sorting based on predicted baseline achievement: for both math and reading, a standard deviation at the school level of .31 and for teachers within schools of .28. Given this evidence that there is pervasive sorting on predicted baseline achievement, the potential for bias in conventional and IV VAM models resulting from sorting on private information about baseline achievement is of particular concern.

The final panel in Table 6 reports similar results for the Kalman residual. The Kalman residual contains several components: the true effect of year t-1 interventions, ν_{it-1}, student-level heterogeneity in growth, α_i, and measurement error in baseline achievement, η_{it-1}. In addition, because the Kalman filter could not perfectly predict baseline achievement, some component of baseline achievement is also in the Kalman residual. Since the unbiasedness of the IV estimator depends on absence of sorting on measurement error in baseline achievement, η_{it-1}, we interpret sorting on the Kalman residual as an upper bound estimate of the sorting on baseline measurement error, η_{it-1} . There is no sorting at the school level (not a surprise, given that we included teacher effects in the Kalman –even though this is lagged, it is likely to

nearly perfectly absorb school effects). There's also very little evidence of sorting at the teacher level, with an SD of .025 (variance .0006) for math and .010 (variance .0001) for reading, implying an upper bound of 6% bias for math and 1% bias for reading due to sorting on measurement error in the baseline score.

In the first three columns of Table 7, we report the correlation among the three sets of estimates—conventional VAM, Kalman and IV. All include fixed effects for current and lagged teacher that are used to estimate the permanent and transitory teacher effects. The correlation in the permanent teacher effects estimated by the three methods is between .85 and .88 for math. The correlation in the transitory effect is even higher, ranging from .91 to .96 for math. As reported in the lower panel, the results are similar for reading. In other words, as long as one is conditioning on current and lagged teacher, the estimated teacher effects are highly correlated under the three methods.

In the last three columns of Table 7, we report the correlation between estimates without and with conditioning on teacher ratings of student mastery and neighborhood income effects. While the left side of Table 7 is estimated for the full sample of six cohorts, the estimates on the right side of Table 7 are limited to the three cohorts for whom teacher ratings are available. The correlation is highest for the IV model, .998 to 1. That is to be expected given that the IV should not be sensitive to teacher private information learned during the prior school year. However, the conventional VAM is also largely unaffected by the inclusion of the additional controls, with a correlation of .955 to .964. This simply reflects the finding from Table 6: while there is considerable sorting to teachers and schools based on expected achievement in the baseline year, there is little sorting based on teacher's private information about students' current state of knowledge or neighborhood income. Comparing the upper and lower panels of Table 6, the results are similar for math and reading.

Table 8 summarizes the main parameters of the state space model. We estimate these only for the IV specification because there is less evidence of bias in this specification and because the other two models have other components in their residuals that making estimating the error components more complex. As reported in the top row, for both math and reading, we estimate that slightly less than 100 percent of the beginning stock of knowledge and persistent innovations each year is passed on as students transfer between grades. ($\hat{\delta}$ is .972 in math and .984 in reading.) In the absence of a vertical scale that extends from one grade to another, we have standardized test scores within each grade and year. Under the state space model, with random innovations arriving each year, the variance in achievement would be growing in each grade. Thus, a one unit increase in achievement in grade g would translate into a less than full unit increase in achievement in grade g+1, even if it were permanent (Cascio and Staiger, 2012).

The next two rows of Table 8 contain our estimates of the variance in measured achievement as well as the measurement error variance. Our estimates imply a test reliability

of about 90 percent for math and 85 percent for reading – in line with the reliability reported in technical reports for NC end-of-grade tests in these years.

The next rows report the variance in true knowledge and the variance in baseline achievement predicted by the Kalman filter. The larger the difference between the two, the greater the scope for private information about a student's true state to lead to bias in the Kalman filter (although that information would not lead to bias in the IV). The estimated variance in baseline achievement predicted by the Kalman is .629 and .619 in math and reading respectively. The true variance in knowledge implied by the model is .759 in math and .730 in reading. The roughly .11 to .13 gap in variance between predicted achievement and true achievement implies that teachers or parents or students may have private information about a student's baseline state of knowledge which, if used to sort into teachers or other interventions could lead to substantial bias. This is not a concern for IV estimates because teacher effects using the IV method are not biased by private information about student baseline achievement.

The next panel refers to the variance in the annual innovations to achievement which persist into future years. The first two rows describe the share of innovations associated with student unobservables and student traits. As noted above, a limited share of the growth from one year to the next is attributable to fixed differences in student growth unrelated to student traits: .001 in math and reading. Following the convention in the value-added literature, we also condition on indicators for student race/ethnicity, gender, economic disadvantage, limited English proficiency and learning disability. When we form an index by multiplying each of those indicators by the coefficients from the value-added models, the variance is .005 in math and .002 in reading. When combined with the heterogeneity in student growth from unobservables (α_i), the combined student level variance is .006 in math and .003 in reading.

For comparison, the variance in persistent teacher effects is .022 in math and .019 in reading—implying a standard deviation in teacher effects of .148 in math and .138 in reading, both of which are in range reported in the prior literature (Koedel et al. (2015)). In addition to teacher effects, we estimate that the variance in other persistent innovations is .040 in math and .037 in reading—yielding a combined variance in persistent innovations of .062 in math and .056 in reading. The variance in growth due to annual persistent innovations is 10-20 times larger than the variance due to student-level factors. In other words, the variance in achievement is growing over time primarily due to the varying quality of the teachers and other idiosyncratic factors that affect students that year—not because some students are fast learners or slow learners and not because of differences in student growth associated with demographic characteristics or income.

The next panel in Table 8 reports on the share of teacher effects which is transitory. Although many have estimated aggregate fade-out of teacher effects, the state space model allows us to estimate both a permanent and transitory effect for each teacher.[18] Our estimate

[18] One exception is Jacob, Lefgren and Sims (2010), who allow teachers to have differing levels of persistence.

of the variance in transitory teacher effects (effects which dissipate after the current year) is .027 in math and .013 in reading, implying a standard deviation in transitory teacher effects of .16 in math and .11 in reading. The model also implies a modest negative correlation in teachers' permanent and transitory effect: those who have larger permanent effects on student achievement tend to have smaller transitory effects. The variance of the sum of current and transitory effects in a single year is .039 in math and .017 in reading (or .197 standard deviations in math and .13 standard deviations in reading.) The implied average fade-out rate (the share of the variance in total effects which persists) is 44 percent in math and 70 percent in reading. However, there is considerable variation in this relationship: the correlation between the total (transitory + persistent) and persistent effect is only .59 for math and .66 for reading.

In Table 9, we estimate the indirect effect of neighborhood income operating through quality of teachers. As reported in Table 5, the direct effect of differences in growth between students from the highest and lowest income neighborhoods, conditional on teacher effects, is quite small. However, income may have an indirect effect through teacher quality. The first two columns of Table 9 report differences across income deciles (relative to the lowest income decile) in average teacher persistent and transitory effects for math. The difference between the top and bottom three deciles of neighborhoods in terms of average persistent teacher effects is .032 in math. In other words, each year, students in the highest income neighborhoods receive .032 standard deviations more in persistent math teacher effects than students in the lowest income neighborhoods. Moreover, students in higher income neighborhoods receive less transitory teacher effects (which disappear after one year) in math (a difference of -.015 between the top and bottom three deciles). The next two columns report the differences in persistent and transitory teacher effects for math after conditioning on school fixed effects. Apparently, much of the observed difference is due to differences in teacher quality by school. The remaining columns repeat the exercise for reading. In contrast to math, there is very little difference in teacher quality (persistent or transitory) between the top and bottom deciles.

Summary and Conclusion

The state space model allows us to take the catch-all concern of "selection on unobservables" and parse it into four parts: heterogeneity in student growth rates, transitory teacher effects, private information about students' baseline knowledge and measurement error in baseline achievement. We find little evidence of heterogeneity of student growth rates in North Carolina. This may explain one of the more surprising findings in the value-added validation studies: that it is not just teacher effects which are forecast unbiased but school effects too (Deming (2014), Angrist et al. (2017), Angrist et al. (2024)). While the value-added researcher could plausibly have access to the same administrative data used to assign students to teachers within schools, that would not be true for the many unobserved factors leading families to sort between schools. One might expect student growth rates to vary, based on unchanging factors such as parents' ability to pay tutors family homework routines, student motivation or family involvement in education. To the extent that fixed family background

factors do matter, though, they may be reflected in students' starting knowledge, not growth rates.

Rather than student-level heterogeneity in growth, our findings imply that the primary challenge in value-added estimation is finding a measure of baseline achievement which is free from the remaining three sources of bias: transitory teacher effects, private information about students' current state and measurement error. Each is a special case of measurement error in baseline achievement. Fortunately, our findings point to ways to resolve each of them:

Transitory teacher effects: Although others have reported transitory effects of teachers on student achievement, such effects have been underappreciated as a potential source of selection. Under the assumption that transitory effects disappear after one year, and that persistent effects become part of a students' stock of knowledge to be passed on into future years, this is perhaps the easiest of the three to resolve: researchers should include fixed effects for current and prior teacher.

Private information about a students' current state of knowledge: Under the state space model, future knowledge is a function of current knowledge plus innovations from teachers and other interventions. Therefore, when we are measuring intervention impacts using growth in imperfectly measured achievement, private information about a students' true knowledge is an obvious source of potential selection bias. Using a unique set of questions available in the NC data, we find that teachers do indeed have private information about students' knowledge not reflected in test score histories. Our model also implies that getting an unbiased estimate of the change in knowledge, by instrumenting with twice lagged achievement, should resolve the problem. It appears to do so. When we instrument for baseline achievement with twice-lagged achievement, the prior grade teachers' subjective ratings are no longer predictive of end of grade achievement.

Measurement error in baseline achievement: Prior research has largely ignored the potential role of measurement error in sorting students to teachers and interventions. The usual solution to measurement error in achievement—such as instrumenting with prior lags—may yield a baseline score which is independent of the measurement error, but it does nothing to resolve the fact that the intervention assignments themselves may also be a function of the baseline measurement error. If there is sorting on the measurement error, there would still be bias in the intervention impact estimates. Fortunately, many states do not provide the test scores used in value-added measures until after teacher and school assignments are made in late summer—thus making it difficult to select based on measurement error. Indeed, the late delivery of scores may be one of the reasons why value-added measures of teacher effects have been forecast unbiased! But sorting on measurement error in baseline achievement would present a bigger problem for measuring the impact of tutoring and other catch-up interventions, which start after the beginning of the year. In such cases, the best approach may be to administer a new test post assignment, and to instrument for that new baseline measure of achievement.

While there are solutions to the latter three sources of selection, addressing the challenge of student-level heterogeneity in growth is more challenging. We find little evidence of student level of heterogeneity in growth in the North Carolina data. The fact that others have found value-added estimates to be forecast unbiased in a variety of settings leads us to believe that the same may be true elsewhere. Nevertheless, we encourage others to use our statistical model to test for student level heterogeneity in growth in their own data.

If the finding of limited heterogeneity in student level growth proves to be generally true, it would set the stage for a much broader effort to evaluate educational products and interventions, not just teachers and schools. Value-added studies could be a first cut, followed up with randomized trials, especially for expensive interventions. But given the value of student achievement for earnings and productivity growth, even a small increase in the rate at which effective interventions are identified and spread would produce large social returns.

Our results also have implications for other analyses of student-level panel data. Estimates of our state-space model suggest strong persistence of prior innovations to knowledge with δ near to 1, which implies that student test scores contain a large unit root component (plus some transitory measurement error). As in the time series literature, 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. Moreover, other aggregations of student achievement, such as school and district averages, may also be following similar patterns – with unit roots in the aggregate measures. We will be exploring those implications in future research.

References:

Andrabi, Tahir, Natalie Bau, Jishnu Das, Asim Ijaz Khwaja 2022 "Heterogeneity in School Value-Added and the Private Premium" NBER Working Paper 30627
http://www.nber.org/papers/w30627

Angrist, J.D., Hull, P.D., Pathak, P.A., Walters, C.R. (2017) Leveraging Lotteries for School Value-Added: Testing and Estimation. The Quarterly Journal of Economics (2017), 871–919. doi:10.1093/qje/qjx001.

Angrist, J.D., Hull, P.D., Pathak, P.A., Walters, C.R. (2024) Credible school value-added with undersubscribed Deming, D. J. (2014). Using School Choice Lotteries to Test Measures of School Effectiveness. *American Economic Review,* 104(5), 406-11.

Bacher-Hicks, Andrew, Thomas J. Kane and Douglas O. Staiger, "Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles" *NBER Working Paper No*. 20657, November 2014.

Blazar, David, Blake Heller, Thomas J. Kane, Morgan Polikoff, Douglas O. Staiger, Scott Carrell, Dan Goldhaber, Douglas N. Harris, Rachel Hitch, Kristian L. Holden, Michal Kurlaender "Curriculum Reform in the Common Core Era: Evaluating Elementary Math Textbooks Across Six U.S. States" Journal of Policy Analysis and Management (2020) Vol. 39, No. 4, pp. 966-1019.

Britton, Jack; Clark, Damon; Lee, Ines (2023) : Exploiting discontinuities in secondary school attendance to evaluate value added, IFS Working Papers, No. 23/24, Institute for Fiscal Studies (IFS), London,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 CALDER Working Paper No. 291-0823

Chetty, Raj, John Friedman and Jonah Rockoff (2014) "Measuring the Impacts of Teachers I: Evaluating
Bias in Teacher Value-Added Estimates" American Economic Review, pp. 2593-2632.

Ehlert, Mark, Cory Koedel, Eric Parsons & 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, DOI:10.1080/2330443X.2013.856152

Glazerman, Steve and Ali Protik (2015) "Validating Value-Added Measures of Teacher Performance" Mathematica Policy Research, February 20, 2015.

Hanushek, Eric A. (1971) Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data. American Economic Review Papers and Proceedings 61(2):280-88.

Hanushek, Eric A. (1979) Conceptual and empirical issues in the estimation of educational production functions. The Journal of Human Resources,14(3),351–388.

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, Vol 65(1) 7–23

Jacob, Brian A., Lefgren, Lars & Sims, D.P. (2010) "The persistence of teacher-induced learning gains." Journal of Human Resources, Vol. 45(4), pp. 915–943.

Kane, Thomas J. and Douglas O. Staiger (2008) "Estimating Teacher Impacts on Student Achievement: An
Experimental Evaluation" *NBER Working Paper 14607*.

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

Koedel, Cory, Kata Mihaly and Jonah Rockoff (2015) "Value-Added Modelling: A Review" Economics of Education Review Vol. 47 pp. 180–195

McCaffrey,D.F., Sass,T.R., Lockwood,J.R.,&Mihaly,K.(2009)."The intertemporal variability of teacher effect estimates." Education Finance and Policy,4(4),572–606.

Meghir, Costas and Steven G. Rivkin (2011) "Econometric Methods for Research in Education," Handbook of the Economics of Education Volume III.

Murnane, Richard J. (1975) The Impact of School Resources on Learning of Inner City Children. Cambridge, MA: Ballinger Publishing.

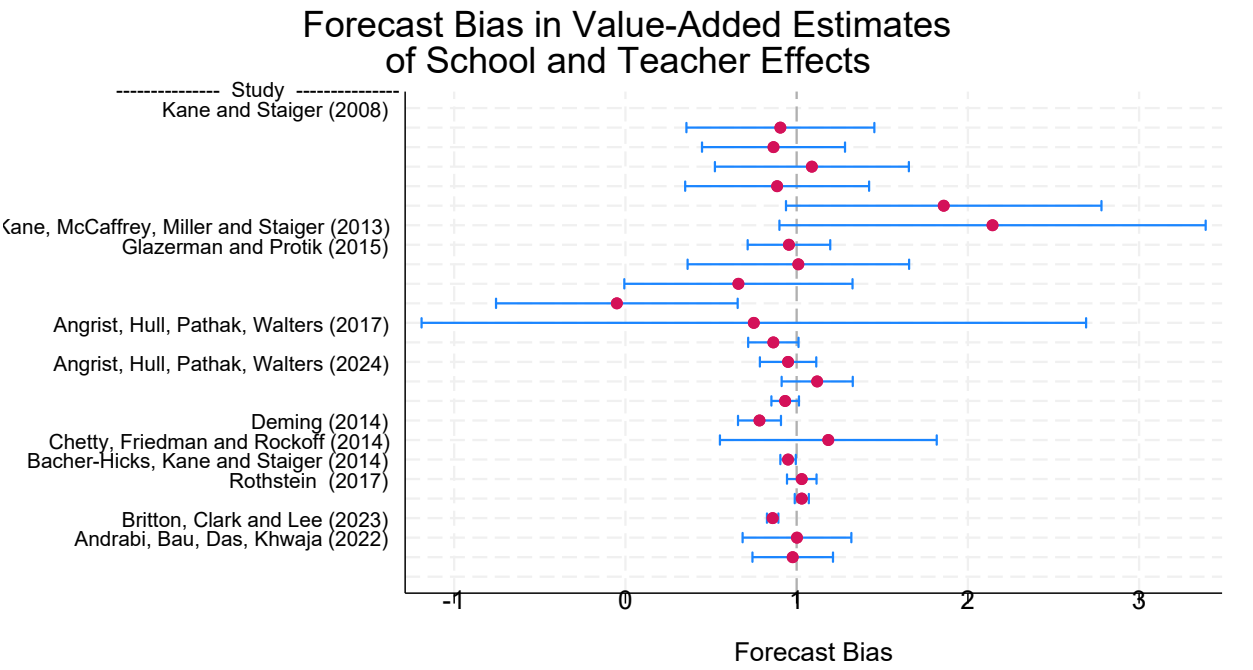
Plecki, Margaret L., Ana M. Elfers, and Yugo Nakamura (2012) "Using Evidence for Teacher Education Program Improvement and Accountability: An Illustrative Case of the Role of Value-Added Measures" Journal of Teacher Education 63(5) 318–334.

Rothstein, Jesse (2010) "Teacher Quality in Educational Production: Tracking, Decay, and Student
Achievement." *Quarterly Journal of Economics* 125(1), February 2010, pp. 175-214.

Rothstein, Jesse (2017). "Revisiting the Impacts of Teachers". IRLE Working Paper No. 101-17. http://irle.berkeley.edu/files/2017/Revisiting-the-Impacts-of-Teachers.pdf

Todd, Petra and Kenneth Wolpin (2003) "On the specification and estimation of the production function for cognitive achievement" The Economic Journal, 113 (February), F3–F33.

Figure 1.



Notes: A coefficient of one implies forecast unbiaseness. See Table 1 for details on the specifications, grade levels, subject and ty

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 peer controls)	Random assignment	.905 (.280)
		Teacher	Elem.	Math	Los Angeles	Gain Score (w peer controls)	Random assignment	.865 (.213)
		Teacher	Elem.	Reading	Los Angeles	Lagged Score (w peer controls)	Random assignment	1.089 (.289)
		Teacher	Elem.	Reading	Los Angeles	Gain Score (w 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 (2013)	Teacher	Elem./ Middle	Math/ Reading	New York, Denver, Dallas Memphis, Hillsborough FL, Charlotte	Lagged Score (w peer controls)	Random assignment	.955 (.123)
3	Glazerman and Protik (2015)	Teacher	Elem.	Math	7 Large Districts	Lagged Score (w demog	Random assignment	1.01 (.33)
		Teacher	Elem.	Reading	7 Large Districts	Lagged Score (w demog	Random assignment	.66 (.34)
		Teacher	Middle	Math	7 Large Districts	Lagged Score (w demog	Random assignment	-.05 (.36)
		Teacher	Middle	Reading	7 Large Districts	Lagged Score (w demog	Random assignment	0.75 (.99)
4	Angrist, Hull, Pathak, Walters (2017)	School	Middle	Math	Boston	Lagged Score (w demog	Random assignment	.864 (.075)
		School	Middle	Math	Boston	Gain Score (w demog	Random assignment	.950 (.084)
5	Angrist, Hull, Pathak, Walters (2024)	School	Middle	Math	New York	Lagged Score (w demog	Random assignment	1.12 (.106)
		School	Middle	Math	Denver	Lagged Score (w demog	Random assignment	.933 (.041)
		School	H.S.	SAT Math	New York	Lagged Score (w demog	Random assignment	.783 (.064)
6	Deming (2014)	School	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 demog controls& 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 demog controls& drift adjustment)	Quasi Exp (shifting teacher assignments)	1.030 (.044)
9	Rothstein (2017)	Teacher	Elem.	Math/ Reading	North Carolina	Lagged Score (w demog controls& drift adjustment)	Quasi Exp (shifting teacher assignments)	1.030 (.021)
		Teacher	Elem.	Math/ Reading	North Carolina	Lagged Score (w demog controls& drift adjustment)	Quasi Exp (shifting teacher assignments)+ change in prior year score	.860 (.017)
10	Britton, Clark and Lee (2023)	School	Middle	Math/ Reading	England	(w demog controls) (No Shrinkage)	Quasi Exp (RD in distance to school)	1.002 (.162)
11	Andrabi, Bau, Das, Khwaja (2022)	School	Elem.	Math, English, Urdu	Pakistan	Lagged Score (w demog controls)	Quasi Exp (School Closures)	.977 (.120)

Table 2. Testing the Kalman Filter: 8th Grade Math

	Conventional VAM (1)	Conventional VAM + Lagged Teacher (2)	Kalman Filter (3)	Kalman + G6 Math Test (4)	Kalman + All Lags Math Test (5)	Kalman + All Lags Math & Reading (6)	Kalman + All Lagged Tests + Full Teacher History (7)
G7 Math	0.754***	0.758***	-	-	-	-	-
	(0.001)	(0.001)	-	-	-	-	-
G7 Kalman Math	-	-	0.925***	0.916***	0.895***	0.755***	1.017***
	-	-	(0.001)	(0.004)	(0.040)	(0.041)	(0.063)
G7 Kalman Math Residual	-	-	0.407***	0.407***	0.454***	0.398***	0.376***
	-	-	(0.002)	(0.002)	(0.038)	(0.038)	(0.002)
G8 Math Tchr	X	X	X	X	X	X	X
G7 Math Tchr		X	X	X	X	X	X
G6 Math				X	X	X	X
G3-7 Math					X	X	X
G3-7 Reading						X	X
G3-7 Math Tchr							X
Adj. R^2	0.7448	0.7520	0.7877	0.7877	0.7877	0.7898	0.7948
Difference in Adj. R2 relative to Col (3)	-0.0428	-0.0357	0	0.0000	0.0001	0.0022	0.0072
p-value of F-test of constraint	-	0.000 Vs. Col (3)	-	0.027 Vs. Col (3)	0.000 Vs. Col (3)	0.000 Vs. Col. (3)	-
N	276184	275828	273139	273139	273139	273139	273019

Note: All specifications include indicators for year, race, gender, economic disadvantage, IEP and LEP status. All models are linear in the covariates. $E[\mu_{i,t-1}|\Omega_{i,t-1}]$ incorporates scores and teachers from both subjects from all periods through $t-1$. Standard errors are in parentheses: * p < 0.05, ** p < 0.01, *** p < 0.001. Sample fluctuates slightly because singletons are dropped with multiple fixed effects.

Table 3. Testing the Kalman Filter: 8th Grade Reading

	Conventional VAM (1)	Conventional VAM + Lagged Teacher (2)	Kalman Filter (3)	Kalman + G6 Math Test (4)	Kalman + All Lags Math Test (5)	Kalman + All Lags Math & Reading (6)	Kalman + All Lagged Tests + Full Teacher History (7)
G7 Math	0.714***	0.707***	-	-	-	-	-
	(0.001)	(0.001)	-	-	-	-	-
G7 Kalman Math	-	-	0.956***	0.947***	0.623***	0.226***	0.562***
	-	-	(0.002)	(0.004)	(0.032)	(0.034)	(0.066)
G7 Kalman Math Residual		-	0.331***	0.331***	0.178***	0.067*	0.301***
	-	-	(0.002)	(0.002)	(0.029)	(0.029)	(0.002)
G8 Math Tchr	X	X	X	X	X	X	X
G7 Math Tchr		X	X	X	X	X	X
G6 Math				X	X	X	X
G3-7 Math					X	X	X
G3-7 Reading						X	X
G3-7 Math Tchr							X
Adj. R^2	0.6623	0.6653	0.7263	0.7263	0.7267	0.7305	0.7331
Difference in Adj. R2 relative to Col (3)	-0.0641	-0.0611	0	0.0000	0.0004	0.0042	0.0068
p-value of F-test of constraint	-	0.000 Vs. Col (3)	-	0.030 Vs. Col (3)	0.000 Vs. Col (3)	0.000 Vs. Col (3)	-
N	276158	275775	273119	273119	273119	273119	272990

Note: All specifications include indicators for year, race, gender, economic disadvantage, IEP and LEP status. All models are linear in the covariates. $E[\mu_{i,t-1}|\Omega_{i,t-1}]$ incorporates scores and teachers from both subjects from all periods through $t-1$. Standard errors are in parentheses: * p < 0.05, ** p < 0.01, *** p < 0.001. Sample fluctuates slightly because singletons are dropped with multiple fixed effects.

Table 4. Role of Prior Teacher Judgement of Student Mastery in Predicting End of 8th Grade Achievement

	Math			Reading		
	Conventional VAM	Kalman Filter	IV	Conventional VAM	Kalman Filter	IV
Prior Teacher Judgement						
Insufficient Mastery	-0.323***	-0.192***	0.002	-0.326***	-0.138***	0.002
	(0.009)	(0.008)	(0.011)	(0.009)	(0.009)	(0.011)
Inconsistent Mastery	-0.179***	-0.108***	0.006	-0.202***	-0.090***	-0.022***
	(0.004)	(0.004)	(0.006)	(0.005)	(0.005)	(0.006)
Consistent Mastery (ref)	-	-	-	-	-	-
Consistently Superior	0.205***	0.108***	-0.001	0.234***	0.078***	0.000
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.006)
None of the Above	-0.219**	-0.191**	-0.037	-0.260**	-0.185*	-0.002
	(0.067)	(0.060)	(0.087)	(0.096)	(0.086)	(0.109)
p-value for F-test	0.000	0.000	0.801	0.000	0.000	0.001
N (Grade 8)	122672	120662	122672	122634	120657	122634
$Var(Z\hat{y}\|Controls)$	0.011***	0.003***	0.000***	0.014***	0.002***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$SD(Z\hat{y}\|Controls)$	0.105	0.059	0.006	0.119	0.050	0.008
N (Grades 5-8)	490255	484751	490255	490015	484451	490015

Note: The coefficient results are based on 8th grade only, while the variance estimates are based on grades 5-8. The Kalman specification includes the quadratic of prior achievement and the VAM and IV specifications the cubic of prior achievement. The p-value is reported for F-test that all teacher judgement categories are zero. Standard errors are in parentheses: * $p < 0.05$,** $p < 0.01$,*** $p < 0.001$.

Table 5. Direct Effect of Neighborhood Income in Predicting End of 8th Grade Achievement

	Math			Reading		
	Conventional VAM	Kalman Filter	IV	Conventional VAM	Kalman Filter	IV
Neighborhood Median Income Decile						
1 (ref.)	---	---	---	---	---	---
2	0.012*	0.007	0.004	0.008	0.003	0.001
	(0.006)	(0.005)	(0.006)	(0.007)	(0.006)	(0.007)
3	0.010	0.005	0.003	0.014*	0.005	0.003
	(0.006)	(0.005)	(0.006)	(0.007)	(0.006)	(0.007)
4	0.010	0.006	0.000	0.009	0.002	-0.004
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
5	0.011	0.005	0.000	0.014*	0.008	-0.002
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
6	0.013*	0.009	-0.002	0.016*	0.002	0.000
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
7	0.013*	0.008	-0.004	0.021**	0.005	-0.003
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
8	0.030***	0.016**	0.008	0.032***	0.01	0.001
	(0.006)	(0.006)	(0.006)	(0.007)	(0.006)	(0.007)
9	0.040***	0.025***	0.010	0.041***	0.011	-0.001
	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.008)
10	0.043***	0.019**	-0.004	0.066***	0.013	0.002
	(0.007)	(0.006)	(0.007)	(0.008)	(0.007)	(0.008)
p-value for F-test	0.000	0.003	0.213	0.000	0.726	0.992
N (Grade 8)	165813	162708	165813	165749	162675	165749
$Var(X\hat{\beta}\|Controls)$	0.00013***	0.00006***	0.00003***	0.00015***	0.00003***	0.00002***
	(0.000000)	(0.000000)	(0.000000)	(0.000000)	(0.000000)	(0.000000)
$SD(X\hat{\beta}\|Controls)$	0.011510	0.007920	0.005400	0.012090	0.005590	0.004350
N (Grades 5-8)	662496	654192	662496	662086	653700	662086

Note: The income decile coefficients presented here are based on only 8th graders. The variance estimates are based on 5th-8th graders. The Kalman specification includes the quadratic of prior achievement and the IV and VAM specifications the cubic of prior achievement. The p-value reported is testing whether all neighborhood income deciles equal 0. Standard errors are in parentheses: * p < 0.05,** p < 0.01,*** p < 0.001.

Table 6. Sorting of Students by Teacher and School

Dependent variable:	Math			Reading		
	Conventional VAM	Kalman Filter	IV	Conventional VAM	Kalman Filter	IV
Index using neighborhood income deciles						
SD of random school effects	0.013***	0.009***	0.006***	0.012***	0.006***	0.003***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
SD of random teacher effects	0.005***	0.005***	0.005***	0.003***	0.004***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	662496	654192	662496	662086	653700	662086
Index using prior teacher judgements						
SD of random school effects	0.051***	0.043***	0.011***	0.057***	0.045***	0.007***
	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)
SD of random teacher effects	0.054***	0.033***	0.011***	0.068***	0.039***	0.013***
	(0.001)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
N	490255	484751	490255	490015	484451	490015
Prior teacher transitory effect (ψ)						
SD of random school effects	0.073***	0.053***	0.086***	0.061***	0.049***	0.064***
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
SD of random teacher effects	0.068***	0.064***	0.079***	0.052***	0.056***	0.057***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	1103000	1370145	1103000	1102699	1369889	1102699
Kalman prediction for baseline score						
SD of random school effects	-	0.306***	-	-	0.308***	-
	-	(0.006)	-	-	(0.006)	-
SD of random teacher effects	-	0.278***	-	-	0.279***	-
	-	(0.002)	-	-	(0.002)	-
N		1097264			1096999	
Kalman residual for baseline score						
SD of random school effects	-	0.000	-	-	0.000	-
	-	(0.000)	-	-	(0.000)	-
SD of random teacher effects	-	0.025***	-	-	0.010***	-
	-	(0.001)	-	-	(0.001)	-
N		1097264			1096999	

Note: The results are hierarchical random effects, estimated for grades 5-8. Sample sizes for the index results are smaller because of limited availability of income deciles and teacher judgement variables. The indices for neighborhood income and prior teacher judgements were derived using coefficients from value-added specifications. All model specifications are non-linear in the prior measure of achievement (cubic for VAM and IV, quadratic for Kalman). Standard errors are in parentheses: * $p < 0.05$,** $p < 0.01$,*** $p < 0.001$.

Table 7. Correlation in Teacher Effects from Conventional VAM, Kalman Filter and IV Specifications

			Standard Controls			Adding Neighborhood Income Deciles and Lagged Teacher Judgements		
			Conventional VAM	Kalman Filter	IV	Conventional VAM	Kalman Filter	IV
Math	Permanent Teacher Effect	Conv. VAM	1.000	-	-	0.955	-	-
		Kalman Filter	0.887	1.000		-	0.981	-
		IV	0.852	0.872	1.000	-	-	0.998
	Transitory Teacher Effect	Conv. VAM	1.000	-	-	0.964	-	-
		Kalman Filter	0.915	1.000	-	-	0.989	-
		IV	0.937	0.961	1.000	-	-	1.000
Reading	Permanent Teacher Effect	Conv. VAM	1.000	-	-	0.958	-	-
		Kalman Filter	0.815	1.000	-	-	0.991	-
		IV	0.762	0.817	1.000	-	-	0.998
	Transitory Teacher Effect	Conv. VAM	1.000	-	-	0.960	-	-
		Kalman Filter	0.829	1.000	-	-	0.994	-
		IV	0.877	0.913	1.000	-	-	1.000

Note: All model specifications are non-linear in the prior measure of achievement (cubic for VAM and IV, quadratic for Kalman). Correlations are computed across grades 5-7, the grades in which it is possible to estimate both transitory and teacher effects. The first three columns use all six student cohorts. The last three columns, which add neighborhood income and teacher judgement indicators to the specification, are limited to the three cohorts with available data.

Table 8: State Space Model Parameters

Component:		Math	Reading
Delta (averaged across grades)	$\hat{\delta}$	0.972	0.984
Measuring Achievement:			
Measured score variance	$Var(Y_{it})$	0.883	0.880
Measured error variance	$Var(\eta_{it})$	0.097	0.137
The gap between true and predictable baseline achievement:			
True score variance	$Var(\mu_{it})$	0.759	0.730
Kalman prediction variance	$Var(\hat{Y}_{it})$	0.629	0.619
Sources of persistent innovations:			
Student-level heterogeneity (unobserved)	$Var(\alpha_i)$	0.001	0.001
Student-level heterogeneity (assoc. with demographics/program use)	$E[Var(W\hat{\gamma}\|grade)]$	0.005	0.002
Persistent teacher effects	$Var(\theta_j)$	0.022	0.019
Other persistent innovations	$Var(v_{it})$	0.040	0.037
Transitory teacher effects:			
Transitory teacher effects	$Var(\psi_j)$	0.027	0.013
Correlation between persistent, transitory teacher effects	$Corr(\theta_j, \psi_j)$	-0.195	-0.456
Overall teacher effect	$Var(\theta_j + \psi_j)$	0.039	0.017
Implied teacher fadeout	$Cov(\theta_j, \theta_j + \psi_j) / Var(\theta_j + \psi_j)$	0.442	0.696
Correlation between total and persistent teacher effects	$Corr(\theta_j + \psi_j, \theta_j)$	0.590	0.660

Note: The IV specification uses the cubic in the lagged score of the same subject. All parameter estimates are estimated across grades 5-8 except for the variance in the permanent teacher effects, which are only available in grades 5-7, and the variance in the teacher transitory effects, which are only available in grades 4-7. The correlation in the teacher permanent and transitory effects is taken over grades 5-7. Kalman predictions use the quadratic in the prior prediction.

Table 9. Differences in Teacher Quality by Neighborhood Income

	Math				Reading			
	Teacher Effects No Controls		Teacher Effects With School FEs		Teacher Effects No Controls		Teacher Effects With School FEs	
Neigh. Income Decile	$\hat{\theta}_j$	$\hat{\psi}_j$	$\hat{\theta}_j$	$\hat{\psi}_j$	$\hat{\theta}_j$	$\hat{\psi}_j$	$\hat{\theta}_j$	$\hat{\psi}_j$
1 (ref.)	---	---	---	---	---	---	---	---
2	0.005**	0.016***	0.002	-.002*	-0.003*	-0.002	0.002	-0.002*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
3	0.006***	-0.005***	0.000	-0.001	-0.007***	0.001	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
4	0.005***	-0.012***	0.001	-0.001	-0.010***	0.008***	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
5	0.006***	-0.012***	0.001	-0.002	-0.007***	0.001	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
6	0.004**	-0.019***	0.001	-0.001	-0.017***	0.000	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
7	0.017***	-0.016***	0.001	0.001	-0.003*	-0.002*	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
8	0.020***	-0.021***	0.002	-0.001	0.001	-0.007***	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
9	0.038***	-0.029***	0.003**	0.000	0.001	-0.006***	0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
10	0.048***	-0.017***	0.003*	0.002	-0.010***	-0.003*	0.000	-0.002
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
N	495807	497168	495807	497168	495385	496912	495385	496912

Notes: The permanent and transitory teacher effect estimates come from an IV regression cubic in the lagged score of the same subject. The sample represented in this table includes grades 5-7, the grades in which we can estimate both permanent and transitory teacher effects with IV. Standard errors are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Appendix

Table A1. Comparing Linear and Non-Linear Specifications of VAM and Kalman Filter for 8th Grade Achievement

	Math		Reading	
	Linear	Non-Linear	Linear	Non-Linear
Achievement Measure	Conventional VAM			
L1 Math	X	X	X	X
L1 Reading	X	X	X	X
L1 Math^2		X		X
L1 Reading^2		X		X
L1 Math^3		X		X
L1 Reading^3		X		X
Adj. R^2	0.760	0.766	0.683	0.686
p-value for non-linear effects=0	-	0.000	-	0.000
N	275829	275829	275776	275776
	Kalman Filter			
L1 Math Kalman Prediction	X	X	X	X
L1 Math Kalman Residual	X	X	X	X
L1 Reading Kalman Prediction	X	X	X	X
L1 Reading Kalman Residual	X	X	X	X
L1 Math^2		X		X
L1 (Math x Math Kalman Residual)		X		X
L1 (Math^2 x Math Kalman Residual)		X		X
L1 Reading^2		X		X
L1 (Reading x Reading Kalman Residual)		X		X
L1 (Reading^2 x Reading Kalman Residual)		X		X
Adj. R^2	0.789	0.793	0.729	0.730
p-value for non-linear effects=0	-	0.000	-	0.000
N	272881	272881	272890	272890

Note: All specifications include indicators for current outcome subject teacher, lagged outcome subject teacher, year, race, gender, economic disadvantage, IEP and LEP status. Standard errors are in parentheses: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.