NATIONAL
CENTER for ANALYSIS of LONGITUDINAL DATA in EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Is It Just a Bad Class?

Assessing the Long-term Stability of Estimated Teacher Performance

DAN GOLDHABER
AND MICHAEL HANSEN

WORKING PAPER 73 • MARCH 2012

Is It Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance

Dan Goldhaber
Center for Education Data and Research
University of Washington

Michael Hansen
American Institutes for Research

Contents

Acknowledgements

The research presented here utilizes confidential data from the North Carolina Education Research Data Center (NCERDC) at Duke University, directed by Clara Muschkin and supported by the Spencer Foundation.

The authors wish to acknowledge the North Carolina Department of Public Instruction for its role in collecting this information. We gratefully acknowledge the Institute of Educational Studies at the Department of Education for providing financial support for this project. This paper has benefited from helpful comments from Cory Koedel, Austin Nichols, Daniel McCaffrey, Tim Sass, Jim Wyckoff, Hamp Lankford, Jesse Rothstein, and participants at the APPAM 2008 Fall Research Conference, the University of Virginia's 2008 Curry Education Research Lectureship Series, and the Economics Department Discussion Group (EDDG) Series at Western Washington University. We also wish to thank Joe Walch for research assistance and Carol Wallace and Jordan Chamberlain for editorial assistance.

This research was supported by the National Center for the Analysis of Longitudinal Data in Education Research (CALDER) funded through Grant R305A060018 to the American Institutes for Research from the Institute of Education Sciences, U.S. Department of Education.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street N.W., Washington, D.C. 20007
202-403-5796 • www.caldercenter.org

Is It Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance

Dan Goldhaber and Michael Hansen

CALDER Working Paper No. 73
March 2012

Abstract

In this paper we report on work estimating the stability of value-added estimates of teacher effects, an important area of investigation given public interest in workforce policies that implicitly assume effectiveness is a stable attribute within teachers. The results strongly reject the hypothesis that teacher performance is completely stable within teachers over long periods of time, but estimates suggest that a component of performance appears to persist within teachers, even over a ten-year panel. We also find that little of the changes in teacher effectiveness estimates within teachers can be explained by observable characteristics.

I. INRTODUCTION

Does a worker's current performance on the job forecast future productivity? Economic theory commonly assumes unobserved worker quality as an individual parameter that is fixed over time, yet empirical evidence supporting this assumption over the long run is sparse.[1] Evidence about the veracity of this assumption is policy relevant for numerous occupations as there are significant fixed costs associated with hiring and job termination, and it is a particularly timely and important issue in the public teaching profession. Education policymakers show a growing interest in raising the level of quality among the teacher workforce by using productivity measures to reward, select, and retain workers in the profession.[2] Though approaches to address teacher quality vary across the country, many attempts to manage teacher quality implicitly rest, in part, on the assumption that teacher quality is a stable attribute in teachers.[3] In other words: a good (or bad) teacher today will be a good (or bad) teacher tomorrow and further into the future.

If a teacher's individual performance turns out to be an extremely stable characteristic (as assumed in the literature), then performance measurement and selectively retaining teachers might best be used to cull the labor force, as is suggested by Gordon et al. (2006). Proposals to incentivize student learning are similarly crafted to mitigate adverse selection across the labor market by rewarding "good" teachers and discouraging "bad" teachers (Podgursky and Springer, 2007). Alternatively, if actual

[1] Models of imperfect information in the labor market including adverse selection (Akerlof, 1970; Greenwald, 1986) and moral hazard with hidden information and hidden action (Myerson, 1982) state unobserved worker quality varies as a given starting point for the model. Models of tournaments (Lazear and Rosen, 1981), career concerns (Hansen, 2009; Holmstrom, 1982), and a myriad of other explicit and implicit incentive models have unobservable fixed worker quality as a baseline assumption of the model (see Prendergast, 1999).

[2] The U.S. Department of Education's recent Race to the Top initiative offered large monetary incentives for states to adopt policies that rely on student test-based measures of teacher effectiveness. For more information on Race to the Top, see: http://www2.ed.gov/programs/racetothetop/index.html.

[3] For instance, Colorado and Tennessee have shifted to selectively retaining teachers at tenure based on past performance; see the *Denver Post* article, "In Bold Move, Colorado Alters Teacher Tenure Rules," Colleen Slevin, June 12, 2010, and the *Nashville Education Examiner* article, "Tennessee Legislature Passes Bill on Teacher Evaluation," Elandriel Lewis, January 19, 2010.

performance (or our measures of it) tends to be an unstable characteristic (i.e., reveals relatively little about future performance), then the usefulness of teacher-based accountability may be limited.[4]

The purpose of this paper is to investigate the long-term stability of teacher performance in the workforce over an extended period, and to investigate what may account for changes in performance within teachers. We use a unique dataset from North Carolina that allows us to match students to their individual teachers to estimate teachers' value-added input into student learning. These measures are used to track their performance over a relatively long period (ten years). The length of the panel allows us to focus on fundamental issues about the nature and stability of teacher performance over an extended time and factors that may influence changes in performance.

To provide some context, we first ask what would be a reasonable expectation for teacher performance stability over time. We inform our priors on teacher performance stability with evidence from workers in other industries—since teachers are simply workers engaging in a specific (albeit complex) task, we presume the stability of their performance over time is likely akin to worker stability in similar occupations. Organizational psychologists have documented the stability of performance with workers across a wide range of industries: student ratings of university professors showed correlation coefficients as high as 0.72 across semesters (Hanges et al., 1990); sewing machine operators' performance across a six-month interval correlate at 0.55 (Deadrick and Madigan, 1990); year-to-year correlations ranged from 0.3 to 0.4 on batting and earned-run averages for a group of baseball players (Hofmann et al., 1992); and similar yearly measures ranged from 0.26 to 0.44 for an objective productivity score for a sample of field-service personnel for a gas utility company (Hoffman et al.,

[4] The generally accepted principal-agent framework models worker output as a function of unobserved quality, effort, and error. Note that policies based on past performance aggregate these components, leaving any selection decisions vulnerable to Type-I and Type-II errors. Though some theoretical models (e.g. moral hazard or career concerns) separate ability from effort in an ex-post Bayesian framework (e.g. Greenwald, 1986; Holmstrom, 1982; Myerson, 1982), this distinction is not made in actual policy proposals. The approach we pursue in this paper, where we only parse teacher effectiveness from error, is consistent with common policy implementation.

1991).[5] This literature suggests performance instability in a worker is due to both changes in individuals (e.g. human capital accumulation, effort) and changes in the actual job to which the worker must constantly adapt (in the case of teachers this might be different students, curricula, leadership, etc.).

Other than college professors, many of the above comparisons may be seen as not terribly relevant in informing our prior on the stability of teacher performance since teaching is a multidimensional, highly complex job (Prendergast, 1999). More relevant, perhaps, are findings from a meta-analysis (Sturman et al., 2005) based on 22 studies of performance stability that estimates stability for objective performance measures (which describe value-added estimates) in a category of highly complex jobs (a categorization that includes teachers), which shows year-to-year performance correlations ranging from 0.33 to 0.40. Therefore, one might reasonably expect only modest correlations in teachers' yearly value-added estimates, based on workers in similar settings using similar measures.

What does prior empirical evidence suggest about the stability of performance for teachers specifically? The topic was first studied in the education literature decades ago, including papers from Rosenshine (1970) and Brophy (1973); however, the topic appears to have been ignored until recent years. With the emergence of a literature estimating teachers' value-added input into student test scores, the stability of estimated teacher performance has again become a subject of interest. Several papers in this literature have addressed the year-to-year stability of teacher effectiveness estimates (although generally as an aside), and find evidence showing teacher performance is certainly correlated over time, though the correlations are generally lower than 0.5 (e.g., Aaronson et al., 2007; Ballou, 2005; Koedel and Betts, 2007).[6]

[5] We searched at length for empirical evidence on the stability of worker performance in the economics literature, but could find none.

[6] Kane and Staiger's (2008) experimental validation of teacher effectiveness estimates stands in contrast to these other studies, suggesting year-to-year correlations of within-pair differences in teacher value-added estimates are not significantly different from 1, after adjusting for non-persistent variability in the measures. Their approach differs from other studies in the literature in two key ways. First, they look at the correlation of within-pair differences over time, not the correlation of value-added estimates across all teachers in both time periods. Second, their adjustment

McCaffrey et al. (2009) utilize a five-year panel of teachers and students from five large school districts in Florida in a study focused explicitly on the intertemporal stability of value-added teacher effectiveness estimates. The authors find the year-to-year correlation of value-added estimates for math teachers ranges from 0.2 to 0.5, and estimate 30 to 60 percent of the variation in effect estimates is due to transitory noise. McCaffrey et al. model teacher quality (net of the sampling error) as the sum of two parts: a permanent component that could be viewed as a fixed teacher quality endowment within teachers, and a transient component that varies randomly within teachers over time with fixed variance. Based on this model, the authors estimate much less than half of a teacher's estimated performance today persists indefinitely into the future.

The magnitude of the intertemporal correlations in the research cited above is viewed by some (e.g. Baker et al., 2010) to be too low to be useful for policy. However, the fact that the correlations are only "moderate" or "small", depending on one's perspective, should not be surprising given that value-added estimates based on just a few years of observation are inherently noisy. The instability is not likely to be due to noise alone as there is evidence that some differences in performance are associated with observable time-varying variables in teachers. The most salient is increased effectiveness from experience (Rockoff, 2004; Clotfelter et al., 2006). Variation in class size (Angrist and Lavy, 1999) or teacher absences (Miller et al., 2008) could also influence real changes in performance within teachers, but are commonly ignored because they occur simultaneously with the estimation of value added (for single-year estimates). Finally, recent research (Jackson, 2010; Jackson and Bruegmann, 2009) suggests teachers' performance is associated with the effectiveness of their peers and match between workers and jobs. How much these and other observable variables account for changes within teachers' performance over time, however, has not been investigated as a share of the total variation in observed performance as we do here.

---

technique effectively removes both measurement error and changes in teacher performance that do not persist. This adjustment technique is discussed in more detail below.

On balance, the results from the studies above indicate that teacher effectiveness estimates show persistence from year to year (correlations ranging from 0.2 to 0.5). What is not clear is how much of the year-to-year variation may be explained by real changes in teacher effectiveness versus measurement error alone. Moreover, the longest panel of data in these studies covers just five years, which limits researchers' ability to analyze the stability of teacher performance for a longer term. Indeed, no study has assessed whether current performance estimates predict future performance beyond more than a few years into the future[7]—a critical omission given policymakers' interest in using these measures to potentially make permanent decisions about teachers' employment. Thus, understanding the long-run stability of performance estimates, and what may be associated with their volatility over time, is crucial.

This paper contributes to the existing literature in several ways. First, this study uses a long panel of student-teacher matched data (twelve years), enabling us to evaluate stability over periods of up to ten years.[8] Based on our longer observation period, we develop a time-series model of teacher performance that estimates the variances of stability and variability in performance measures over an extended time (rather than a short-term range of a few years). Based on this model, we suggest much of a teacher's performance changes over time, though a component of teacher quality appears to be fixed (at least over the ten years of estimates we analyze). This suggests managing teacher quality across the workforce based on estimates of effectiveness could have some long-run effect on the overall quality in the teacher workforce, though is not as large as that estimated in prior studies in the teacher quality

[7] Aaronson et al. (2007), Ballou (2005), Koedel and Betts (2007), and McCaffrey et al. (2009) all address the stability of teacher value-added estimates by comparing adjacent-year estimates only. One approach to performing an empirical Bayes adjustment uses the correlation of performance over time to isolate the persistent component of teacher quality, which could include using correlations from pairings of estimates across greater intervals. Kane et al. (2006) use such a technique, and in a footnote say they found no evidence of weaker correlations across longer intervals of time. Aside from the results in this footnote, we are not aware of any studies that have specifically investigated the stability of teacher performance beyond adjacent years.

[8] As described in Section II, we use a student's full history of test scores (two years) to compute the teacher fixed effects. With twelve years of data, we are able to compute ten years of teacher effect estimates, spanning 1997–2006.

literature. Next, we investigate associations between explanatory variables and changes in estimated teacher performance within teachers.

We find only a small portion of the variation in performance within teachers is explained by observable factors such as teacher experience and absences; indeed, though estimated performance appears to be quite dynamic, we can explain only a small amount of it with observable covariates. Finally, we analyze the role of sorting in our value-added estimates and compute an upper-bound on the persistence of sorting bias observed in our data, and find evidence that any potential sorting decreases stability. In the next section, we describe our estimation approach for the value-added measures. Section III describes the data we employ, Section IV presents our analytic methods and findings, and Section V concludes.

II. ESTIMATING TEACHER PERFORMANCE

One approach to measuring teacher performance—*identifying an individual teacher's value-added contribution*—is more complicated than it sounds. This is well illustrated by a growing body of literature that critically examines the foundation for whether value-added models (VAMs) can obtain unbiased, causal estimates of teacher inputs into student learning (Ballou et al., 2004; Kane and Staiger, 2008; Koedel and Betts, forthcoming; McCaffrey et al., 2004; Rothstein, 2010; Todd and Wolpin, 2003). Many issues in estimating teacher VAMs are not resolved and are still being validated. [9]

One of the most common VAM approaches for estimating teacher effectiveness typically predicts growth in student test scores and includes a fixed effect for each teacher spanning multiple periods in the panel data to soak up the teacher-specific residual growth. By construction, this approach estimates an average teacher effect over that time period.[10] While such a measure is informative, it

[9] See Hanushek and Rivkin (2010) or Staiger and Rockoff (2011) for recent synopsis on the state of current research in value-added estimation.

[10] For the sections of the analysis that investigate multi-year effects, we first isolate all teachers observed teaching in all of the years within the multi-year span and estimate Equation 1 with the additional inclusion of year-specific fixed effects.

cannot inform us about changes within teachers over that time period. To get a more fine-grained picture of teacher performance, we estimate the following VAM, which provides a teacher-year effect estimate for each year a teacher is observed teaching in the classroom:

$$\mathbf{\Delta A} = \mathbf{A_{prior}}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{T}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \quad (1)$$

This VAM specification predicts current student gains, Δ*A*, as a function of: a vector representing a student's history of prior test scores (containing two lags in our case) in both math and reading, A_{prior}; a vector of student and family background characteristics (gender, race and ethnicity, learning disabilities, free or reduced-price lunch status, and parental education level), *X*; a teacher fixed effect (where *T* represents a vector of indicator variables for each teacher); and a random error, ε.[11] We only estimate VAM estimates for 5th grade teachers in order to incorporate two years of prior testing performance from grades 3 and 4. Note that this model is estimated separately by year, so teachers' performance measures across years are all from separate samples and are combined after estimation. Estimates from this model provide the basis for our investigation into the stability of teacher performance.[12]

Aside from our intentional omission of school fixed effects, this model is the VAM specification that Rothstein (2009) finds contains the least amount of bias under various sorting rules by incorporating as much information as possible about the student's past.[13] We omit school fixed effects because an explicit goal of this investigation is to compare teacher estimates over time, and most teachers change schools over the course of their careers.[14] The model also intentionally does not include student fixed effects, as some VAMs do. We do not include them because their inclusion inhibits

[11] We estimate with robust standard errors to account for possible heteroskedasticity, and adjust our resulting value-added estimates for measurement error in a second stage. This is described in detail below.

[12] All estimated teacher effects are de-meaned prior to making comparisons across years.

[13] Rothstein (2009) computed no bias in this VAM (which he labels as "VAM4") only under certain, restrictive sorting rules; most sorting rules Rothstein analyzed showed a positive level of bias and under some extreme sorting rules the bias was larger than the actual teacher effects. Regardless of the sorting rule employed (whether random conditioned on observables or based on unobservables), however, VAM4 showed a consistently smaller bias than any of the other VAMs he investigated.

[14] Excluding school fixed effects facilitates intertemporal comparisons, though an additional part of our analysis below investigates how much of the variation in estimated performance over time is due to school characteristics compounded in these teacher-by-year effects.

explanatory power and appears to introduce bias that is not evident in a model that includes student characteristics only (Kane and Staiger, 2008). Also, note we include neither teacher nor classroom covariates in the model as these are fixed within a teacher-year and cannot be identified separately from the teacher-year effect (further analysis presented in the Section IV attempts to use these teacher and classroom covariates to explain variation in value-added estimates over time). Finally, while we cannot claim that the estimates computed from this model will be unbiased estimates of causal teacher inputs, in Section IV we address how any residual sorting in our value-added estimates may bias our estimates of performance stability.

The analyses of the teacher-year effectiveness estimates ($\hat{\boldsymbol{\tau}}_{j,t}$) generated from Equation 1 above can be broken into two groups: the first series of analyses attempts to characterize the stability of these estimates over time and investigate how multi-year effectiveness estimates might affect observed stability; the second series attempts to determine what accounts for the observed fluctuation in these performance estimates. To facilitate the narrative of these analyses and our results, we present our analytic methods and our results for each investigation together in Section IV below.

III. DATA

The administrative data we use is from the North Carolina Department of Public Instruction (NCDPI), which is compiled and managed by Duke University's North Carolina Education Research Data Center (NCERDC). The data include information on student achievement on standardized tests in math and reading (in grades 3 through 5) that are administered as part of the North Carolina accountability system.[15] We utilize data for teachers and students from school years 1994-1995 through 2005-2006.[16]

[15] One issue that arises in VAM estimation is the possibility that estimates may be sensitive to test ceilings (Koedel and Betts, 2008). The data used here show little evidence of a test ceiling. For instance, the skewness of the distributions on test scores ranges between -0.392 and -0.177 in reading and -0.201 and 0.305 in math (skewness = 0 for symmetric distributions). The authors find minimum competency tests (skewness ranging from -2.08 to -1.60)

The North Carolina data does not explicitly match students to their classroom teachers; rather, it identifies the person administering the class's end-of-grade tests. At the elementary level, the majority of those administering the test are likely the classroom teacher; however, we also take several precautionary measures to reduce the possibility of inaccurately matching non-teacher proctors to students. First, we restrict our sample to those matches where the listed proctors have separate personnel file information and classroom assignments that are consistent with their teaching the specified grade and class for which they proctored the exam. Because we wish to use data from classes most representative of typical classroom situations, we further restrict the data sample to self-contained, non-specialty classes, and impose class size restrictions to no fewer than 10 students (for a reasonable level of inference) and no more than 29 students (the maximum for elementary classrooms in North Carolina). Finally, our analysis is restricted to 5th grade teachers, to incorporate two years of prior test scores in both math and reading in our estimation (students in NC are not tested prior to grade 3). Students with missing current or prior test scores are omitted from the sample, implicitly selecting a more stable segment of the student population.

These restrictions leave us a sample of 541,552 student-year observations and 28,931 teacher-year observations spanning ten years, representing 9,961 unique teachers. The median teacher in the sample is observed two times, and the average is nearly three observations. In Panel A of Table 1, we compare the unrestricted NCERDC data from all 5th grade students over the ten-year analysis window between the 1996–97 and 2005–06 school years against the restricted sample of students we use to compute teacher effectiveness estimates. Comparing these means shows some differences between the unrestricted data and our sample: fewer minority students are observed, fewer students are FRL

have the most consequential impacts on teacher effectiveness estimates. The impacts are limited in tests with only modestly skewed distributions like those in the North Carolina data.

[16] Over this period, North Carolina introduced various versions of the end-of-grade tests. We investigated whether our value-added measures showed any systematic difference around the introduction of the new tests, and we could find no evidence of an effect.

eligible, more students have parents with at least a bachelors degree, and scores in both math and reading are slightly above the standardized average for the grade. T-tests indicate this is not a random sample, as expected based on the inclusion restrictions described above (which selects a reasonably stable sample of students).

In Panel B of Table 1, we report descriptive statistics for all 5th grade teacher-year observations over the ten-year window compared with those in the sample. As shown, teachers in the sample are primarily white and female, over 25 percent hold a master's degree or higher, over 75 percent have an unrestricted teacher license (i.e. not an emergency, temporary, or provisional license), and average about ten years of experience. Descriptive demographic variables from the sample show little variation from the unrestricted data from which it is drawn. For the sample of teachers, we also report the variances of the raw and measurement-error adjusted teacher effectiveness estimates in both reading and math (details on the calculation of measurement-error are described further below). The magnitudes of these adjusted variances are consistent with prior estimates in the literature (see Hanushek and Rivkin, 2010). For the remainder of this paper, we present all of the results using value-added measures in math only.[17]

IV. RESULTS

A. Characterizing the Long-run Stability of Teacher Effectiveness Estimates

We begin our investigation by conducting several tests to evaluate whether *true* teacher performance is stable over time. To be fair, we do not know of any individuals or studies that explicitly argue that teacher effectiveness is truly stable over time; in fact, there is ample evidence to suggest otherwise. Yet, an operating assumption of using value-added estimates for teacher policies requires some level of stability in order to affect overall workforce quality. In this section, we seek to characterize

[17] All of the analyses reported here were also estimated using value-added estimates in reading. Overall, the results were qualitatively similar in support of dynamic changes in teacher performance over time that are locally more highly correlated than what is observed globally. The full set of results for reading is reported in a prior version of this paper, disseminated as a working paper (Goldhaber and Hansen, 2010).

the long-run stability of these teacher-year effect estimates to understand how effective such policies might be in practice.

Is teacher performance stable?

To aid in identifying real changes in signal apart from noise, we use Pearson's correlation coefficient ($\rho_{t,t+1} = \frac{\sigma_{t,t+1}}{\sigma_t \sigma_{t+1}}$) to measure the correlation of these estimates over time. Direct pairwise correlations on adjacent years of observations on teachers provide an estimate of the stability of performance within teachers; however, teacher effectiveness is measured with error in each year. Thus, the directly calculated correlation coefficient reflects instability in both performance and measurement. The correlation of real performance can be recovered in a fairly straightforward manner. Consider two successive measures of teacher effectiveness (we use $\hat{\tau}_{j,t}$ to represent the value-added estimate of teacher *j*'s performance in year *t* from Equation 1 above), where estimated performance is true performance and a random error:

$$\begin{aligned} \hat{\tau}_{j,t} &= \tau^0_{j,t} + \varepsilon_{j,t} \\ \hat{\tau}_{j,t+1} &= \tau^0_{j,t+1} + \varepsilon_{j,t+1} \end{aligned} \tag{2}$$

The correlation coefficient based on these two measures takes the following form:

$$corr(\hat{\tau}_{j,t}, \hat{\tau}_{j,t+1}) = \hat{\rho}_{t,t+1} = \frac{\mathrm{cov}(\tau^0_{j,t}, \tau^0_{j,t+1})}{\sqrt{\mathrm{var}(\tau^0_{j,t}) + \mathrm{var}(\varepsilon_{j,t})}\sqrt{\mathrm{var}(\tau^0_{j,t+1}) + \mathrm{var}(\varepsilon_{j,t+1})}} \tag{3}$$

Assuming the random errors in estimating teacher performance are not serially correlated, Equation 3 shows the correlation coefficient on estimated performance between two adjacent years isolates the covariance of actual performance in the numerator. The denominator represents the noisy estimates of performance in both time periods. By isolating the signal variance of performance in each

time period through removing the error variance, we calculate the correlation of latent (true) performance over time across teachers, which we denote $\widetilde{\rho}_{t,t+1}$:

$$\widetilde{\rho}_{t,t+1} = corr(\tau^0_{j,t}, \tau^0_{j,t+1}) = \frac{\text{cov}(\tau^0_{j,t}, \tau^0_{j,t+1})}{\sqrt{\text{var}(\tau^0_{j,t})}\sqrt{\text{var}(\tau^0_{j,t+1})}} \tag{4}$$

As detailed in Aaronson et al. (2007) and Rothstein (2010), we estimate the error variance in each period with the weighted mean of the standard errors across all fixed effects within a given sample-year.[18] Removing the error variance from the estimated effect variance leaves us with the adjusted variance of teacher quality; these adjusted variances in each period comprise the denominator. Note that using the estimated standard errors to estimate overall measurement error as we do here is also a common first step in performing an empirical Bayes adjustment on value-added estimates, which shrinks raw value-added estimates towards the population mean in proportion to the reliability of the estimate. An alternative, less common approach to performing the empirical Bayes adjustment uses the correlation of yearly teacher value-added estimates over time to estimate the noise (this method is used in Kane et al., 2006, and Kane and Staiger, 2008); implicitly this approach isolates only time-invariant components to teacher performance and discards yearly fluctuations in estimated performance as noise. As a result, these latter kinds of empirical Bayes adjustments will result in smaller teacher effect variances overall, with the caveat that the estimates are intended to capture permanent differences in teacher effectiveness (see Staiger and Rockoff, 2010).[19]

[18] We use heteroskedasticiy-robust standard errors of the fixed effects, which are estimated with the Stata user-written command `fese`, written by Austin Nichols (2008). In our data, heteroskedasticity-robust standard errors result in the estimate of measurement error variance that is largest (compared to either regular OLS or clustered standard errors); therefore, we choose to use these standard errors to provide an upper-bound estimate on measurement error. We conducted a small-scale simulation to determine which standard errors would be the best predictor of measurement error variance, and both OLS and robust standard errors had virtually equivalent mean squared errors; cluster-robust standard errors were markedly worse at predicting measurement error.

[19] A central assumption to this empirical Bayes approach is that teacher effectiveness is stable, net of yearly performance fluctuations. If teacher performance is dynamic beyond yearly fluctuations (as we provide evidence of below) this approach will likely overestimate the permanent component of teacher quality.

In the ten adjacent years of value-added estimates, we calculate the observed and latent correlation coefficients nine times, treating each adjacent-year pair as a different performance realization to see the extent to which the true performance is stable over time.[20] The correlations are presented in Table 2. The first column reports the correlation (and its confidence interval) over time when pooling across all years; the second and third columns report the minimum and maximum yearly correlations observed among all of the nine adjacent-year pairs in the data. Consistent with prior estimates (e.g., McCaffrey et al., 2009), the correlation of observed performance is significantly positive, but smaller than 1. Rows 4 through 6 report the same measures after removing the estimated error variance from the correlation coefficients, providing an estimate of the correlation of true performance ($\widetilde{\rho}_{t,t+1}$, see Equation 4). These correlations suggest actual performance is also significantly lower than 1, implying the instability in performance estimates over time is unlikely to be due to measurement error alone. [21]

While not surprising, the implications of these findings are important. Previous studies evaluating whether teacher value-added estimates are statistically reliable have commonly placed teacher-year estimates side-by-side in a transition table, as evidence of reliability in the estimates.[22] Criticisms of value-added measures, conversely, use the same "test-retest" comparison to demonstrate the poor reliability of estimates and argue against using such a blunt measure to evaluate teacher performance (Hill, 2009). Contrary to both views, however, measurement error may not be the only random piece from one year to the next; but teacher performance itself also appears to vary over the interval.

---

[20] Throughout the paper, the reported confidence intervals around the estimated correlation coefficients are calculated using the following equation: $\frac{\tanh\left[\sqrt{n-3}\operatorname{arctanh}(\hat{\rho})\pm 1.96\right]}{\sqrt{n-3}}$.

[21] Boyd et al. (2008) present the case that gain-score measurement error is larger than commonly assumed, and part of these errors may be shared at the classroom level (e.g., the hypothetical "barking dog"). We cannot distinguish classroom-level errors from yearly fluctuations in teacher performance with our data as they are observationally equivalent in settings where only one measurement is taken per year.

[22] See, for instance, Aaronson et al. (2007), Ballou (2005), and Koedel and Betts (2007).

The estimates presented here on short-run teacher stability are reasonably consistent with our prior distribution. Recall our prior, based on Sturman et al.'s (2005) meta-analysis, suggested correlations for objective performance measures at yearly intervals in a high-complexity job range from 0.33 to 0.40. Our calculated correlations in Table 2 show our correlations using raw math teacher-year effect estimates are notably higher than this range. The same study also estimates the correlations of actual performance (removing measurement error, analogous to $\widetilde{\rho}_{t,t+1}$ above) over a yearly interval range from 0.76 to 0.88; our calculations in Table 2 are at the lower end of this range.[23] Thus, the instability of teacher's performance over time is not at all unique to the teaching profession.

How stable is performance over a longer period?

We next seek to characterize how teacher performance estimates change over a long period of time, as observations over a longer period might inform whether and how teacher performance may evolve over time. To guide our investigation, we calculate the correlation coefficients of performance at greater intervals of time (rather than adjacent years only) in Figure 1, which places years between observation on the x-axis and the correlation coefficient on the y-axis. The observed correlations over these intervals are plotted with a solid line, which shows the correlations between performance estimates decrease over longer intervals of time (i.e., estimated performance is more highly correlated locally than it is globally). In other words, current estimates of performance are increasingly less predictive of future performance with each additional year out. Note, however, these correlations also

[23] While this study focuses specifically on VAMs, an objective estimate of productivity, we wish to note studies in this organizational psychology literature find subjective performance measures are slightly more stable within workers than objective measures, attributable to more reliability in the test instrument (Sturman et al., 2005). Studies that have compared principal or mentor ratings of teacher performance with VAM estimates (Jacob and Lefgren, 2008; Harris and Sass, 2009; Rockoff and Speroni, 2010) have shown some correlation between the measures, though both objective and subjective measures orthogonally predict future teacher performance. We are not aware of empirical evidence that has been presented on the stability of subjective measures of teacher performance over time, though early findings from the Gates Foundation's Measures of Effective Teaching Project find subjective student perception measures are more highly correlated across different class sections taught by the same teacher in a year, compared with value-added estimates on the same sections (The Gates Foundation, 2010). These subjective measures could feasibly be used in tandem with objective performance measures to reduce measurement error, which could in turn enhance estimated stability.

show a "long memory" in that they appear to level off and do not decay entirely. Figure 1 also graphs three other plausible time-series models of teacher quality: stable within-teacher performance, stable performance net of random yearly performance fluctuations, and geometric decay over time (a random walk). As shown, the simulation-based confidence intervals of all three models fail to capture the relationship observed in the data.[24] The primary models that have been used in the literature to this point (implicitly or explicitly) are the two that are stable over time.[25] Based on this figure, these two models clearly overstate the long-run stability in VAM estimates.

The observed pattern appears that it could be a hybrid of these various time-series processes. We propose a time-series model where current teacher quality is the sum of three components—a persistent component of teacher quality (φ_j), a dynamic component ($\gamma_{j,t}$), and a transient component ($\nu_{j,t}$):

$$\tau_{j,t} = \varphi_j + \gamma_{j,t} + \nu_{j,t} \tag{5}$$

Assume the dynamic component follows a first-order auto-regressive process (random walk):

$$\gamma_{j,t} = \beta\gamma_{j,t-1} + \eta_{j,t} \tag{6}$$

Embedding Equation 6 into Equation 5 and adding the measurement error associated with all value-added measures results in teacher effectiveness estimates of the following form:

$$\hat{\tau}_{j,t} = \varphi_j + \beta\gamma_{j,t-1} + \eta_{j,t} + \nu_{j,t} + \varepsilon_{j,t} \tag{7}$$

Note there are three sources of variation in performance estimates in Equation 7: innovations in teacher quality related to the dynamic component ($\eta_{j,t}$), innovations in quality that are purely transient

[24] The expected values and confidence intervals for the various time-series models depicted in Figure 1 were calculated by simulation: using the variance decomposition, 999 data samples (of 3,000 teachers each) were generated that conformed to each of the competing models.

[25] Hanushek (2009) and Gordon et al. (2006) implicitly use the stable model as the basis for their calculations of the effect of using value-added in policy and tenure decisions. McCaffrey et al. (2009) and Staiger and Rockoff (forthcoming) both use the model with stable performance net of random yearly fluctuations. Because none of the studies above present intertemporal correlations over extended time intervals (beyond adjacent-year correlations), we cannot determine whether or not a fading relationship exists in the data they analyze.

observed only in that period of measurement ($v_{j,t}$), and measurement error ($\varepsilon_{j,t}$). We assume these three sources of variability are all orthogonal. Though these three sources of variation may seem cumbersome, this approach accommodates the potential sources of shocks to the time series. For instance, innovations in the dynamic component of teacher effectiveness fades over time as more current innovations occur—one might think of this as professional development where it has an impact in the time period received, but the learned skills fade over time. The transient component is a difference in real performance that shifts randomly from one period to the next, but is only observed in that year—this could be "good chemistry" in a particular class or a teacher randomly being sick for a week during flu season. Finally, measurement error is the error associated with the test instrument itself.

The model in Equation 7 allows for current teacher quality estimates to predict future performance, but its predictive power fades with time (note Equation 7 projected to the $t+n$ period would include θ^n as the coefficient on the dynamic component, which converges to zero). As a result, the long-run relationship between current performance and that multiple years into the future is only as large as the component that is persistent within teachers.

We wish to quantify the magnitude of these different sources of variability in our model that would be consistent with the patterns we observe in the estimated teacher-year effects. To do this, we take the total variation and the estimated error variance (described above) as given, and fit values for beta and the variance of the stable, dynamic, and transient components that are consistent with the observed correlation. Parameter values are estimated by minimizing the sum of squared errors between the predicted and observed lines from this model.

There are only nine correlation points along which we are fitting our time-series model so this is only an approximation of the various components in the model. Further, we make some simplifying assumptions of stationarity in the time series. Caveats aside, these estimated parameters make for a

close approximation of the actual observed time series of, as depicted in Figure 2.[26] The resulting parameter estimates suggest the total variance of teacher performance estimates are approximately 29 percent stable, 34 percent dynamic, 15 percent transient, and 21 percent error. The permanent component of teacher effectiveness, by these estimates, is 0.132 student standard deviation units in math achievement. Note this estimate is notably smaller than the adjusted standard deviation reported in Table 1 (0.217). This is critical because the long-run effect of workforce policies that use VAMs to select or retain teachers will operate directly on this permanent component of teacher effectiveness; other parts of performance might be important for a time, but after a few years these other components are little more than noise.

How do multi-year estimates affect stability?

The analysis presented above is based on single-year estimates of performance in math; but many policymakers appear interested in using multiple years of estimated performance given the improvement in statistical power and the apparent reduction in sorting bias.[27] McCaffrey et al. (2009) and Staiger and Rockoff (2010) both consider how using multiple years may influence the reliability of the resulting teacher effect estimate.[28] Based on this time-series model derived in Equation 7, we can evaluate how multi-year estimates of teacher performance increase our ability to predict future

[26] The sample of teachers used in this analysis constitute an unbalanced panel, and one may worry that non-random sampling may drive the pattern observed in Figures 1 and 2. This is not the case: we isolated a subset of teachers that are observed in the sample for all 10 years, another subset for teachers observed in the same school for all 10 years, and another subset for teachers in the same school teaching students of similar backgrounds (based on t-tests using race, free lunch status, and prior test scores) for all 10 years. In all three teacher subsets, similar patterns of decreasing correlations over longer time intervals were observed. Supplementary details on estimating these time series parameters and the stable teacher sample are available in an appendix of Goldhaber and Hansen (2010).

[27] The improvement in power was first noted in Ballou (2005), who showed that less than a third of teachers had teacher effects significantly different (based on an alpha level of 0.10) from the average in math based on one year of performance; but using a three-year estimate, over half of all teachers had effects that were statistically different from the average. A recent study from Koedel and Betts (forthcoming) shows using multiple years in estimation better controls for year-to-year variation in the composition of teachers' classrooms that Rothstein (2010) calls bias, as multi-year estimates do not show evidence of bias..

[28] Note, however, both of these studies implicitly assume stable teacher effectiveness, net of yearly performance fluctuations.

classroom effectiveness. For instance, a VAM estimate based on *n* years of observation simply becomes an average effect, described with the following:

$$\hat{\tau}_{j,tn} = \frac{1}{n}\left(n\varphi_j + (\sum_{i=1}^{n}\beta^{i-1})\gamma_{j,t-n-1} + \sum_{i=1}^{n-1}(\sum_{k=1}^{i}\beta^{k-1})\eta_{j,t-i+1} + \sum_{i=1}^{n}\nu_{j,i} + \sum_{i=1}^{n}\varepsilon_{j,i}\right) \quad (8)$$

Assuming a stationary time series, the total variance of this *n*-year VAM estimate evaluates to:

$$\mathrm{var}(\hat{\tau}_{j,tn}) = \mathrm{var}(\varphi_j) + \frac{n + \sum_{i=1}^{n-1} 2i\beta^{n-i}}{n^2}\mathrm{var}(\gamma_{j,t}) + \frac{1}{n}\mathrm{var}(\nu_{j,t}) + \frac{1}{n}\mathrm{var}(\varepsilon_{j,t}) \quad (9)$$

This overall variance gets mechanically smaller as more years are considered in creating VAM estimates, though the variance of the stable component stays fixed. The implication of this result is that the stable component will take a progressively larger share of the total variance in multi-year VAMs, though at a decreasing marginal rate. This is demonstrated in Table 3, which reports some of the properties of these multi-year VAMs based on the parameter estimates from the estimated time-series model above. Column 1 reports the share of the stable component in the total variance of the *n*-year VAM estimates; this is equivalent to the *n*-year VAM's reliability in identifying *permanent* teacher effectiveness. As shown, the reliability of the permanent component increases from 0.29 with a one-year VAM to 0.52 with a six-year VAM.

While using additional years of observation in the VAMs shows a marked improvement in reliability, the extra years do not show such a clear improvement in the utility of VAMs to predict future performance. Column 2 shows the expected correlation coefficient between the *n*-year VAM estimate and the very next year of performance; in other words, the predictive validity for the next out-of-sample year. While adding years into the VAM does improve predictive validity, the marginal effect is relatively small. These moderate gains in predictive validity are the result of using past performance to predict future performance when performance is dynamic over the period. In this situation, relying on

additional information from the past to predict the future offers little traction and introduces a bias towards performance that does not persist.[29]

Column 3 represents the calculated predictive validity for the average performance of the next three out-of-sample years, which is higher across the board due to the (out-of-sample) three-year VAM's increased reliability on the permanent component. Note, however, the marginal improvements in predictive power are again highest when moving to the two- or three-year measures, and improve little with four or more years. Column 4 presents the correlation coefficients that are empirically observed in the data (along with confidence intervals), which are analogous to the three-year, out-of-sample predictions. Note the expected correlations presented in column 3 (derived analytically based on the estimated time-series parameters) are contained within the confidence intervals estimated empirically, lending some face validity to the estimated time series parameters described above.

The evidence presented here shows more years of teacher observation in VAM estimates do in fact improve the reliability of the permanent component of teacher quality, but do not improve our ability to predict future performance in a linear fashion. Instead, the gains in predictive validity are nonlinear and most are realized once three years of observation are incorporated into the VAM estimate.

B. Investigating the Sources of Instability in Teacher-Year Value-Added Estimates

The next series of analyses attempt to identify potential sources of instability in estimated teacher-year value-added estimates. We investigate this instability in three different ways: first, we test for statistically significant associations with other observable variables in the data that may be related to changes in performance; second, we use a quantile regression approach to address

[29] McCaffrey et al. (2009) note this tradeoff between precision and bias in multi-year VAMs when performance is shifting over the period. However, they do not offer any empirical evidence on the tradeoff between the marginal predictive validity and increasing bias with each additional year of past performance included in the VAM.

whether teacher performance is more or less stable at different points in the distribution; and finally, we analyze the role bias may play in affecting the stability of our performance estimates.[30]

Are changes in estimated performance associated with observable covariates?

To begin, we ask how observable changes in the teaching environment might be related with estimates of teacher performance. We hypothesize instability may arise from actual changes within the teacher (e.g. gains due to experience, changes in absences over time), or it may be a result of external changes in a teacher's job to which a teacher must adapt (e.g. changes in class size, the effectiveness of colleagues). We estimate the relative magnitudes of the variation in teacher-year estimates arising from these various sources. For this analysis, the vector of estimated teacher-year effects ($\hat{\boldsymbol{\tau}}$) are dependent variables regressed on teacher fixed effects (α) and time-varying school and classroom (S_s), peer ($P_{j,t}$) and teacher ($X_{j,t}$) characteristics.

$$\hat{\boldsymbol{\tau}} = \mathbf{T_j}\boldsymbol{\alpha} + \mathbf{S_s}\boldsymbol{\beta_1} + \mathbf{P_{j,t}}\boldsymbol{\beta_2} + \mathbf{X_{j,t}}\boldsymbol{\beta_3} + \boldsymbol{\nu} \tag{10}$$

Ordinary least-squares estimation of this relationship produce standard errors that are too small, because the dependent variable is an estimate. Instead, we use a generalized least squares approach that accounts for the uncertainty in the dependent variable by weighting observations in proportion to the reliability of each individual estimated teacher-year effect (Aaronson et al., 2007; Borjas and Sueyoshi, 1994; Koedel and Betts, 2007). The resulting estimates are presented in Table 4. The first column includes observable school and classroom-level variables including the percentage of minority students in the class, the percentage eligible for free or reduced-price lunch in the class, class

[30] Additionally, we investigated how the measurement error associated with class size contributes to this instability (omitted here for brevity). In particular, we selected a subsample of classrooms with 23 or more students and treated their value-added estimates using the full class as the "true" effect. Next, we ran a series of Monte Carlo simulations in which we randomly removed students from the classroom to create an increasingly smaller class to estimate the additional measurement error associated with the incrementally smaller class sizes. The motivating logic behind these simulations is it removes productivity differences that may due to a smaller class size from the incremental changes in measurement error. See Goldhaber and Hansen (2010) for more details.

size, and a variable indicating a new school principal. [31] The second column incorporates the teacher fixed effect, and the third column additionally incorporates a vector of teacher characteristics including experience, education, and absences as well as a vector of peer variables on experience, absences, and effectiveness (prior-year estimated effects of other teachers in same grade-school-year). Because we are concerned that including teachers with few yearly observations will attribute too much of the across-teacher variation to the teacher fixed effects, the sample is limited to teachers with at least four years of effectiveness estimates.[32] We interpret these results as descriptive (non-causal), and not necessarily representative of a definitive parsing of teacher effectiveness within teachers.

The school- and classroom-level variables in column 1 explain 1 percent of the overall variation in the estimated effects across teachers.[33] Keep in mind that the estimated measurement error in these estimates accounts for approximately 21 percent of the total estimated variance (implying the upper bound of the R-squared values should be near 0.79). Therefore, these school and classroom variables appear to explain no more than 2 percent of the variation in effectiveness estimates, net of error. Judging by the increase in the R-squared by including the teacher fixed effect in column 2, approximately 70 percent of the total variation in the teacher-year effect estimates (net of measurement error) is explained by individual heterogeneity. Note, however, that the student demographic variables are still significant, suggesting performance estimates vary *within teachers* due to

[31] While school fixed effects would be a better method to remove school-level differences, our sample includes teachers that were observed at least four times in the data and in many cases only one teacher from a given school is represented in the sample. Using school fixed effects in this case would confound teacher with school heterogeneity.

[32] We restricted the sample to teachers with at least four years of observations based on the evidence that teachers' performance changes over time and is less well correlated at longer intervals; using teachers with more observations attributes less variation in the estimated effects to individual heterogeneity. We also estimated this model with the sample of teachers with two or more observations and the inclusion of teacher fixed effects (in column 2) explained slightly more of the variation in estimates over time, as we expected.

[33] It is interesting to note that the academic literature is not consistent on whether teacher effectiveness estimates, based on multiple years of matched student-teacher data, include classroom level or school covariates. As it turns out, while some of these covariates are statistically significant, it appears to make little difference in the teacher effectiveness estimates themselves as the correlation between estimates based on two years that include the covariates in column 1 of Table 4 and those that do not include these covariates is over 0.97 in math instruction and 0.91 in reading instruction.

changes in the composition of the classroom; whether this is due to estimation bias or actual changes in effectiveness across different classrooms is unclear.

Column 3 presents the estimates when we add a vector time-varying teacher and peer variables to the model. Focusing on the point estimates in column 3, the prior estimated effectiveness of a teacher's peers (same school-grade-year) are significant predictors of within-teacher increases in performance, consistent with the peer effects identified in Jackson and Bruegmann (2009). We also include covariates on peer teacher absences, because shirking could arise in teachers when peers show higher levels of absence (Bradley et al., 2007); and we find evidence of a significant negative association between peer absences and a teacher's own performance, even keeping the teacher's own absences constant, suggestive of a potential change in behavior above and beyond increased absences.[34] A teacher's own absences, both anticipated (vacation and administrative leave) and unanticipated (sick and personal leave) also show a significant within-teacher association with decreases in estimated teacher effectiveness, consistent with prior studies on the issue (Clotfelter et al., 2009; Miller et al., 2008). And, as expected, gains in experience have a positive association with estimates, while gaining an advanced degree shows no significant effect.[35]

While many of these covariates are statistically significant predictors of changes in teacher performance, these included teacher and peer variables together explain an increment representing less than 2 percent of the total variation in teacher estimates (net of measurement error), or approximately 5 percent of the variation in estimated effectiveness *within* teachers (net of error). Presumably the remaining 95 percent of variation within teachers not explained by any of these factors

[34] Though the estimates in Table 5 are not causal, a causal link between high absenteeism among one's peer teachers and a teacher's performance (holding a teacher's own absences constant) could operate through at least two channels: the non-absent teacher's efforts could be diverted to assist substitutes or high peer absenteeism may signal a lax working environment in which a teacher may choose to shirk while on the job. Investigating the channels through which these peer effects may operate is beyond the scope of this paper.

[35] The pattern of coefficient estimates on experience are fairly consistent with those presented in Clotfelter et al. (2006), which also uses the North Carolina data; our estimate on getting an advanced degree, though not significantly different from zero, is also not statistically different from the -0.028 coefficient they report.

is due to variation in unobservable factors that can vary over time such as effort, motivation, class chemistry, or possibly bias. In addition, other sources of instability in teacher performance are plausible, but cannot be investigated in this analysis (given our data sample); for instance, Ost (2010) finds lower returns to experience among teachers who change grades they are teaching, and early findings from the Gates Foundation (2009) suggest greater performance stability among departmentalized teachers teaching the same course to different classes compared to those teaching multiple courses. In short, instability in teacher performance appears to arise from many sources; yet, like teacher quality itself, *changes* in teacher quality within teachers over time appears to be largely unobservable.

Does stability vary over the distribution of estimated performance?

We also wish to investigate whether performance for teachers at one end of the distribution is more stable that of teachers at the other end. Prior studies have suggested teachers at either the high or low end of the performance distribution may be more stable in their performance than those in the middle (Aaronson et al., 2007; Koedel and Betts, 2007), but did not formally investigate this finding. We do so here by modeling the current teacher performance estimate as a function of the lagged estimated effect in a quantile regression framework, where the estimated parameters minimize the sum of absolute deviations between the observed and predicted values, evaluated at the conditional means (see Koenker and Hallock, 2001). Observations are again weighted by the inverse of the estimates' standard errors to reflect our imprecision in estimated current teacher performance in the dependent variable.

The results are presented in Table 5, where each row (labeled 10^{th} percentile, etc.) represents the point in the distribution of prior-year estimated effectiveness where the conditional mean was evaluated. Column 1 reports the results of the quantile regression evaluated on all teachers in the data for which two adjacent-year observations exist; these columns show a small but very consistent monotonic increase in the stability of estimated teacher performance as teachers perform at the higher

end of the distribution. Specifically, the magnitude of the lagged value-added measure suggests that a low-performing teacher's classroom performance is less predictive of the following year's performance, relative to that of a high-performing teacher. The magnitude of these marginal increases in the estimated coefficients is small enough that when moving up an incremental decile (e.g. moving from the 30th to the 40th percentile) the difference is not significantly different. Only differences spanning at least 50 percent or more of the distribution are statistically different.

Part of the increasing stability over the distribution, however, may be driven by novice teachers: novice teachers have disproportionately lower performance estimates and are generally becoming more productive with experience (i.e. predictably larger year-to-year variation in performance). Thus, to assess whether it is the improvement of novice teachers driving our results, we re-estimate the quantile regression parameters among a subsample of teachers that have five or more years of experience, when prior literature suggests the marginal returns to experience decrease substantially (Clotfelter et al., 2006; Rivkin et al., 2005). The results for this sub-sample of teachers are reported in column 2. At every decile there is slightly greater estimated stability for the more experienced subset of teachers (column 2) than the full sample (column 1), but the monotonically increasing stability over the entire distribution reflects the general finding that stability in estimated performance increases with the level of estimated teacher effectiveness. Put another way, the best teachers are more predictable over time, and we would argue that the magnitude of the difference is consequential; the stability coefficient increases by 0.06, or roughly improving predictive power by 10 percent, when moving from the bottom to the top of the teacher performance distribution.

Does sorting potentially bias estimated stability?

The level of stability that we document above could potentially be misstated if the estimated VAM performance estimates are biased. For instance, a principal who rewards favored teachers with "better" (along dimensions unobserved in the data) classes every year will artificially increase the

stability of performance estimates by introducing a sorting bias that overstates stability over time. On the other hand, if a principal "takes turns" among teachers by assigning low-achieving students (again along unobserved dimensions) to a given teacher in one year and switches in the following year, then the VAM estimates for the teachers in that school will have sorting biases that switch directions each year. This unstable sorting in the VAM estimates will understate true stability (downward bias) of teacher performance in this case.

Throughout the analysis, we have made no adjustment for the unknown (but possible) presence of sorting bias in our estimates, due to the potential non-random processes that matches students to teachers (Rothstein, 2010).[36] We are careful to state, however, that we have no way to verify the actual level of bias in these estimates, which depends on the degree of sorting in the data between teachers and students on unobservable characteristics. Though we cannot definitively determine whether bias exists, we can understand how sorting may potentially bias our stability results through a few simple tests.

Consider a teacher-year effect estimate that is generated from the hybrid model we propose above, with the addition of an extra parameter representing this unobserved sorting ($\delta_{j,t}$):

$$\tau'_{j,t} = \varphi_j + \gamma_{j,t} + \nu_{j,t} + \delta_{j,t} + \varepsilon_{j,t} \tag{11}$$

Because this sorting parameter is a latent component of the data-generating process for teacher-year effect estimates, we cannot parse it out from the other components without making additional assumptions. But, we can determine the how sorting may influence our stability estimates by looking at the correlation of the resulting estimated teacher-year effects over time:

$$corr(\tau'_{j,t}, \tau'_{j,t+1}) = \rho'_{t,t+1} = \frac{\text{var}(\varphi_j) + \beta_1 \text{var}(\gamma_{j,t}) + \beta_2 \text{var}(\delta_{j,t})}{\sqrt{\text{var}(\tau'_{j,t})}\sqrt{\text{var}(\tau'_{j,t+1})}} \tag{12}$$

[36] Though we make no explicit steps to address bias to this point, the VAM estimates used for the analysis in this paper were generated using a model that is shown to be subject to the least amount of bias when compared to several competing models (Rothstein, 2009). See discussion in Section II above.

The first two terms in the numerator of Equation 12 are derived from the properties of the hybrid model shown in Equation 7. The difference in this case is the third term in the numerator ($\beta_2 \operatorname{var}(\delta_{j,t})$), which indicates the influence of this sorting parameter in calculating stability. If $\theta_2 > 0$, then sorting is positively correlated over time, which will bias our estimate of the stability of true teacher performance upwards; the reverse is also true. If $\theta_2 = 0$, this indicates sorting is not persistent and is observationally equivalent to random error over time. We can compute an upper-bound value of θ_2 in Equation 12 above by comparing the correlation coefficients of VAMs that are known to provide more biased estimates of performance.

VAMs that are known to suffer from sorting bias will increase the variance component of the third term, while only indirectly affecting the numerator. If the correlation coefficients from these more biased VAMs are greater (less) than those of the VAM we use in the analysis, this provides evidence about the value of θ_2.[37] Using the same sample of teachers, we compute teacher-year effect estimates based on two different VAMs: a gains model with a single lagged prior achievement score, and another that predicts gains without controlling for prior achievement. Rothstein (2009) argues these models produce VAM estimates that are necessarily more biased due to dynamic sorting that is not captured when omitting the lagged achievement scores.[38] Ordering all of the models by increasing sorting bias we have: first, our VAM used throughout this analysis (a gains model with two lags of prior achievement); second, the single-lagged achievement model; and third, the gains model without lags.

Figure 3 presents the correlation coefficients of our model (gains with 2 lags) against those from the more-biased estimates, calculated for the nine adjacent-year pairs in the data. The figure shows the year-to-year correlation in the estimated teacher-year effects universally decreases with more biased

[37] Increasing bias in the estimates has both a direct and indirect effect on the observed correlation coefficients. We therefore cannot unambiguously state whether β_2 is positive or negative, but we can compute an upper-bound value. See Appendix A for detail.

[38] The gains model that includes a single lagged prior achievement score is analogous to Rothstein's (2009) VAM3 model; the model without prior achievement is analogous to Rothstein's VAM2 model. Our departure from Rothstein is the intentional omission of the school fixed effects in our analysis (see Section II).

estimates, and the reduction in stability is largest for the model with the most bias (the gains model without lags).

The confidence intervals between the most and least biased estimates are mutually exclusive in most of the nine adjacent-year pairs. Given this statistically significant reduction in correlation coefficients when sorting bias in the estimates is increased, we know sorting bias cannot have a high level of persistence, if any. By setting the derivative of Equation 12 with respect to $\text{var}(\delta_{j,t})$ equal to zero, we compute the upper-bound value of θ_2. Even under some extreme (unlikely) levels of bias, θ_2 must be less than 0.171. This comparison shows any sorting bias in the value-added estimates is largely transient, and may even be compensatory over time (if the actual value of θ_2 is negative). This finding is consistent with the evidence presented in Koedel and Betts (2010), who show dynamic sorting bias cannot be detected in VAM estimates based on multiple years of observation. In light of this evidence, we conclude our estimates of positive correlations in teacher performance over time cannot be driven by sorting between teachers and students; given the possibility that θ_2 could even be negative, our estimates on the stability of teacher performance can be interpreted as a lower bound on true performance stability.

To take this investigation of sorting one step further, we conduct this same comparison of intertemporal stability with estimates that appear to show no detectable sorting pattern across classrooms (arguably less biased than our original sample). To do this, we isolate a subsample of teachers in our data that appear to be assigned students that are randomly sorted within schools at each grade level. The idea here is that in some schools the matching processes that govern the assignment of teachers and students to particular classrooms may result a sample of students that are

more or less randomly assigned to teachers, and, if true, this sample could be used to derive unbiased teacher effectiveness estimates.[39]

We do know the underlying processes that govern the matching of teachers and students to classrooms, but we can test for what appears to be random assignment based on observable student characteristics. We do this by conducting a chi-square test in each school-year-grade block to see whether the distributions of class averages of seven exogenous observable student characteristics – gender, ethnicity, limited English proficiency, eligibility for the free and reduced-price lunch program, parental education, and prior-year achievement in reading and in math – are comparable to the school-grade-level aggregate in that year (Clotfelter et al., 2006). Those schools where students appear to be distributed non-randomly across classrooms using one or more of these observable measures are removed from the sample, leaving a sample of schools that have classrooms that are balanced, at least along observable lines.

The results from the analysis of the subsample of "random schools" is shown in Figure 4, and it provides additional evidence that our estimates of stability are not strongly influenced by the within school sorting of teachers and students into classrooms. Specifically, there is little difference in the observed year-to-year estimated teacher effectiveness correlations between the full sample and the subsample of "random" schools.

V. CONCLUSION

This analysis ultimately seeks to inform policies that rely on the stability of teacher performance by investigating the observed stability of estimated effectiveness measures over time. We present evidence showing instabilities in estimated performance change over time within teachers, and these changes are unlikely to be due to measurement error alone. But importantly, while the observed time

[39] This process would not account for non-random matches of teachers and students to schools.

path of correlations over time with teachers suggests that some performance changes over time, some part of teacher quality does not change, even over a span of ten years. Because not all estimated performance persists beyond a few years, this permanent component of teacher quality is essentially the leverage point for workforce policies based on value-added estimates. Consequently, the effectiveness of these policies pivots on the extent that performance measures (subjective or objective) correctly identify this permanent value of teacher quality.

Must VAM estimates of performance be stable in order to be useful? We argue the answer is no. The imperfect information problem associated with teacher quality potentially engenders two sister market failures: adverse selection and moral hazard. Adverse selection deals with quality control across the whole labor market; considerable research in teacher quality has documented this problem (Hoxby and Leigh, 2004; Murnane et al., 1988; Corcoran et al., 2004). Workforce selection policies (e.g. using the estimates to retain teachers) are prescribed based on these findings of adverse selection. Stable teacher effectiveness estimates would enhance the expected effect of such policies, but there is a tradeoff in that they increase the risk of moral hazard. High stability in effectiveness estimates will reduce the expected productivity for exerting a marginal unit of effort in a given teacher, thus decreasing any motivational effect for policies that use VAMs as a way to incentivize teachers. Were policies constantly rewarding the same teachers, those left without rewards would quickly learn that they could be better off by withholding effort.[40] We thus recommend carefully approaching VAM-based policies, fully cognizant that efforts to target adverse selection could potentially be undermined through moral hazard.

It is also true that the source of instability of teacher performance estimates may critically inform policy. For instance, we noted in the introduction that recent research (Jackson and Bruegmann,

[40] Unfortunately, little research has been done on the moral hazard problem in teaching (i.e. teachers intentionally withholding effort because performance is not rewarded). See Hansen (2009) for evidence of teachers responding to career concerns (a multi-period moral hazard model).

2009) suggests that teachers' performance is associated with the effectiveness of their peers. Consequently, some instability in performance estimates may be associated with the match between teachers and schools (Jackson, 2010). This may not be relevant for policies that focus on rewarding or selecting teachers who remain (or are likely to remain) in a particular school. But the component of stability associated with the match between teacher and school is critically relevant for policies that would involve school switches, for instance a superintendent who might want to move high-performing teachers into low-performing schools.

Regardless of how they are used, the political calculus of using effectiveness estimates in teaching is complicated by the possibility that these estimates may be biased. We address this issue by calculating an upper-bound value of the influence of sorting bias in our stability estimates. The evidence in Section IV suggests any sorting in the value-added estimates appears to be largely transitory; in other words, the observed stability of these performance measures cannot be due to a sorting rule consistently applied to teachers.

Finally, we investigated the sources of instability in these measures; particularly looking for associations with observable school and teacher characteristics. We find evidence that teacher performance measures are associated with changes in school demographics, peer effectiveness, experience, and the absences of both teachers and their peers; however, much of the within-teacher variation in effectiveness over time is still not explained by these observable variables. In contrast, the majority of the total variation in teacher-year estimates over time is due to differences in individual teachers. Further, we presented evidence that performance stability is higher at the upper end of the teacher performance distribution (i.e. having a high performance estimate appears to be a more reliable signal of future productivity than a low performance estimate).

Two of our results presented here speak directly to policymakers. First, our findings suggest there is a permanent component of teacher quality that is stable in teachers over long periods of time—

implying workforce policies selecting teachers based on VAMs could effectively improve student achievement. But, importantly, our time-series model suggests the permanent component of performance is considerably smaller than that which is typically used to estimate workforce policy impacts, leading these studies to overestimate the long-run effects of these workforce selection policies. For instance, Staiger and Rockoff (forthcoming) estimate an improvement of 0.08 standard deviations of student achievement in math could be obtained through an aggressive screening process that only retained teachers shown to be above the labor market average after the first year of teaching.[41] Applying our smaller estimates of the reliability of the permanent component to their work, the long-run effect of such a policy would likely be between 0.05 to 0.06 standard deviation units (a reduction of between 25 to 38 percent).

Second, we also investigate how multi-year VAMs enhance the stability of teacher performance. Based on our estimated time-series model of teacher performance, we find the reliability of VAMs increases substantially with additional years of observation; however, most of the gains in the ability of multi-year VAMs to predict future performance are gained with the three-year VAM. The notion that incorporating more prior information into VAMs will give more reliable estimates of future performance is only true to a point, and incorporating too much prior information increases the risk of bias from performance that does not persist over time.

The results presented here are based on observed teacher performance in North Carolina over a twelve-year time period and may not generalize to different states using different testing instruments; however, the methods presented here provide tools for researchers to evaluate the time series of teacher performance in different contexts to determine how VAM estimates can provide the most information in identifying teacher effectiveness across the workforce.

[41] Effect estimates from Hanushek (2009) and McCaffrey et al. (2009) may likewise overstate the long-run stability of selectively retaining teachers by 25 percent or more.

References

Aaronson, D., Barrow, L., and Sanders, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25:95–135.

Abowd, J., Kramarz, F., and Margolis, D. (1999). High wage workers and high wage firms. *Econometrica* 67:251–333.

Akerlof, G. (1970). The market for "Lemons": Quality, uncertainty and the market mechanism. *Quarterly Journal of Economics* 84:488–500.

Angrist, J. and Lavy, V. (1999). Using Maimondes' Rule to estimate the effect of class size on scholastic achievement. *Quarterly Journal of Economics* 114:533–575.

Austin, J., Humphreys, L., and Hulin, C. (1989). Another view of dynamic criteria: A critical reanalysis of Barrett, Caldwell, and Alexander. *Personnel Psychology* 42:583–596.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. L., Linn, R. L., et al. (2010). Problems with the use of student test scores to evaluate teachers. Washington, DC: The Economic Policy Institute.

Ballou, D. (2005). Value-added assessment: Lessons from Tennessee. In *Value-added Models in Education: Theory and Application*, ed. R. Lissitz. Maple Grove, MN: JAM Press.

Ballou, D., Sanders, W. and Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Education and Behavioral Statistics* 29:37–65.

Borjas, G., and Sueyoshi, G. (1994). A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64:165–182.

Boyd, D., Grossman, P, Lankford, H., Loeb, S. and Wyckoff, J. (2008). Measuring effect sizes: the effect of measurement error. In *National Conference on Value-Added Modeling*. University of Wisconsin-Madison.

Bradley, S., Green, C. and Leeves, G. (2007). Worker absence and shirking: Evidence from matched teacher-school data. *Labour Economics* 14:319–334.

Brophy, J. (1973). Stability of teacher effectiveness. *American Educational Research Journal* 10:245–252.

Clotfelter, C., Ladd, H. and Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41:778–820.

———. (2009). Are teacher absences worth worrying about in the United States? *Education Finance and Policy* 4:115–149.

Corcoran, S., Evans, W. and Schwab, R. (2004). Women, the labor market, and the declining relative quality of teachers. *Journal of Policy Analysis and Management* 23:449–470.

Cullen, J. and Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In *Improving School Accountability*, ed. Timothy J. Gronberg and Dennis W. Jansen. Amsterdam: Elsevier Science.

Deadrick, D. and Madigan, R. (1990). Dynamic criteria revisited: A longitudinal study of performance stability and predictive validity. *Personnel Psychology* 43:717–744.

Figlio, D. (2006). Testing, crime, and punishment. *Journal of Public Economics* 90:837–851.

The Gates Foundation. (2010). Initial findings from the measures of effective teaching project. The Bill & Melinda Gates Foundation.

Goldhaber, D., and Hansen, M. (2010). Is it just a bad class? Assessing the stability of measured teacher performance. CEDR Working Paper 2010-3. University of Washington, Seattle, WA.

Gordon, R., Kane, T., and Staiger, D. (2006). Identifying effective teachers using performance on the job. Hamilton Project White Paper. Washington, D.C.: Brookings Institution.

Greenwald, B. (1986). Adverse selection in the labour market. *The Review of Economic Studies* 53:325–347.

Haney, W. (2000). The myth of the Texas miracle in education. *Educational Policy Analysis Archives* 8. Available at: http://epaa.asu.edu/ojs/article/view/432.

Hanges, P., Schneider, B., and Niles, K. (1990). Stability of performance: An interactionist perspective. *Journal of*

Applied Psychology 75:658–667.
Hansen, M. (2009). How career concerns influence public workers' effort: Evidence from the teacher labor market. CALDER Working Paper #40.
Hanushek, E. (2009). Teacher deselection. In *Creating a New Teaching Profession*, ed. Dan Goldhaber and Jane Hannaway. Washington, DC: Urban Institute Press.
Hanushek, E, Kane, T., and Rivkin, S. (2004). Why public schools lose teachers. *Journal of Human Resources* 39:326–354.
Hanushek, E. and Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review* 100:267–271.
Harris, D. and Sass, T. (2009). What makes for a good teacher and who can tell? CALDER Working Paper #30.
Hill, H. (2009). Evaluating value–added models: A validity argument approach. *Journal of Policy Analysis and Management* 28:700–709.
Hoffman, C., Nathan, B., and Holden, L. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self– versus supervisor ratings. *Personnel Psychology* 44:601–619.
Hofmann, D., Jacobs, R. and Baratta, J. (1993). Dynamic criteria and the measurement of change. *Journal of Applied Psychology* 78:194–204.
Hofmann, D., Jacobs, R., and Gerras, S. (1992). Mapping individual performance over time. *American Psychological Association* 77:185–195.
Holmström, B. (1982). Managerial incentive problems: A dynamic perspective. In *Essays in Economics and Management in Honor of Lars Wahlbeck*. Helsinki: Swedish School of Economics.
Hoxby, C. and Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review* 94:236–240.
Jackson, K. (2010). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. NBER Working Paper #w15990.
Jackson, K. and Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics* 1:85–108.
Jacob, B. (2004). Accountability, incentives, and behavior: The impact of high–stakes testing in the Chicago Public Schools. *Journal of Public Economics* 89:761–796.
Jacob, B. and Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics* 26:101–136.
Kane, T., Rockoff, J., and Staiger, D. (2006). What does certification tell us about teacher effectiveness? Evidence from New York City. Cambridge, MA: NBER.
Kane, T. and Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA: NBER.
Koedel, C. and Betts, J. (2007). Re-examining the role of teacher quality in the educational production function. San Diego, CA: University of Missouri.
———. (2008). Value-added to what? How a ceiling in the testing instrument influences value-added estimation. National Center on Performance Incentives Working Paper 2008-21.
———. (Forthcoming). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein Critique. *Education Finance and Policy*.
Koenker, R. and Hallock, K. (2001). Quantile regression. *Journal of Economic Perspectives* 15:143–156.
Lazear, E. and Rosen, S. (1981). Rank order tournaments as optimum labor contracts. *Journal of Political Economy* 89:841–864.
Lewis, E. (2010). Tennessee legislature passes bill on teacher evaluation. *Nashville Education Examiner*, January 19, 2010.
McCaffrey, D., Koretz, D., Lockwood, J.R., Louis, T., and Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* 29:67–101.
McCaffrey, D., Sass T., Lockwood, J.R., and Mihaly, K. (2009). The intertemporal variability of teacher effect

estimates. *Education Finance and Policy* 4:572–606.
Miller, R., Murnane, R. and Willett, J. (2008). Do teacher absences impact student achievement? Longitudinal evidence from one urban school district. *Educational Evaluation and Policy Analysis* 30:181–200.
Murnane, R., Singer, J., and Willett, J. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational Researcher* 17:22–30.
Myerson, R. (1982). Optimal coordination mechanisms in generalized principal-agent problems. *Journal of Mathematical Economics* 10:67–81.
Nichols, A. (2008). fese: user-written Stata command.
Podgursky, M., and Springer, M. (2007). Teacher performance pay: A review. *Journal of Policy Analysis and Management* 26:909–949.
Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature* 37:7–63.
Rockoff, J. (2004). The impact of individual teachers on students' achievement: Evidence from panel data. *American Economic Review* 94:247–252.
Rockoff, J. and Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review* 100:261–266.
Rosenshine, B. (1970). The stability of teacher effects upon student achievement. *Review of Educational Research* 40:647–662.
Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy* 4:537–571.
———. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics* 125:175–214.
Schmidt, F., Hunter, J., Outerbridge, A., Goff, S. (1988). Joint relation of experience and ability with job performance: Test of three hypotheses. *Journal of Applied Psychology* 73:46–57.
Sherry, A. (2007). Salary system luring faculty. *Denver Post*, June 18, 2007.
Slevin, C. (2010). In bold move, Colorado alters teacher tenure rules. *Denver Post*, June 12, 2010.
Staiger, D. and Rockoff, J. (Forthcoming). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*.
Sturman, M., Cheramie, R., and Cashen, L. (2005). The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology* 90:269–283.
Todd, P. and Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113:F3–F33.

Tables and Figures

Table 1. Descriptive Means and Standard Deviations

Panel A. Student Characteristics		
	Unrestricted	Sample
Female	0.490	0.501
	(0.500)	(0.500)
Black	0.295	0.283
	(0.455)	(0.451)
Hispanic	0.049	0.036
	(0.219)	(0.186)
Other Non-White	0.051	0.041
	(0.220)	(0.197)
Free Lunch Eligible	0.461	0.336
	(0.498)	(0.472)
Parents' Bachelor's Deg. Or Higher	0.159	0.166
	(0.366)	(0.372)
Standardized Reading*	0.000	0.086
	(1.000)	(0.962)
Standardized Math*	0.000	0.100
	(1.000)	(0.972)
Observations (grade 5 students)	1,029,259	541,552
Panel B. Teacher Characteristics		
Female	0.879	0.879
	(0.326)	(0.326)
Black	0.151	0.135
	(0.358)	(0.342)
Hispanic	0.005	0.004
	(0.068)	(0.065)
Other Non-White	0.011	0.008
	(0.104)	(0.091)
Master's Degree or Higher	0.272	0.260
	(0.445)	(0.439)
Approved NC Education Program	0.393	0.390
	(0.488)	(0.488)
Full Licensure	0.772	0.781
	(0.420)	(0.414)
Yrs. Of Experience	10.446	10.501
	(9.877)	(10.061)
Yearly Absences (Personal+Sick Leave)	9.350	8.947
	(9.770)	(9.429)
Observed SD of Teacher-Year Effects-Reading		0.147
Adjusted SD of Teacher-Year Effects-Reading		0.106
Observed SD of Teacher-Year Effects-Math		0.244
Adjusted SD of Teacher-Year Effects-Math		0.217
Observations (teachers)	22,871	9,961

Note: Standard deviations in parentheses.

Table 2. Adjacent-year Correlations of Teacher Performance Estimates in Math

	10 year min	Pooled	10 year max
Observed Correlation	0.49	0.55	0.58
Conf. Interval	[0.46-0.53]	[0.54-0.56]	[0.56-0.61]
Adjusted Correlation	0.60	0.69	0.76
Conf. Interval	[0.55-0.64]	[0.68-0.70]	[0.73-0.80]

Note: Pooled estimates calculated across all adjacent-year pairs observed in data (multiple observation pairs for teachers spanning three or more years). The 10-year minimum and max measures report the lowest and highest values observed among the nine adjacent-year pairs within the 10-year span (one pair of observations per teacher).

Table 3. Properties of Multi-year VAM Estimates: Math

		Calculated		Observed
	Reliability of Permanent Component	Corr. with Next Year's Performance	Corr. with Next 3 Year Avg. Performance	Corr. with Next 3 Year VAM Out of Sample
1-year VAM	0.292	0.548	0.590	0.583 [0.568-0.597]
2-year VAM	0.377	0.586	0.637	0.632 [0.616-0.647]
3-year VAM	0.427	0.590	0.647	0.645 [0.626-0.664]
4-year VAM	0.463	0.587	0.648	0.649 [0.625-0.672]
5-year VAM	0.493	0.580	0.645	0.657 [0.626-0.685]
6-year VAM	0.518	0.574	0.642	0.658 [0.615-0.698]

Note: The first three columns represent reliability and correlation coefficients calculated analytically based on parameter estimates from hybrid time-series model. The fourth column presents the correlation coefficient empirically observed in data.

Table 4: Observed and Unobserved Factors of Teacher-Year Effect Estimates

	School and Class Variables	Teacher Fixed Effect	Teacher Variables
	1	2	3
Class size	-0.004**	-0.005**	-0.005**
	-(0.001)	-(0.001)	-(0.001)
Class percentage of students FRL eligible	-0.079**	-0.015*	-0.022**
	(0.008)	(0.008)	(0.008)
Class percentage of minority students	0.031**	0.058**	0.059**
	(0.008)	(0.015)	(0.015)
New School Principal	-0.007	-0.002	-0.001
	(0.005)	(0.004)	(0.004)
Peer Experience			-0.001**
			(0.000)
Peer Teacher Effectiveness (Reading)			0.012
			(0.028)
Peer Teacher Effectiveness (Math)			0.171**
			(0.019)
Peer Sick & Personal Absences (x10)			-0.011**
			(0.004)
Peer Other Absences (x10)			-0.014*
			(0.007)
Teacher Exper. (1-2 Yrs.)			0.087**
			(0.011)
Teacher Exper. (3-5 Yrs.)			0.099**
			(0.011)
Teacher Exper. (6-12 Yrs.)			0.100**
			(0.012)
Teacher Exper. (13+ Yrs.)			0.114**
			(0.014)
Advanced Degree			-0.01
			(0.013)
Teacher Sick & Personal Absences (x10)			-0.012**
			(0.002)
Teacher Other Absences (x10)			-0.009*
			(0.004)
Teacher 1st year in new school			0.004
			(0.007)
Year Indicators	YES	YES	YES
Teacher Indicators		YES	YES
Teacher Variables			YES
Observations	18,130	18,130	18,130
R-squared	0.01	0.54	0.55

Note: *, ** $p \leq 0.05, 0.01$, respectively. Robust standard errors in parentheses. Sample includes all teachers with four or more years of teacher-year effect estimates. Estimation through generalized least squares where observations are weighted in proportion to the precision of the estimate used as the dependent variable.

Table 5. Stability Across Distribution of Estimated Teacher Effectiveness

Decile of Teacher Effectiveness	Full Sample	Experience ≥ 5
1st decile	0.555	0.560
	(0.013)	(0.014)
2nd decile	0.567	0.574
	(0.009)	(0.011)
3rd decile	0.580	0.586
	(0.008)	(0.009)
4th decile	0.579	0.583
	(0.008)	(0.008)
5th decile	0.581	0.585
	(0.008)	(0.009)
6th decile	0.589	0.592
	(0.008)	(0.009)
7th decile	0.596	0.597
	(0.009)	(0.010)
8th decile	0.603	0.610
	(0.009)	(0.010)
9th decile	0.615	0.621
	(0.014)	(0.014)

Note: Robust standard errors are displayed in parentheses. Estimates based on quantile regression in which observations weighted in proportion to the precision of the estimated dependent variable.

Figure 1.

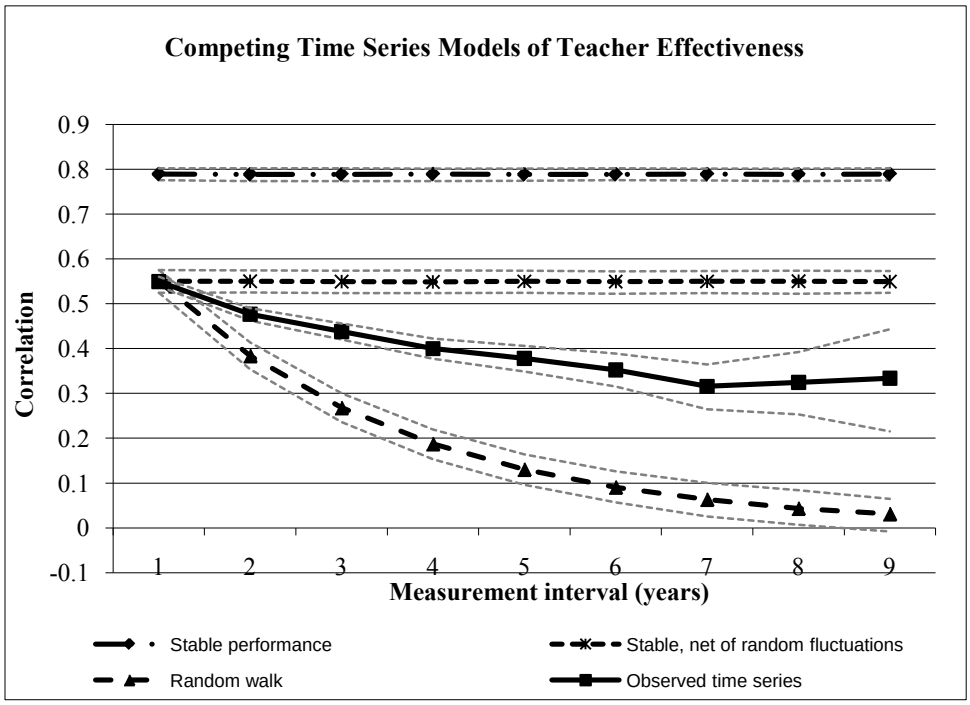


Figure 2.

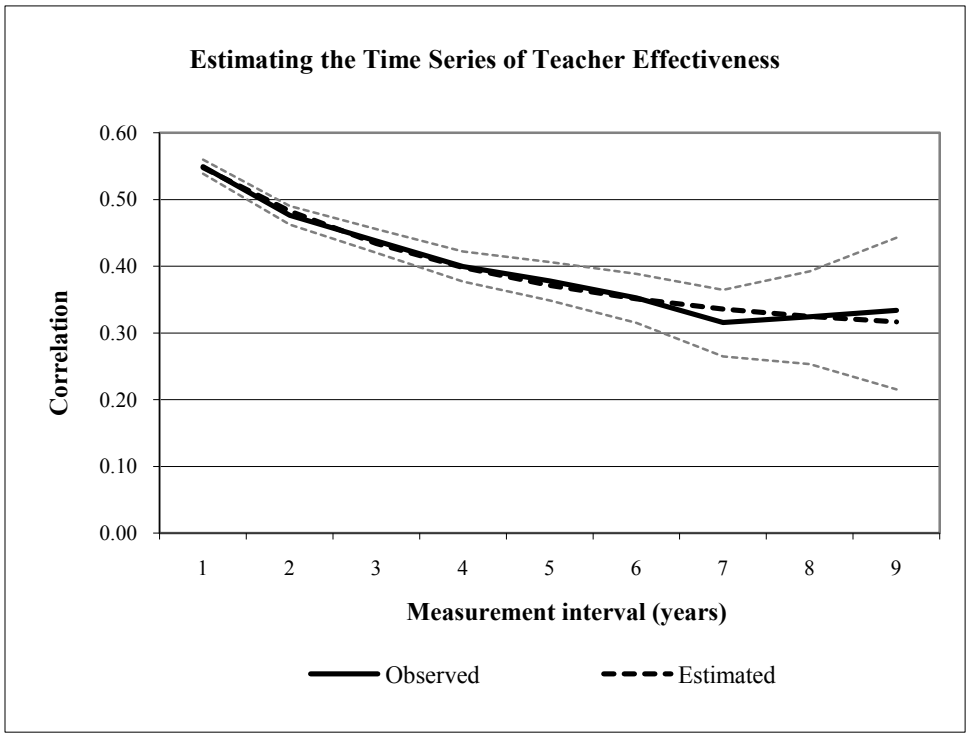


Figure 3.

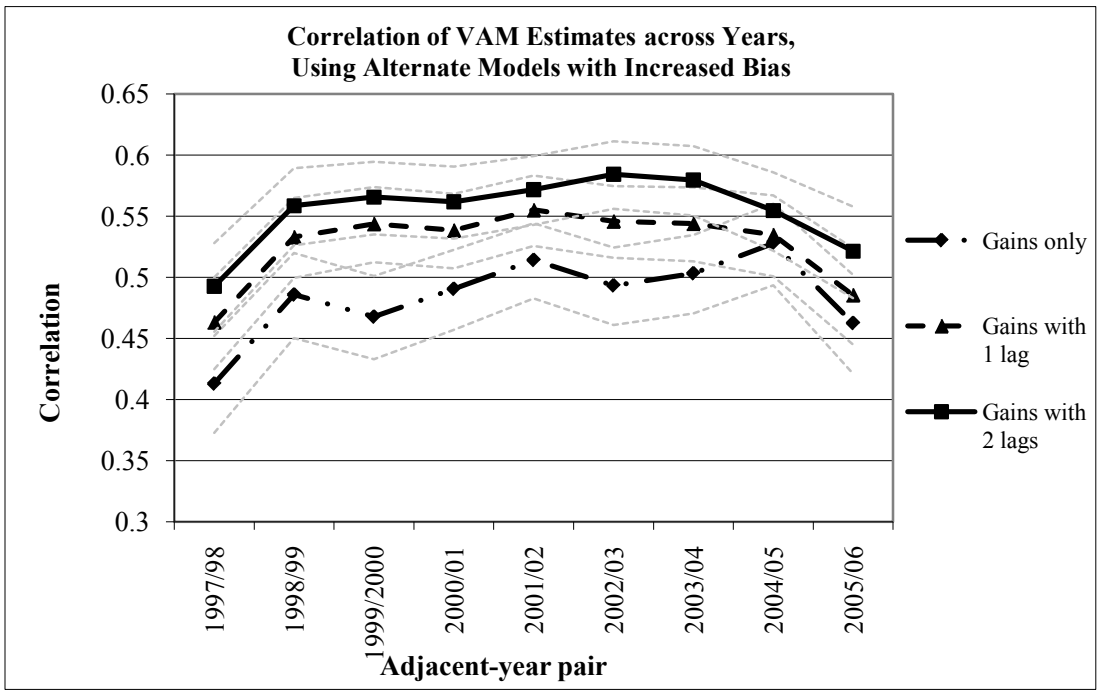
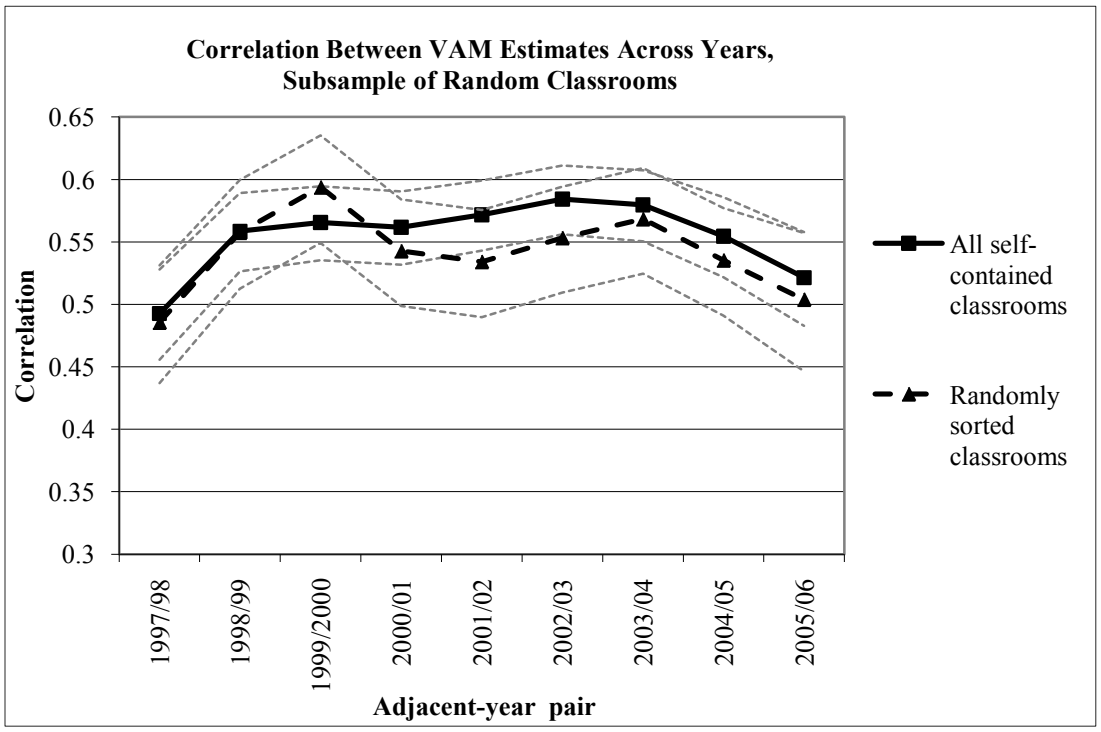


Figure 4.



Appendix

A. The Influence of Bias

In our discussion of the influence of bias on our estimates, we compare the correlation coefficients under various VAMs where the relative size of the bias is known. How the increase in the variance of the bias influences the correlation coefficients can be derived by applying the product rule to Equation 10, which results in the following:

$$\frac{\partial \rho'_{t,t+1}}{\partial \operatorname{var}(\delta_{j,t})} = \frac{\beta_2}{\sqrt{\operatorname{var}(\tau'_{j,t})}\sqrt{\operatorname{var}(\tau'_{j,t+1})}} - \left(\frac{1}{2}\right)\frac{\operatorname{var}(\varphi_j) + \beta_1 \operatorname{var}(\gamma_{j,t}) + \beta_2 \operatorname{var}(\delta_{j,t})}{\left(\sqrt{\operatorname{var}(\tau'_{j,t})}\right)^3\sqrt{\operatorname{var}(\tau'_{j,t+1})} + \sqrt{\operatorname{var}(\tau'_{j,t})}\left(\sqrt{\operatorname{var}(\tau'_{j,t+1})}\right)^3} \quad \text{(A5)}$$

Imposing the simplifying assumption of stationarity in the time series over time allows the derivative to be rendered as:

$$\frac{\partial \rho'_{t,t+1}}{\partial \operatorname{var}(\delta_{j,t})} = \frac{\beta_2}{\operatorname{var}(\tau'_j)} - \frac{\operatorname{var}(\varphi_j) + \beta_1 \operatorname{var}(\gamma_{j,t}) + \beta_2 \operatorname{var}(\delta_{j,t})}{2\left(\operatorname{var}(\tau'_j)\right)^2} \quad \text{(A6)}$$

The first term on the right-hand side is the direct effect of the bias in estimating stability; β_2 determines its influence on the correlation coefficient. The second term on the right-hand side is the indirect effect, which will be negative for the most plausible values of the parameters. Note that increasing the bias does not necessarily isolate the effect of β_2; even if $\beta_2 = 0$, we would expect a slight reduction in the correlation coefficient because it increases the noise in each period without affecting the signal (or alternatively, increasing the denominator of Equation 10 while holding the numerator constant).

Because of this indirect effect, we cannot unambiguously state whether β_2 is negative; however, we can calculate its upper-bound value. For there to be no observed change in the correlation coefficients, the direct effect must equal the indirect effect in magnitude, but with an opposite sign. Setting the derivative equal to zero, and isolating β_2 gives the following relationship:

$$\beta_2 = \frac{\operatorname{var}(\varphi_j) + \beta_1 \operatorname{var}(\gamma_{j,t})}{4\operatorname{var}(\tau'_j) - \operatorname{var}(\delta_{j,t})} \quad \text{(A7)}$$

This provides the upper-bound value of β_2; which is a function of the unknown variance of the bias in our original estimates. Given the parameters from the model, however, we can calculate the implied upper-bound value of β_2 under various magnitudes of the bias in our estimates. The values of the other variance components (aside from the bias component) are calculated using the

estimated parameter values presented in Appendix I of Goldhaber and Hansen (2010). We present the range of implied upper-bound values of β_2 in Table 6 below. Even in the extreme (and unlikely) case of our VAM estimates capturing as much bias they do signal, the upper-bound values of β_2 must be relatively small in order to show no increasing correlation when bias is added.

Table 6. Upper-bound Values of β_2

Math		
Variance of Bias	Percentage of Signal Variance	Upper-bound
0.000	0%	0.137
0.005	10%	0.140
0.009	20%	0.143
0.014	30%	0.145
0.019	40%	0.149
0.023	50%	0.152
0.028	60%	0.155
0.033	70%	0.159
0.038	80%	0.163
0.042	90%	0.166
0.047	100%	0.171

Given the evidence presented in the text, suggesting that the correlations are significantly lower under increased bias (not simply equivalent), we conclude β_2 must be lower than these values (and may potentially be negative).