CALDER

National
Center for Analysis of Longitudinal Data in Education Research

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

Urban Institute

A program of research by the Urban Institute with Duke University, Stanford University, University of Florida, University of Missouri-Columbia, University of Texas at Dallas, and University of Washington

Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions

Dan Goldhaber
and Michael Hansen

WORKING PAPER 31 • FEBRUARY 2010 (UPDATED VERSION)

Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions

Dan Goldhaber
Center on Reinventing Public Education
University of Washington

Michael Hansen
The Urban Institute
CALDER

Corresponding authors: Dan Goldhaber (dgoldhab@u.washington.edu), University of Washington University and Michael Hansen (mhansen@urban.org), National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute.

The authors are grateful to Philip Sylling and Stephanie Liddle for research assistance. This paper also benefitted from helpful comments and feedback by Dale Ballou, Cory Koedel, Hamilton Lankford, Austin Nichols, Jesse Rothstein, Daniel McCaffrey, Tim Sass, and Duncan Chaplin, and from helpful comments from participants at the APPAM 2009 Fall Research Conference, the University of Virginia's 2008 Curry Education Research Lectureship Series, and the 2008 Economics Department Seminar Series at Western Washington University. The research presented here is based primarily on 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 and making it available. The authors are also grateful for support from the National Center for Analysis of Longitudinal Data in Education Research (CALDER), supported by Grant R305A060018 to the Urban Institute from the Institute of Education Sciences, U.S. Department of Education.

CALDER, The Urban Institute
2100 M Street N.W., Washington, D.C. 20037
202-261-5739 • www.caldercenter.org

CONTENTS

Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions
Dan Goldhaber and Michael Hansen
CALDER Working Paper No. 31
February 2010

ABSTRACT

Reforming teacher tenure is an idea that appears to be gaining traction with the underlying assumption being that one can infer to a reasonable degree how well a teacher will perform over her career based on estimates of her early-career effectiveness. Here we explore the potential for using value-added models to estimate performance and inform tenure decisions. We find little evidence that the variation of teacher effects change over teacher careers, but strong evidence that prior year estimates of job performance predict student achievement, even when there is a multi-year lag between the two.

I. USING TEACHER EFFECTS ESTIMATES FOR HIGH-STAKES PERSONNEL DECISIONS

Well over a decade into the standards movement, the idea of holding schools accountable for results is now being pushed to a logical, if controversial, end point: the implementation of policies aimed at holding individual teachers (not just schools) accountable for results. An idea that has gained traction is the notion that some high-stakes personnel decisions ought to be based more on estimates of teacher outputs than on paper credentials like certification and degree level. In particular, *every state grants tenure* – which comes with considerable job protections – to teachers after a specified number of years, and evidence suggests that the tenure review process is not very rigorous.[1] Reforming teacher tenure is a policy option that appears to be gaining traction with the underlying assumption being that one can infer to a reasonable degree how well a teacher will perform over her career based on estimates of her early-career effectiveness (Hanushek, 2009; Gordon et al. 2006). Clearly this in turn presumes some degree of stability of job performance over time.

The focus on teachers and the stability of their job performance is supported by three important findings from teacher quality research. First, teacher quality, measured by value-added models (VAMs), is the most important school-based factor when it comes to improving student achievement. For example, Rivkin et al. (2005) and Rockoff (2004) estimate that a one standard deviation increase in teacher quality raises student achievement in reading and math by about 10 percent of a standard deviation – an achievement effect that is on the same order of magnitude as lowering class size by *10 to 13 students* (Rivkin et al. 2005).[2] Second, teacher quality appears to be a highly variable commodity. Studies typically find that less than 10 percent of the variation in teacher effectiveness can be attributed to readily observable credentials like degree and experience levels (e.g. Aaronson et al. 2007; Goldhaber et al. 1999;

[1] The median for states in granting tenure is three years (National Council on Teacher Quality, 2008).
[2] Other estimates of the effect size of teacher quality are even larger: Koedel and Betts (2007).

Goldhaber and Hansen, 2010; Koedel and Betts, 2007; Hanushek et al. 2005; McCaffrey et al. 2009).[3] Third, while the evidentiary base is thin, it appears that only a strikingly small percentage of teachers are ever dismissed (or non-renewed) as a consequence of documented poor performance.[4]

But while focusing accountability on individual teacher performance may seem sensible, it is easier said than done. Empirically derived estimates of teacher effectiveness (i.e. VAMs) involve making some strong assumptions about the nature of student learning (Todd and Wolpin, 2003). It is not entirely clear, for example, how teacher value-added effect estimates are influenced by the inclusion or exclusion of adjustments for differences in the backgrounds of a teacher's students, or the extent to which statisticians can adjust for the assignments of students and teachers to particular classes (Ballou, 2005b; Ballou et al. 2004; McCaffrey et al. 2004; Rothstein, 2009a). Moreover, researchers have shown that there is a substantial amount of "noise" resulting from test measurement error or the luck of the draw in students that is associated with measures of teacher effectiveness (we use the term "performance" interchangeably with "effectiveness" throughout) (McCaffrey et al. 2009).

In this paper we explore the potential for using VAM estimates as the primary criteria for rewarding teachers with tenure; a policy reform currently under consideration in several states.[5] The impacts of such a policy depends on at least three things: the distribution of teacher workforce quality over teacher careers; the stability of within-teacher job performance; and the extent to which early-career job performance serves as a signal of performance later in teacher careers. In this paper we present the results of an empirical examination of these three issues.

[3] For example, new research shows that even with comprehensive information on teachers – including measures of cognitive ability and content knowledge, personality traits, reported self-efficacy, and scores on teacher selection instruments – researchers can only explain a small proportion of the variation in teacher effectiveness. Specifically, Rockoff et al. (2008) find that the predicted value-added from a comprehensive set of teacher measures is just over 10 percent of the variance of the expected distribution of teacher effectiveness.

[4] Few very tenured teachers are ever fired (The New Teacher Project 2009). As an example, only 44 of over 100,000 Illinois' tenured teachers were dismissed from 1991 to 1997 (Goldstein 2001).

[5] See, for instance, The Tennessee Journal (2009), McGuinn (2010), Harris (2009), and The New York Times, http://www.nytimes.com/2009/11/26/education/26teachers.html?_r=2&ref=education.

Our findings are based on a unique dataset from North Carolina that allows us to match students – who are tested in math and reading on an annual basis – to their individual teachers. The relatively long panel (11 years) allows us to focus on fundamental issues about the nature and stability of teacher performance that inform a wide array of teacher policies that rely on the accuracy and stability of VAM job performance measures. We find statistically significant relationships between teachers' value-added effectiveness measures and the subsequent achievement of students in their classes. This suggests that VAM teacher effectiveness estimates provide information to policymakers relevant to consider for making personnel decisions like tenure.

II. BACKGROUND: VAMS AND THE STABILITY OF INDIVIDUAL TEACHER PERFORMANCE ESTIMATES

There is a growing body of literature that examines the implications of using value-added models (VAMs) in an attempt to identify causal impacts of schooling inputs, and indeed the contribution that individual teachers make toward student learning gains (e.g. Ballou et al. 2004; Kane and Staiger, 2008; Koedel and Betts, forthcoming; McCaffrey et al. 2004; Rothstein, 2009a, 2009b). Most of these studies focus on model specification and whether empirical estimates of teacher effects are unbiased. In summary of these studies, VAM estimates of teacher performance appear to be correlated with actual teacher quality, though it is not clear that these estimates are unbiased. If the VAM estimates we present are biased this would significantly reduce the implications of the results we present here. However, we perform various robustness tests which suggest similar results hold for subsamples of our data (or alternative specifications) where bias is less likely. This suggests that bias is not a significant driver of our findings.

More on point is the issue of intertemporal stability of estimated teacher effects, and only a handful of studies have addressed the issue. Aaronson et al. (2007), Ballou (2005), and Koedel and Betts

(2007) all generate estimates of teacher quality from different data sets, using different models, and using different numbers of years of observation to generate these estimates. The authors then assess the stability of teacher rankings (using either quartiles or quintiles) over time, and observe considerable numbers of teachers jumping between groups over time. In spite of the divergence in their approaches, however, all of these authors reject the hypothesis that cross-year rankings are random.

One study that focuses on indentifying the stable component of teacher quality is McCaffrey et al. (2009). The authors model VAM estimates as having three components: a persistent component of teacher quality that is fixed for each teacher, a transitory component of quality that is realized each year, and a random error term. Decomposing the variation in teacher quality this way implies only the persistent component of quality will be stable over time, and both the transitory component and the error are the noise in predicting future teacher performance. Using a 5-year panel of data, the authors find the year-to-year correlation of teacher effects of elementary school teachers in math ranges from 0.2 to 0.5 depending on model specification. However, they also find that multi-year estimates are considerably more stable: using a two-year average increase in the stability of estimates of a teacher's long-term average effect by about 50 percent relative to a single year measure, and adding a third year increases the stability by approximately an additional 20 percent.

McCaffrey et al. (2009) also devote a brief discussion to the implications of using estimated teacher effect estimates in the context of making decisions about which teachers receive tenure.[6] They estimate that if districts were to institute policies whereby teachers falling in the bottom two quintiles of *true* effectiveness were precluded from receiving tenure, then the average effectiveness level of the teacher workforce would increase by about 4 percent of a standard deviation of student achievement on standardized tests. Moreover, the overall improvement, which they deem to be small, is little affected by

[6] Note that most of this discussion is in their technical appendix.

the fact that teacher effects are measured with error. For example, were the tenure decision to be based on a two year average of *estimated* teacher effectiveness the effectiveness level of the workforce would improve by slightly less, about 3 percent of a standard deviation.

McCaffrey et al.'s (2009) estimates on the teacher workforce are based on a three-component model, however, analysis by Goldhaber and Hansen (2010) suggests this model may be misspecified.[7] Using data from elementary teachers in North Carolina, they show teacher effectiveness estimates have a "long memory" when correlating estimates across increasing intervals—the observed decay in the correlation coefficients is significantly slower than geometric decay from a random walk and rejects the hypothesis of no decay (this no-decay hypothesis is the approach McCaffrey et al. use). Goldhaber and Hansen instead adopt a model in which teacher quality is composed of a persistent component, an auto-regressive transient component, and a random error term. While this is only a small change in the model, it is an important change as it allows for teacher quality to change within teachers in ways that are consistent with the observed path of teacher quality estimates over time. Specifically, it allows time-specific innovations in teacher quality (through professional development, productive effort, etc.) to persist into future time periods; however, the magnitude of the effect gets progressively smaller with time. We address this further in the following sections.

On balance, the results from the above studies indicate that teacher quality estimates show some degree of persistence from year to year, but hardly an overwhelming amount, and it is unclear the degree to which the estimates may be contaminated by the inability of VAMs to fully account for the match between teachers and students. As we describe below, we have a much longer panel of matched teacher

[7] McCaffrey et al. (2009) note an attempt to model changes in teacher effectiveness over time with a drift component (i.e. teacher quality has a component that shifts with some variation from year to year rather than staying constant over time), and report the data fit the constant model better. They, however, do not report correlations on teacher effectiveness measures between larger intervals other than adjacent years—which would reveal whether that assumption of the model is valid.

and student data than has previously been used for analyses of VAMs. This longer panel allows us to investigate the changes in the stability of teacher estimates over a longer time frame, assess the stability of multiple year estimates of teacher effects, and examine the degree to which early career estimates of teacher effects predict the achievement of students taught later in a teacher's career, specifically pre- and post-tenure. All of these investigations inform the feasibility of using VAM estimates in making tenure decisions for public school teachers.

III. DATA AND ANALYTIC APPROACH

Data

In order to assess the stability of estimated teacher performance over time, it is necessary to have data that links students to their teachers and tracks them longitudinally. [8] The data we utilize is collected by the North Carolina Department of Public Instruction (NCDPI) and managed by Duke University's North Carolina Education Research Data Center (NCERDC). These data, based on administrative records of all teachers and students in North Carolina, include information on student performance on standardized tests in math and reading (in grades 3 through 8) that are administered as part of the North Carolina accountability system.[9] We utilize data for teachers and students from school years 1995-1996 through 2005-2006.

Unfortunately, the North Carolina data does not explicitly match students to their classroom teachers. It does, however, identify, the person who administered each student's end-of-grade tests, and

[8] Recent research illustrates how these data can be used for analyzing the effects of schools and teachers on students (Clotfelter et al. 2006; Goldhaber and Anthony 2007; Goldhaber 2006a, 2006b; Rothstein 2009a, 2009b).

[9] One issue that arises in the context of using VAMs to estimate teacher effects is the possibility that value-added teacher effectiveness estimates may be sensitive to ceilings in the testing instrument (Koedel and Betts 2008). Our data show little evidence of a test ceiling, so we do not feel it should pose a problem in our estimation. 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 have skewness measures ranging from -2.08 to -1.60, and these have the most consequential impacts on teacher effectiveness estimates and rankings. The impacts are fairly small in tests with only moderately skewed distributions, such as the tests we use here.

at the elementary level there is good reason to believe that those individuals administering the test are generally the classroom teachers. We utilize this listed proctor as a student's classroom teacher, but also take several precautionary measures to reduce the possibility of inaccurately matching non-teacher test administrators 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 them teaching the specified grade and class for which they proctored the exam. Because we wish to use data from classes that are most representative of typical classroom situations, we also restrict the data sample to self-contained, non-specialty classes, and impose class size restrictions to no fewer than 10 students (to obtain a reasonable level of inference in our teacher effectiveness estimates) and no more than 29 students (the maximum for elementary classrooms in North Carolina). Finally, we restrict our analyses to 4th and 5th grade teachers, because these classroom arrangements are most common in the elementary grades (students are not tested prior to grade 3 and the VAM we employ requires prior testing information).

These restrictions leave us a sample of 19,586 unique teachers and 62,588 unique teacher-year observations spanning 11 years (most teachers are observed more than once in the data). For part of our analysis, we will use a subset of the data in which we can identify teachers for multiple periods both before and after receiving tenure. This subset of the data is limited to 4th and 5th grade teachers for whom we observe (at minimum) the first 2 years of teaching in a district before becoming eligible for tenure, and at least one year after tenure. These stipulations provide us with a subset of 556 unique teachers, and 3,442 unique teacher-year observations. Throughout our analysis, we use various sub-samples of the restricted dataset described above; we describe inclusion criteria for the various sub-samples where relevant.

In Panel A of Table 1, we compare the unrestricted NCERDC data from all 4th and 5th grade students against the restricted sample of students we use to compute teacher effectiveness estimates, and

the group of students for whom we have effectiveness estimates and at least one year of data in which teachers are tenured. The observations reported represent unique student 4th and 5th grade student observations.

Comparison of the means shows some slight differences between the unrestricted data and the sample used for the analysis: in our sample, fewer minority students are observed, fewer students are FRL 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; however, these differences are expected, as inclusion in the sample requires valid sequential observations and therefore implicitly selects a relatively stable group out of the student data.

In Panel B of Table 1, we report descriptive statistics for teachers in 2006 (the last year in our sample), which is approximately representative of cross-sectional means over other years in the sample. As shown, teachers are primarily white and female. In terms of credentials, a minority holds master's degrees or higher or certifications from an approved North Carolina education program; a far higher proportion of the sample are fully licensed (that is, those teachers not holding a temporary or provisional license). The percentiles represent the one-year value-added teacher effect for teachers in each subject (the units are standard deviation units of student achievement in reading or math, and the method we used to estimate these effects are described in the next subsection of the paper). Comparison of the magnitudes of these effect estimates shows a considerably higher variance in the distribution of teacher quality in math relative to reading.

Value-added Measures of Teacher Effectiveness

A common modeling approach used in the VAM literature estimates a teacher fixed effect based on multiple years of observation, using observed classroom and school characteristics as controls. If one is

willing to assume that teacher effectiveness does not change within a teacher over time, then such an approach would provide an estimate of that teacher's future effectiveness.[10] This study, however, does not impose such a strong assumption; our reasons for this are two fold. First, Goldhaber and Hansen (2010) analyze the correlation of teacher effectiveness measures across increasing intervals of time and find these measures decrease as the time between measurement increases; and second, the policy motive of potentially attaching high stakes to value added estimates rests (in part) on the presumption that teachers will respond to these incentives, thereby changing performance over time. Thus, we allow teacher effectiveness to vary in each time period by using the following model:[11]

$$A_{ijkst} = \alpha A_{is(t-1)} + X_{it}\gamma + \tau_{jt} + \varepsilon_{ijkst} \qquad (1)$$

In this equation, current student learning (A_{ijkst}) is a function of students' lagged learning outcomes in both subjects ($A_{is(t-1)}$), observable characteristics (X_{it}), and a teacher-specific input (τ_{jt}).[12] The value-added of a teacher is estimated through using fixed effects methods to obtain these teacher-specific parameter estimates ($\hat{\tau}_{jt}$).

Equation (1), when estimated separately for each grade and year, imposes no inter-temporal restrictions on teacher quality. This flexibility, however, comes at the cost of some potentially important identifying information: the teacher-year effect is now confounded with classroom and school contributions to student learning.[13] Further, any bias in the estimates due to principals' allocation of students across classrooms will be captured in these estimates, however, as we describe below, various

[10] Though this assumption seems innocuous enough, it ignores changes within a teacher over time—some of which may be observable (i.e. returns to additional experience) and some may be unobservable (changes in effort levels).
[11] While some models (see Hanushek and Rivkin (forthcoming 2010), for a review), include school fixed effects, we opt not to. However, as we report below, our main findings are largely unaffected by their inclusion.
[12] Observable student characteristics include sex, race, ethnicity, free/reduced price lunch and limited English proficiency status, parents' education, and base-year test scores in math and reading.
[13] Recent research suggests there is also great variation in principal effectiveness, which could potentially be captured in these estimates. See Branch et al. (2009) and Clark et al. (2009).

tests of robustness suggest that our findings on the predictive ability of early career teacher performance to predict student achievement are not driven by biased effect estimates.

Research on VAMs have investigated alternative VAM specifications, often using a student fixed effects approach to control for time invariant characteristics in students (in theory this approach removes the influence of nonrandom sorting of students across schools and teachers that is based on time-invariant student factors). We do not pursue this approach as our primary VAM specification for two primary reasons. First, Rothstein (2009a) shows this approach is not necessarily robust to "dynamic sorting," i.e. the match of students to teachers based on unobserved attributes that are time-varying. Second, models that use student fixed effects generally have low power on the estimation of the student fixed effects themselves (due to data limitations from observing students in just a few years), and tests of the joint hypothesis that all student effects are non-zero commonly fail to reject. This implies more efficient estimation is possible through dropping the student-level effects (or pursuing a random effect strategy). Moreover, a recent paper by Kane and Staiger (2008) not only finds that a student fixed effects specification understates teacher effects in a VAM levels specification (where the dependent variable is the achievement of students in year t and the model includes a control for prior student achievement in year t-1), but also that the student fixed effects were insignificant in a VAM gains specification (where the dependent variable is the gain in student achievement from one year to the next). By contrast, they find a specification that includes a vector of prior student achievement measures produces teacher effects quite similar to those produced under conditions where teachers are randomly matched to their classrooms.

Analyzing Teacher Effectiveness Estimates for Tenure Decisions

A primary contribution of this paper is its investigation of using VAM estimates in making tenure decisions for teachers. We present evidence from three specific inquiries; the methods of each line of inquiry are outlined below.

First, rewarding teachers with tenure is a one-time decision that remains in force for the remainder of a teacher's employment with a school district, which in many cases may be the duration of a teacher's entire career. Thus, while investigating tenure decisions, we feel compelled to take a descriptive look at the variation in teacher effectiveness estimates over a teacher's career. Many studies have investigated the variance in teacher quality over the workforce (e.g. Hanushek et al. 2005; Aaronsen et al. 2007) and have investigated how mean performance changes with a teacher's experience in teaching (e.g. Rockoff 2004); however, no study has investigated how *the variation in estimated teacher quality* changes with experience.

Whether there is a convergence or divergence of effectiveness over a teacher's career likely influences the efficacy of any tenure policy adopted. For instance, if there is a high-degree of convergence it may not make sense to use VAM in the context of tenure decisions as teachers downstream would end up with performance closely bunched around some mean level. On the other hand, should there be a divergence of effectiveness over time using VAM effects to inform tenure might be even more important; e.g. if those teachers who are poor performers early in their careers are likely to be even worse, relative to the mean, as they progress through their careers.

To estimate the variation in teacher quality over a career, teachers are binned by experience and, within each experience bin, the adjusted variance of teacher quality in the workforce is calculated in both reading and math by netting out the measurement error, as is common in the teacher quality literature

(e.g. Koedel and Betts 2007; Rothstein 2009a). Additionally, because experience in a district or school may be similarly important, we make comparisons along these dimensions as well.

The second line of inquiry we pursue investigates the correlation of VAM estimates at increasing time intervals. We adopt Goldhaber and Hansen's (2010) approach in modeling performance estimates as having three components:

$$\hat{\tau}_{j,t} = \varphi_j + \gamma_{j,t} + \varepsilon_{j,t} \qquad (2)$$

The estimate of teacher effectiveness from Equation 1 ($\hat{\tau}_{j,t}$), has a teacher-specific persistent component of quality (φ_j), a transitory component of teacher quality ($\gamma_{j,t}$), and a random error ($\varepsilon_{j,t}$). The transitory component is autoregressive as a random walk:

$$\gamma_{j,t+1} = \beta\gamma_{j,t} + \eta_{j,t+1} \qquad (3)$$

Here, the current transient component of teacher quality is a function of the last period's realization, and a random error ($\eta_{j,t+1}$) that is orthogonal to all other model components. For example, one might think of this transient component as professional development: it has an impact in the time period received, but over time newly learned skills fade and have a lesser impact in future years. Projecting Equation 2 forward one period, and substituting Equation 3 in for the second component yields:

$$\hat{\tau}_{j,t+1} = \varphi_j + \beta\gamma_{j,t} + \eta_{j,t+1} + \varepsilon_{j,t+1} \qquad (4)$$

This model allows for teacher quality to predict future performance, but its predictive power fades with time to the component that is persistent within teachers. This model is consistent with the observational evidence on teacher effect estimates Goldhaber and Hansen (2010) present.

In this study, we are particularly interested in whether the additional stability of VAM estimates based on multiple years of observation are more predictive of long-term outcomes, compared to those based on one year only. Estimates based on multiple years will be based on higher numbers of student observations, decreasing the relative magnitude of sampling error in the estimates. Moreover, the signal in these estimates is averaged over multiple years, providing a more precise, and potentially less biased (Koedel and Betts, 2009) estimate of teacher quality. Not all of the signals identified in any of these estimates are persistent, however, and some of the estimated effectiveness of past performance will fail to be identified in future estimates of teacher quality.

In Table 2 we present the functional form of 1-year, 2-year, and 3-year VAM estimates, along with their variances, and covariances (with one-year VAM estimates *n* years in the future). Note the relative magnitude of the persistent component in the variance of the multi-year estimates increases with additional years of observation because the variance in the sampling error and importance of the temporary component diminish when looking across multiple years.[14] Likewise, the value of the covariance terms also converges to the variance of the persistent component with time (as *n* increases). Our primary tool of empirically estimating the variation in these various components of teacher quality is a comparison of the Pearson correlation coefficients. Correlating teachers' VAM estimates over time allows us to isolate the more stable parts of teacher quality, by netting out those that change between observation periods. Given the observed correlations over multiple periods, we isolate the magnitudes of each of these variance components in the data, which in turn inform us of the predictive power of using these estimates in making tenure decisions.

[14] Empirically, this decreasing variance with multi-year VAM estimates is also observed. In our data set, the standard deviation of one-year VAM estimates in math is 0.218 and the standard deviation of three-year estimates is 0.187. At the same time, the estimated signal component of these estimates increases by adding years of observation. A similar pattern is also observed in the VAM estimates in reading.

The third line of inquiry is the extent to which past performance measures predicts student achievement. Our results in this section use a basic model of estimating student achievement, but instead of using fixed effects to control for teacher quality as in Equation 1 above, we insert a vector of teacher quality, TQ, explanatory variables:

$$A_{ijkst} = \alpha A_{is(t-1)} + TQ_{jst}\beta + X_{it}\gamma + \delta_g + \varphi_t + \varepsilon_{ijkst} \quad (5)$$

The teacher quality vector includes a teacher's licensure status, experience and degree levels, college selectivity, and average licensure scores, in addition to VAM performance estimates from a prior year of observation; X_{it} is a vector of student characteristics; δ_g is an indicator variable on grade; and φ_t is a vector of year dummies. In this section, we separately include raw one-year effectiveness estimates and analogous estimates that have been shrunk using the empirical Bayes adjustment, shrinking estimated teacher performance to the grand mean in proportion to the reliability of the teacher-specific estimate. Furthermore, we isolate the sample of teachers for whom we observe both pre- and post-tenure performance, and estimate post-tenure student learning using pre-tenure estimates of teacher performance as covariates in the regression.

Finally, we check the robustness of our results by recreating the final analysis above predicting student achievement. The most consequential critique of the VAM estimates is that they are not free of bias from non-random matching of students to their teachers. To assess whether our findings may be biased, we estimate all of the models described above using various teacher subgroups and specifications that should be less likely to suffer from this type of matching bias. Specifically, we isolate teachers in schools with new principals, where presumably any pre-existing sorting norms would be disrupted with the introduction of a new principal; we isolate 5th grade teachers in our sample and include additional lags of student test performance in estimating teacher effectiveness, which shows less bias in Rothstein

(2009b); and we isolate schools where students appear to be distributed randomly across classes, based on observable student characteristics as outlined in Clotfelter et al. (2006). The results of these tests are presented in Part D of the following section.

IV. EMPIRICAL RESULTS

Variation and Stability of Effects Over Teacher Careers

There are several reasons to believe that *true* teacher effects and the consistency of job performance might not be stable over a teacher's career. There is, for instance, good evidence that the acquisition of classroom management or other skills leads teachers to become more productive as they initially gain classroom experience (Clotfelter et al. 2006; Hanushek et al. 2005; Rockoff, 2004). Moreover, we might also expect increasing experience to coincide with a narrowing of the variation in job performance since teachers who are less productive may be counseled out of the teaching profession while the most productive teachers may be attracted to outside opportunities (Boyd et al. 2007; Goldhaber et al. 2007; Krieg, 2006; West and Chingos, 2008). This suggests a narrowing due to the sorting of individuals in the workforce, but beyond this there are reasons to believe that the consistency of job performance would increase as familiarity with job tasks instills job behaviors that permit a smoother reaction to changes in job requirements (Deadrick and Madigan, 1990).[15] Furthermore, one might imagine that teachers, as they settle into a particular setting, tend to adopt the practices of that setting (see Zevin, 1974), or adjust their effort to converge to the average effort level of their peers (Kandel and Lazear, 1992).

This would suggest a general convergence in teacher effectiveness as teachers become socialized into the norms of a school, district, or the profession.

[15] The notion of converging behavior is common (see, for instance, Dragoset (2007), for a brief review of various studies testing income convergence over time).

We investigate changes in the effectiveness of the workforce by grouping teachers according to experience level, and length of tenure in a district or in a school, and look for changes in the estimated average teacher effect and the estimated standard deviation of teacher-effect estimates conditional on experience grouping.[16]

We report the results of this exercise in Figure 1a (for teaching experience), Figure 1b (for experience in district), and 1c (for experience in school).[17] The estimates presented in these Figures are adjusted for the estimated sampling error in each experience bin, using the adjustment method used commonly in the literature (e.g. Aaronson et al. 2007; Koedel and Betts, 2007);[18] however, instead of adjusting the estimates based on Equation 2 from the entire workforce as is commonly done, we apply the adjustment to each experience bin separately.

The effect of changes in teacher quality varies somewhat by teacher experience, but is generally in the realm of 10 percent of a standard deviation for reading and just over 20 percent of a standard deviation for math. These magnitudes are roughly equivalent to estimates by Kane and Staiger (2008) – who estimate comparable models that include student covariates – for math achievement and somewhat lower than their finding for achievement in reading (18 percent of a standard deviation). And, consistent with the literature, the average teacher effect increases by statistically significant levels early on in a

[16]The experience groupings are: 0-1 yrs, 2-3 yrs, 4-5 yrs, 6-7 yrs, 8-10 yrs, 11-13 yrs, 14-16 yrs, 17-19 yrs, 20-22 yrs, 23-25. Tenure in district or tenure in school is covered by the first five bins (our panel of data does not allow us to see when teachers with a tenure of more than 10 years arrived at the school or district).

[17] We also test for convergence based on teachers possessing an MA degree or higher, teachers in the top and bottom quartile of licensure test scores, and teachers in the top and bottom quartile based on a school's proportion of students receiving free/reduced price lunch. We find no evidence of convergence along any of these dimensions.

[18] This adjustment approach assumes teacher quality is measured with error: $\hat{\tau} = \tau + \nu$. The variance of true teacher quality is recovered by subtracting the estimated sampling variance from the observed variance of the estimated teacher quality parameters: $Var(\hat{\tau}) - Var(\nu) = Var(\tau)$. The sampling variance is estimated by taking the mean standard error for each of the estimated teacher fixed effects. We use heteroskedasticity-robust estimates of the standard errors for this adjustment.

teacher's career (Clotfelter et al. 2006; Hanushek et al. 2005; Rockoff, 2004), and this is true for all types (overall, within, district, and within school) of experience.[19]

More interesting is the finding that there is little obvious narrowing of the distribution of teacher effects. The distribution of teacher effects appears to be very stable in the case of overall experience; in the case of *district* experience, it is stable for reading effects but widens considerably for the case of math experience; and in the case of *school* experience, the variability of teacher effects remains constant across same-school experience.[20]

As we suggested above, one of the arguments for why we see little evidence of a narrowing of job performance with experience is that teacher effects are estimated relative to other teachers in the workforce and the comparison group of teachers changes over time. To account for this, we identify a cohort of 556 teachers who are observed during five consecutive years and plot the standard deviations of teacher effects conditional on experience. We do not report these results, but they too show little evidence that there is behavioral convergence leading to a narrowing of the distribution of teacher effectiveness over teacher careers. Thus, unlike findings from other contexts (Kandel and Lazear, 1992), our findings for teachers do not support the notion of behavioral convergence. One explanation might be that teachers, who are under little direct performance pressure in that they typically have a high degree of

[19] We formally test this by regressing our estimates of teacher-year effects against time-varying observable teacher characteristics. Only two teacher variables were found to be statistically significant predictors of within-teacher variation in effectiveness: a teacher's experience level and a teacher's number of discretionary absences. We also test the within versus between school variation in teacher effects and find that the within-school teacher effect variance is 83% of the total variance of teacher effects in reading and 79% in math.

[20] We further assessed the extent to which the stability of teacher-performance estimates may vary over the course of a teacher's career by separately predicting future performance on lagged effectiveness estimates for teachers in each experience bin. The prediction coefficients were then compared across experience levels. As above, this test showed no evidence of time dependence in the coefficients – thus providing no counter-evidence to the hypothesis on stable job performance over the course of a teacher's career.

job security, do not feel as compelled to make sure their performance is comparable to their peers as do employees in other sectors of the economy (Jacob, 2010).[21]

Multi-Year Estimates and the Intertemporal Stability of Teacher Effects

In the literature on using VAMs to assess teacher effectiveness, the primary reason for using multiple years of observations (rather than those using a single year only) to estimate teacher performance is to improve statistical power in estimating teacher effectiveness. A natural consequence of spanning multiple years of teacher observations is the increase in sample size used to estimate a teacher's value-added effect; thus, multi-year estimates will naturally lower the standard error associated with each teacher's performance. This result was 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. Another potential benefit associated with using multi-year VAM effect estimates is that these estimates are less likely to be biased due to student sorting across teachers (Rothstein forthcoming). This finding is illustrated in a recent paper from Koedel and Betts (forthcoming), who show that while single year VAM estimates of teacher effectiveness fail the so called "Rothstein" test, multi-year estimates do not; this suggests that bias from student sorting is at least partly transitory conditions. Combining multiple years, however, necessarily aggregates two periods in which performance is not necessarily constant. McCaffrey et al. (2009) briefly discuss the consequences of this aggregation, and note that the increase in statistical power is mirrored with an increasing bias from those components of performance that do not persist within teachers.

[21] As described above, tenure attaches considerable job protections to public school teachers. Jacob (2010) analyzes the behavioral effect of probationary public school teachers losing job security through changes in collectively bargained agreements, and finds significant reductions in teacher absences, particularly among those with the highest expected levels of absence.

Piecing the results of these studies together, VAM estimates based on multiple years have some appealing features (statistical power, less sorting bias), but are not flawless estimates of performance either (bias from components of performance that are not permanent). We wish to investigate how using one-year estimates versus multi-year estimates differ in the context of rewarding tenure to teachers. To our knowledge only McCaffrey et al. (2009) have discussed the use of multi-year VAMs to impose a hypothetical teacher quality minimum prior to granting teachers tenure. As discussed previously, they suggest removing the bottom two quintiles of teachers based on true teacher performance would result in an increase in the workforce of 4 percent of a standard deviation of student learning. Recall that we employ a model of teacher quality that differs from McCaffrey et al.'s in the specification of the transient component of teacher quality. As an additional point of departure, the correlations of effectiveness within teachers over time presented in McCaffrey et al. are notably lower than what we observe with the North Carolina data.[22] We wish to compare the predicted effect of imposing a tenure rule on the market using our estimates against what McCaffrey et al. suggest.

Figures 2a and 2b show the correlations of VAM estimates based on one-year, two-years, and three-years of observation over increasing intervals of time. Applying a quality filter at the time of tenure would perform an analogous function to this: it uses observed past performance to predict performance far into the future. In both reading and math, an obvious downward decay is evident in all of the effects (whether using one year or multiple years). In spite of the decay, however, the predictive power of multi-year effects is still large: even five years after the original VAM estimation period the three-year effects still have observed correlations above those of the one-year estimates after just one year.

[22] Using Florida data, the authors report correlations on adjacent-year VAM estimates for elementary teachers in math (using a roughly analogous estimation model to ours) range from 0.30 to 0.46. The data we employ from North Carolina shows a statewide average correlation of 0.53 for elementary teachers in math.

Note that the correlations, even up to nine years later, still do not fall to zero, but appear to level out. This observed pattern is consistent with the model employed, where eventually the permanent component of teacher quality would be the only common component in performance over time. Using these observed correlations multiple years out, we generate two estimates of the variance of this persistent component of teacher quality.[23] The first estimate is based on the point where the correlations hit their lowest observed value (at year 7), and take this as convergence to the permanent component only. The second is based on when the subsequent correlation coefficient is no longer significantly different from the year prior (at year 4), and take this to represent the point of convergence. The first measure is a conservative measure, likely underestimating the true variation of this persistent component; the second measure likely overstates its variance.

Once the persistent components of teacher quality are estimated, we can estimate values for β and $Var(\gamma_j)$ using the estimated sampling error (see footnote 22) to compute the reliability of VAM estimates, given the number of years of observing performance. The results of these calculations are presented in Table 3 (Panel A is based on the conservative estimate of the permanent component of teacher quality, Panel B is based on the more liberal estimate). The first column reports the reliability of the estimate in correctly identifying teacher quality over the time period on which the estimate is based (this is the total teacher quality signal over the total variance). The second column reports the reliability of identifying the persistent component only. As expected, given the graphics presented in Figure 2, more years of observation increase both measures of reliability. Moreover, in spite of using two different

[23] In Table II, we show the covariance between one-year estimates is $Var(\varphi_j) + \beta^n Var(\gamma_j)$, where *n* is the number of years between measurement. As *n* gets large, the second part of this covariance goes to zero. When the second term is small enough to be ignorable, the observed correlation coefficient between VAM estimates multiplied by the standard deviations of the estimated effects in both periods produces estimates of $Var(\varphi_j)$.

estimates of the persistent component of teacher quality (one likely under-estimating the magnitude, the other likely over-estimating), both reliability measures are approximately the same.

In the final five columns of Table 3 we also present the predicted increase in the average level of teacher quality that could be expected by imposing a quality check when rewarding tenure. The rule we impose here is removing the lowest 25 percent of teachers based on observed (noisy) performance.[24] We report the average teacher quality (in student-learning standard deviation units) that we expect to observe in a cohort of teachers at different time intervals (1, 2, 3, 5, and 10 years) following the application of such a tenure rule. Consistent with the observed pattern in the correlations, these calculations show large immediate impacts of the rule that fade somewhat with time (recall the correlations of VAM estimates within teachers are highest when observed with less time between measurement). These calculations predict most of the fade occurs within the first 3 years, and virtually all of it is observed in the first 5 years.

In spite of this fade, the long-run effect of imposing such a rule appears that it can be consequential. While the VAM estimates based on one year of performance predict 10-year effect sizes of 0.012 and 0.015 standard deviations in reading and 0.034 to 0.037 standard deviations in math, using three years of performance in VAM estimation increases the effect size in both subjects by approximately 30 percent (ranging from 0.017 to 0.021 in reading and 0.044 to 0.047 in math). Even though we use two approaches to calculate these predicted effects, the magnitudes are reasonably consistent across the panels. Comparing our predicted effects with those calculated in McCaffrey et al. (2009), we see our long-run predicted effect sizes in math (even after the initial fade in effectiveness) are slightly larger. We return to the relevance and magnitude of this finding in Section V.

[24] In their technical appendix, McCaffrey et al. (2009) derive the formula for calculating the average teacher quality in a truncated normal distribution, given the uncertainty in identification. We use that method for these calculations here.

Predictive Power of Earlier Career Performance Estimates

Next we turn our attention to the question of whether past teacher performance is a good predictor of future results. We know from the correlations of teacher effect estimates above that there is a relationship but in this section we quantify it. We begin by reporting, in Panel A of Table 4 the results of a model regressing student achievement in year t against a standard set of observed teacher and student controls and, in some specifications, estimates of each teacher's immediate past year's VAM estimate (consistent with equation 4 above).[25] Columns 1 and 4 show the results for specifications that include observable teacher variables, columns 2 and 5 include estimates of past teacher performance (in the same subject as the test), and columns 3 and 6 include both observable teacher variables as well as past performance estimates.

Focusing first on columns 1 and 4, we see that, consistent with the literature, explanatory explain more of the variation in student achievement in math than in reading, and F-tests show that the observed teacher variables in both subjects are jointly significant. However, of the individual teacher variables, only teacher experience and performance on licensure tests are statistically significant with the expected signs.

The results in columns 2 and 5 show the results utilizing teachers' prior VAM estimates. We report the results from models that utilize the EB teacher effectiveness estimates, but as it turns out the findings differ little if the unadjusted effects are used instead.[26] Since most elementary teachers are responsible for instructing students in both reading and math, we can estimate a separate lagged VAM effect for each subject. Both student achievement and the teacher effect estimates included in the regressions are standardized by grade and year to zero mean and unit variance so the point estimates show the estimated effect size of a one standard deviation of prior teacher effectiveness on student achievement.

[25] When teacher characteristics are not included in the model, r-squared for math is .71 and for reading is .69.
[26] This is not surprising given that the correlation between the EB and unadjusted teacher effect estimates is 0.97 or higher for all year-grade combinations.

Were it the case that teacher effectiveness did not vary over time and was measured without error, we would anticipate same-subject (e.g. teacher math VAM estimate in student math achievement model) coefficient estimates in the range of 0.1 in the reading model and 0.2 in the math model as these are roughly the estimates for the effect sizes for teacher effectiveness reported in subsection A.

There is good evidence that these prior VAM estimates do predict student performance.[27] And, interestingly, the lagged VAM effects in both math and reading show up as being statistically significant in models predicting student achievement in both subjects. In other words, not only do we see that teachers who demonstrate success in instructing students in a subject tend to be successful a year later in instructing students in that *same* subject, but teachers who are more successful in instructing students in math tend to be more successful in the subsequent year in instructing students in reading and vice versa. The point estimates suggest that, on student reading achievement (column 2), a 1 standard deviation increase in a teacher's lagged effectiveness in reading increases students' reading scores by about just about 4 percent of a standard deviation, and a 1 standard deviation increase in a teacher's lagged effectiveness in math increases reading scores by just over 3 percent of a standard deviation.[28] In the math achievement models, our estimates suggest that a 1 standard deviation increase in a teacher's lagged effectiveness in reading increases students' math scores by about 1 percent of a standard deviation, and a 1 standard deviation increase in a teacher's lagged effectiveness in math increases math scores by about 12 percent of a standard deviation.[29]

[27] While adding more teacher effect estimate lags (e.g. the VAM from time period t-2) does increase the explanatory power of the model, most of the explanatory power possible was achieved from observing the most-recent year's prior VAM estimates. That said, the pattern of effects is far more consistent for math, where each year's performance estimate that is further back has a coefficient estimate that is smaller; all of the prior reading-performance estimates are positive, but they are not all statistically significant and the coefficients show no clear pattern in terms of magnitude.

[28] Recall that these are the Empirical Bayes effect estimates so have been shrunken in proportion to the reliability of the teacher-specific estimate.

[29] Dropping the cross-subject VAM from the model has only a small impact on the own subject VAM coefficient estimates (increasing the magnitude slightly).

Finally, in columns 3 and 6, we report on specifications that include both observed teacher variables and prior VAM estimates. In these models the teacher quality variables are no longer jointly significant and the estimates of the predictive power of lagged teacher effects are little changed. It is worth noting that the estimated VAM teacher effect coefficients from these models should be treated as a lower bound on the impact of *true* teacher quality, since our regressors are *estimated* performance and thus subject to measurement error.[30]

The above results confirm that estimated *prior-year,* estimated teacher performance is a good predictor of estimated future performance – an important finding in the context of thinking about using them for policy purposes. However, using VAM estimates to help inform tenure decisions – an option that is often floated in policy discussions given the perceived (or actual) difficulty of removing ineffective teachers once they are afforded the job protections that come from being tenured – likely would require a higher standard since there would be a lag between the time that VAMs could be estimated and tenure decisions were made.

In North Carolina, teachers are typically awarded tenure after four successive years of teaching in the same school district (the specific time required varies depending on whether a teacher has been tenured in another school district and/or the specific license a teacher holds for each year of teaching).[31] The data do not capture a variable on tenure status, so we use the rules governing tenure to classify teachers as tenured or not and estimate models similar to those discussed above that only include

[30] We also estimated models that use percentile ranking instead of the EB teacher effect. The results from those models differ somewhat in magnitude but are otherwise qualitatively similar.
[31] The requirements for achieving tenure in North Carolina are described in Section 1803 of Joyce (2000).

observations for students with teachers who we calculate have received tenure.[32] The only distinction in specification between these models (reported in Panel B) and those in Panel A is that the teacher VAM estimates included in these tenure models are based on teachers' first two years in the classroom (and we drop the early career experience dummies). In theory school districts in North Carolina could conceivably obtain three or four years of pre-tenure teacher VAM estimates prior to making tenure decisions, but in practice this is unlikely given the lag time for obtaining testing results and for estimating the VAM effects. Moreover, North Carolina is one of the few states that requires teachers be in the classroom for more than three years before they are eligible for tenure.[33]

Columns 1 and 4 show the estimates for observable teacher variables. In these models no teacher variables are statistically significant and F-tests indicate that they are not jointly significant. The coefficient estimates on the pre-tenure teacher VAM estimates (columns 2 and 5) for own-subject (e.g. student achievement in reading and teacher VAM reading effects) are of a very similar magnitude to those we observe when using the prior-year lagged VAM estimates (in Panel A), but the cross-subject prior VAM estimates are only about half as large.[34] The consistency of the same-subject VAM coefficients is somewhat surprising since there is a *three-year lag* between the student achievement we are

[32] It is worth noting that the sample of teachers for this part of the analysis represents a very select group of teachers, implying one should be cautious about drawing strong inferences about the teacher workforce in general. While there are nearly 20,000 unique 4th and 5th grade teachers for whom we can estimate teacher effectiveness, we observe only 1,363 unique novice teachers prior to 2003 (the last year for entering teachers for whom we could also observe post-tenure performance). Of these, only a small percentage stay in the sample (a teacher may stay in the workforce, but would only remain in our sample if they were teaching in either the 4th or 5th grade levels in experience year 5) long enough to observe post-tenure performance: 609 for whom we observe both post-tenure performance and performance estimates for their first two years of teaching.

[33] The mode of states grant tenure in the third year of teaching and several grant it after only one or two years in the classroom. For more information, see the Teacher Rules, Roles, and Rights database, managed by the National Council on Teacher Quality, available at http://www.nctq.org/tr3/.

[34] The VAM teacher effect coefficients are slightly larger in models that only include the own-subject VAM estimates. If we restrict the sample to just teachers in their 5th year, the pattern of results is similar to those reported in Panel B. Similarly, the results differ very little when we use three years of teacher classroom performance to estimate effects rather than two (all of these results are available upon request). While both cross-subject estimates are significant in Panel A, in Panel B teacher prior-effectiveness in reading is not significant in student math models and teacher prior-effectiveness in math is only marginally significant in student reading models.

estimating and the estimates of teacher effectiveness and, there is significant attrition out of the sample, likely implying a restricted range of teacher quality and a downward bias in these coefficients (Killingsworth, 1983). However it is possible that these two year estimates help mitigate for the potential of student-teacher sorting bias as is found by Koedel and Betts (2009). Lastly, when we include both observable teacher variables and the VAM pre-tenure effects (columns 3 and 6), the VAM coefficients remain nearly identical.

Tests of Robustness

In this section we describe the analyses we preformed to assess the robustness of our results. Specifically, we attempt to account for the possibility that our estimates of teacher effectiveness might be biased due to the student-to-teacher assignment process that might lead to a violation of the assumption that the student assignment to teachers is random conditional on the other variables included in the VAM model (Rothstein, 2009a).

We test whether our results are robust by estimating them for three subsamples of our data. The first subsample is 5th grade teachers for whom we have a vector of prior student achievement scores in both math and reading tests at the end of 3rd grade and the end of 4th grade. For these teachers, we estimate teacher effects (in Equation 1) using two years of lagged student performance in both subjects, rather than using just one year of lagged performance only. Rothstein (2009b) shows that this VAM specification is likely to have less bias (than the VAM with only one lagged year of performance) since the vector of twice-lagged prior achievement scores explain a significant portion of the variation in 5th grade achievement, and Kane and Staiger (2008) also use a specification similar to this and find that it produces teacher effect estimates that are similar to those produced under experimental conditions.

The second subsample we utilize is the set of teachers in schools with a new (to the school) principal.[35] The notion here is that principals influence the student-teacher assignment process; they may, for instance, reward their "good" teachers with choice classes or, alternatively, assign them to teach the more difficult students. Incumbent principals are likely to be consistent in their assignment strategies but a new principal may break from those of their predecessors (Koedel and Betts, 2009). While the new student-teacher assignment process may not be random, it is likely to result in different estimates of teacher VAM effects if it differs from the previous assignment process.[36]

Finally, we estimate our models on a sample of students and teachers that appear to be randomly matched based on the distribution of observable student characteristics (gender, ethnicity, free and reduced price lunch and limited English proficiency status, parental education, and base year reading and math test scores) across different school-year-grade units.[37] From our original sample we omit schools from the analysis if any of the chi-square tests reject the hypothesis that students are randomly distributed across classrooms.[38]

Table 5 replicates the analyses used to generate columns 3 and 6 of Table 4.[39] These results show that the coefficients on the lagged teacher VAM effects are robust to sample and model specification; in fact, the estimates in these specifications, for reading and math student achievement models, differ from those reported in columns 3 and 6 in Panel A of Table 4 by less than 0.006.

[35] About 20 percent of teachers in our sample are working in schools in which there is a new principal.
[36] Given that this break may not occur in the first year that a principal takes the helm at a school, we also estimate the models for teachers in schools with principals in their second year. We do not report these results, but they are nearly identical to the first-year principal results presented here.
[37] As an alternative approach to control for non-random sorting within schools, we inserted a dummy variable to flag those schools that appear to be sorted randomly on observable dimensions (rather than isolating the subsample as reported in table V). We omit these results for brevity, but note that the random flag was not significant in the reading model but was significant in the math model, with a coefficient of .008.
[38] For more detail on this process, see Clotfelter et al. (2006).
[39] We do not test the teacher tenure models for robustness given the tenured sample is already quite small and these specifications further restrict sample sizes.

V. POLICY IMPLICATIONS

The results we present above in Table 4 strongly imply that VAM teacher effect estimates serve as better indicators of teacher quality than observable teacher attributes, even in the case of a three year lag between the time that the estimates are derived and student achievement is predicted. But the use of VAM estimates, for instance to inform on tenure decisions, is not costless, politically or otherwise. Thus, for policy purposes it is useful to better understand the extent to which these estimates outperform other means of judging teachers. We explore this issue by comparing out of sample predictions of student achievement based on models with observable teacher characteristics and predictions of achievement based on teacher effectiveness, to actual student achievement.

Specifically, we use the coefficient estimates from panel B of Table 4 to predict student achievement in school year 2006-07 for those students who were enrolled in classes taught by teachers in the sample used to generate the results reported in Table 4.[40] For each student we obtain two different estimates of achievement in reading and two in math. The first is based on using teacher characteristics in the model (all those characteristics that are reported in columns 1 and 4 of Table 4) and the second is based on the pre-tenure VAM measure of teacher effectiveness (in columns 2 and 5 of Table 4). If anything this exercise understates the relative value of the VAM estimates as compared to teacher characteristics since in the overwhelming majority of states and school districts, teacher employment eligibility is determined solely by certification status, whereas we are utilizing all the teacher characteristics in the model for the student achievement predictions.

Not surprisingly, t-tests of the differences in mean absolute error between the observed student achievement and the predictions from the two different models suggest the pre-tenure VAMs to have superior out-of-sample predictive power to the model based on teacher characteristics in both reading

[40] Note that, due to attrition, the number of unique teachers in the sample drops from 609 in Table III to 525 for this exercise.

and math.[41] To get a better sense of whether the differences between the VAM estimates and teacher observable estimates are meaningful, we plot the mean absolute error against actual student achievement in reading and math. Figure 3 shows the mean absolute error of predictions from both models for each percentile of reading and math achievement. There are 10,127 total predictions or about 100 per percentile.

As might be expected, the results of this exercise show that both models do a relatively poor job of predicting student achievement far from the mean (i.e. where the average mean absolute error is larger). It also shows that the VAM effects model is superior to the teacher characteristics model throughout the distribution of math achievement. This is not always true for the reading predictions where the mean absolute error is similar for the two predictions (hence there is significant overlap in the lines).

What would it mean to use VAM estimates in practice for informing teacher "de-selection" decisions at the tenure point (Gordon et al. 2006)? McCaffrey et al. (2009) examine the extreme case where tenure decisions are based solely on VAM estimates. Using their derived estimates of the intertemporal stability of teacher effectiveness and assuming the persistent components of teacher quality are normally distributed, they estimate that a "de-selection" of the most ineffective 40 percent of teachers would increase the average effectiveness of those 60 percent of teachers remaining in the workforce by just over 3 percent of a standard deviation (in student achievement terms). In Table 3, we calculated the effect sizes of a similar rule using the observed estimates from the data. We imposed a slightly lower bar in our case, though, removing only the lowest 25 percent of teachers (compared with removing 40

[41] In reading, the mean absolute error (MAE) for predictions of the teacher characteristics model and VAM effects model are 0.455 and 0.451, respectively, while those for math are 0.561 and 0.419. T-tests of mean equality are strongly rejected in both subjects.

percent above).[42] Even with the lower bar, Table 3 shows that imposing this hypothetical rule could have an educationally significant effect on the distribution of teacher quality for those teachers who remain in the profession: using three-year VAM estimates, the mean level of teacher quality among teachers retained in the market are conservatively predicted to be 0.017 standard deviations higher in reading and 0.044 in math, relative to the distribution with no filter, which, as we describe above appears consistent with current policy.

Taking this hypothetical rule one step further, we can simulate what the pre- and post-selection distributions can look like using teachers observed in our sample. Specifically, we de-select teachers based on their pre-tenure reading and math effects and report the distributions (in Figure 4 of the 5th year post-tenure effectiveness estimates in those subjects.[43] The three distributions in Panel A (reading) and Panel B (math) show the estimated post-tenure effects for de-selected teachers (the lowest 25 percent), the remaining selected teachers (the upper 75 percent), and the pooled distribution of all teachers (imposing no selection rule). In reading, the de-selected teachers are estimated to have student achievement effectiveness impacts that are 10 percent of a standard deviation of student achievement below those teachers who are not de-selected, and the difference between the selected distribution and a distribution with no de-selection is over 2 percent of a standard deviation of student achievement. In math, the de-selected teachers are estimated to have impacts that are over 11 percent of a standard deviation of student achievement lower than selected teachers, and the difference between the selected distribution and a

[42] This truncation is not as large of a reduction in the distribution as may appear at first blush since early career teacher attrition is relatively high. Thirty-eight percent of first-year teachers leave the profession within two years, of which 28% are in the bottom quartile of initial effectiveness. Thus, a 25% deselection rule would require districts to deselect about an additional 14% of teachers, after accounting for natural attrition.

[43] The post-tenure reading distribution is based on pre-tenure selection on reading effects only, and the post-tenure math on pre-tenure math only.

distribution with no de-selection is almost 3 percent of a standard deviation of student achievement.[44] When we take this thought-experiment a step further and replace de-selected teachers with teachers who have effectiveness estimates that are equal to the average effectiveness of teachers in their first and second years, the post-tenure distribution average are 0.016 in reading and 0.025 in math.

The above results show that deselection might be a cost-effective alternative to current policies designed to increase student achievement, such as reducing class-size.[45] Moreover, new evidence (Hanushek et al. 2008; Hanushek, 2009) suggests that even small changes to the quality of the teacher workforce can have profound impacts on aggregate country growth rates.[46]

VI. CONCLUDING THOUGHTS: IN THE EYE OF THE BEHOLDER

Our study has investigated the stability of VAM estimates of teacher job performance and their implications for a deselection policy to the teacher labor market. The evidence presented here shows no detectable evidence of the variation in teacher quality changing over time and it is reasonably stable within teachers over time. We also show VAM estimates based on multiple years of observation are more reliable in predicting long-term job performance, and early-career performance reliably signals post-tenure performance. These findings do not appear to be due to conflated biases in VAM estimates due to student sorting.

[44] These estimates are slightly lower than those reported by McCaffrey et al. (2009). Keep in mind, however, that we deselected a smaller percentage of the teacher workforce, so the smaller magnitude is reasonable. These estimates also vary from the predicted effects calculated in Table III (the observed difference in Figure 4 is larger in reading than what is calculated in Table III, and vice versa for math). This may arise because this figure focuses on the sample of teachers observed on both sides of the tenure point, whereas the results in Table III are based on estimates across the entire workforce. Moreover, sample size is quite small in Figure 3, so the true differences in effectiveness may vary with a larger sample.

[45] For instance, using class-size reduction estimates from Rivkin et al (2005), applying this deselection rule would result in an increase in student achievement roughly equivalent to reducing class size by 2.3 to 6.1 students, depending on whether math or reading estimates are used.

[46] Using the same procedure as Hanushek (2009), we estimate our deselection rule to have an impact on student achievement that would result in an additional .5 to 1 percent point increase in GDP per year. This estimation method assumes an open economy and a 13-year schooling career.

What does all this mean for personnel decisions and tenure policy in particular? We suspect the results presented here will tend to reinforce views on both sides of the policy divide over whether VAM estimates of teacher job performance ought to be used for high-stakes purposes like determining tenure. Those opposed to the idea might point to the finding that the multi-year correlations in teacher effects are modest by some standards, and that we cannot know the extent to which this reflects true fluctuations in performance or changes in class or school dynamics outside of a teacher's control (such as the oft-mentioned dog barking outside the window on testing day). Further, the observed fade in the predictive ability of VAMs at increasing time intervals weakens the effect of any policy intervention based on these VAMs with time.

On the flip side, supporters of VAM-based reforms might note that these inter-temporal estimates are very much in line with findings from other sectors of the economy that do use them for high-stakes personnel decisions (Goldhaber and Hansen, 2008). Perhaps more importantly, there is good evidence that school systems are not very selective in terms of which teachers receive tenure and, while VAM estimates are noisy, our calculations suggest that using them to inform de-selection policies has the potential to affect the quality of the teacher workforce in economically meaningful ways.

Keep in mind that our calculations are only based on a partial equilibrium analysis. There is a question of whether a change in tenure policy might have far reaching consequences for who enters the teacher labor force and how teachers behave. Teaching jobs appear to be relatively secure and changes to the security of the occupation might shift the number or quality of prospective teachers. All of this suggests that we cannot know the full impact of using VAM-based reforms without conducting assessments of actual policy variation, but the results presented here indicate that teacher effect estimates are far superior to observable teacher variables as predictors of student achievement, suggesting that these estimates are a reasonable metric to use as a factor in making substantive personnel decisions.

REFERENCES

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

Ballou, Dale. 2005. *Value-Added Assessment: Controlling for Context with Misspecified Models*. Paper presented at the CALDER Conference, The Urban Institute, Washington, D.C. March 2005.

Ballou, Dale, William Sanders, and Paul S. Wright. 2004. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1):37–66.

Boyd, Donald J., Hamilton Lankford, Susanna Loeb, and James H. Wyckoff. 2005. "Explaining the Short Careers of High-Achieving Teachers in Schools with Low-Performing Students." *American Economic Review* 95(2):166–71.

Boyd, Donald J., Pamela L. Grossman, Hamilton Lankford, Susanna Loeb, and James H. Wyckoff. 2005. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." Working Paper 11844. Cambridge, MA: National Bureau of Economic Research

———. 2007. *Teacher Attrition, Teacher Effectiveness and Student Achievement.* Paper presented at the Annual Conference of the American Education Finance Association, Baltimore, MD, March 2007.

Branch, Gregory F., Eric A. Hanushek, and Steven G. Rivkin. 2009. "Estimating Principal Effectiveness." Working Paper 32. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute.

Clark, Damon, Paco Martorell, and Jonah E. Rockoff. 2009. "School Principals and School Performance." Working Paper 38. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41(4):778–820.

Deadrick, Diana L., and Robert M. Madigan. 1990. "Dynamic Criteria Revised: A Longitudinal Study of Performance Stability and Predictive Validity." *Personnel Psychology* 43(1): 717–44.

Dragoset, Lisa M. 2007. "Convergent Earnings Mobility in the U.S.: Is this Finding Robust?" U.S. Census Bureau.

Goldhaber, Dan. 2006. "Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?" Working Paper 9. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute.

———. 2006. "National Board Teachers Are More Effective, But Are They In The Classrooms Where They're Needed The Most?" *Education Finance and Policy* 1(3): 372–82.

Goldhaber, Dan, and Emily Anthony. 2007. "Can Teacher Quality be Effectively Assessed? National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89(1): 134–50.

Goldhaber, Dan, Betheny Gross, and Daniel Player. 2007. "Are Public Schools Losing Their 'Best'? Assessing the Career Transitions of Teachers and Their Implications for the Quality of the Teacher Workforce." Working Paper 2007-2. Seattle, WA: Center on Reinventing Public Education.

Goldhaber, Dan and Michael Hansen. 2008. "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions." Policy Brief 3. Washington, D.C.: National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute.

———. 2010. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Working Paper. University of Washington.

Goldstein, Andrew. 2001. "Ever Try To Flunk A Bad Teacher?" *Time* (June 24, 2001).

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." White Paper 2006-01. Washington, D.C.: The Brookings Instiution, The Hamilton Project.

Hanushek, Eric A.. 2009. "Teacher Deselection." In *Creating a New Teaching Profession*, edited by Dan Goldhaber and Jane Hannaway (165-80). Washington, D.C.: The Urban Institute Press.

Hanushek, Eric A., and Steven G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* 100(2), forthcoming.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2004. "Why Public Schools Lose Teachers." *Journal of Human Resource* 39(2): 326–54.

Hanushek, Eric A., John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2005. "The Market for Teacher Quality." Working Paper 11124. Cambridge, MA: National Bureau of Economic Research.

Hanushek, Eric A., Dean T. Jamison, Eliot A. Jamison, and Ludger Woessmann. 2008. "Education and Economic Growth: It's Not Just Going to School but Learning That Matters." *Education Next* 8 (2): 62–70.

Hoffman, David A., Rick Jacobs, and Steve J. Gerras. 1992. "Mapping Individual Performance Over Time." *American Psychological Association* 77(2):185–95.

Hoffman, David A., Rick Jacobs, and Joseph E. Baratta. 1993. "Dynamic Criteria and the Measurement of Change." *Journal of Applied Psychology*, 78(2):194-204.

Jacob, Brian A. 2010. "The Effect of Employment Protection on Worker Effort: Evidence from Public Schooling." Working Paper 15655. Cambridge, MA: National Bureau of Economic Research.

Jacob, Brian A., and Lars J. Lefgren. 2005. "Principals as Agents: Subjective Performance Measurement in Education." Working Paper 11463. Cambridge, MA: National Bureau of Economic Research.

Joyce, Robert P. 2000. *The Law of Employment in North Carolina's Public Schools*. University of North Carolina Chapel Hill, Institute of Government.

Judiesch, Michael K., and Frank L. Schmidt. 2000. "Between-Worker Variability in Output under Piece-Rate Versus Hourly Pay Systems." *Journal of Business and Psychology* 14(4): 529–52.

Kandel, Eugene, and Edward Lazear. 1992. "Peer Pressure and Partnerships." *The Journal of Political Economy* 100(4): 801–17.

Kane, Thomas J., and Douglas O. Staiger. 2001. "Improving School Accountability Measures." Working Paper 8156. Cambridge, MA: National Bureau of Economic Research.

———. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *The Journal of Economic Perspectives* 16(4): 91–114.

———. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2006. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." Working Paper 12155. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas J., Douglas O. Staiger, and Jeffrey Geppert. 2002. "Randomly Accountable." *Education Next* 2(1): 57–61.

Killingsworth, Mark. 1983. *Labor Supply*. Cambridge, UK: Cambridge University Press.

Koedel, Cory, and Julian R. Betts. 2007. "Re-Examining the Role of Teacher Quality In the Educational Production Function." Working Paper 0708. University of Missouri, Department of Economics.

———. 2008. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." Working Paper 2008-21. Nashville, TN: National Center on Performance Incentives.

———. 2009. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." Working Paper 09-02. University of Missouri, Department of Economics.

Krieg, John M. 2006. "Teacher Quality and Attrition." *Economics of Education Review 25* (1):13–27.

Lockwood, J. R., Daniel F. McCaffrey, Laura S. Hamilton, Brian Stecher, Vi-Nhuan Le, and Jose F. Martinez. 2007. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Education Measurement* 44(1): 47–67.

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

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton. 2004. *Evaluating Value-Added Models for Teacher Accountability*. Santa Monica, CA: RAND Corporation.

McGuinn, Patrick. 2010. "Ringing the Bell for K-12 Teacher Tenure Reform." Washington, D.C.: Center for American Progress.

National Council on Teacher Quality. 2008. "State Teacher Policy Yearbook: Progress on Teacher Quality."

Rivkin, Steven G. 2007. "Value-Added Analysis and Education Policy." Policy Brief 1. Washington, D.C.: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Students' Achievement: Evidence from Panel Data." *American Economic Review* 94(2): 247–52.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. "Can You Recognize an Effective Teacher When You Recruit One?" Working Paper No. 14485. Cambridge, MA: National Bureau of Economic Research.

Rothe, Harold F. 1978. "Output Rates Among Industrial Employees." *Journal of Applied Psychology* 63(1): 40–6.

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

Rothstein, Jesse. 2009. "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy* 4(4): 537–71.

Sanders, William L., James J. Ashton, and Paul S. Wright. 2005. *Comparison of the Effects of NBPTS Certified Teachers with Other Teachers on the Rate of Student Academic Progress*. Final Report requested by the National Board for Professional Teaching Standards.

Sass, Tim R. 2008. "The Stability of Value-Added Measures of Teacher Quality and Implications for Teacher Compensation Policy." Policy Brief 4. Washington, D.C.: The Urban Institute, National Center for Analysis of Longitudinal Data in Education Research.

The Tennessee Journal. 2009. "Bresden Will Give Legislators a Week to Pass Education Reform." T*he Tennessee Journal* 35(1–2).

Todd, Petra E., and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113(485): F3–F33.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness." Brooklyn, NY: *The New Teacher Project*.

West, Martin R., and Matthew M Chingos. 2009. "Teacher Effectiveness, Mobility, and Attrition in Florida." In *Performance Incentives: Their Growing Impact on American K-12 Education,* edited by Matthew G. Springer (251–71). Washington, D.C.: Brookings Institution Press.

Zevin, Jack. 1974. "In Thy Cooperating Teacher's Image: Convergence of Social Studies Student Teachers' Behavior Patterns with Cooperating Teachers' Behavior Patterns." Education Resources Information Center, ERIC# ED087781.

Tables and Figures

TABLE I
Descriptive Means and Standard Deviations

	Unrestricted	Sample
Panel A. Student Characteristics		
Female	0.488	0.499
	(0.500)	(0.500)
Black	0.298	0.284
	(0.457)	(0.451)
Hispanic	0.052	0.039
	(0.222)	(0.193)
Other Non-White	0.052	0.043
	(0.221)	(0.204)
Free Lunch Eligible	0.464	0.329
	(0.499)	(0.470)
Parents' Bachelor's Deg. Or Higher	0.152	0.155
	(0.359)	(0.362)
Standardized Reading*	0.000	0.066
	(1.000)	(0.971)
Standardized Math*	0.000	0.080
	(1.000)	(0.976)
Observations (Students)		
Grade 4	1,122,586	667,621
Grade 5	1,029,259	541,801
Total	2,151,845	1,209,422
Panel B. Teacher Characteristics		
Female		0.901
		(0.296)
Black		0.132
		(0.339)
Hispanic		0.004
		(0.065)
Other Non-White		0.010
		(0.010)
Master's Degree or Higher		0.241
		(0.428)
Approved NC Education Program		0.413
		(0.492)
Full Licensure		0.752
		(0.432)
Years Of Experience		8.798
		(9.403)
25^{th} Percentile-Reading		-0.109
75^{th} Percentile-Reading		0.114
25^{th} Percentile-Math		-0.161
75^{th} Percentile-Math		0.161
Observations (Teachers)		
Grade 4		11,854
Grade 5		7,732
Total		19,586

*Standard Deviations in Parentheses

TABLE II
Properties of VAM Estimates, Given the Number of Years of Observation Used

	1 year	2 years	3 years
$\hat{\tau}_t =$	$\varphi_j + \gamma_{j,t} + \varepsilon_{j,t}$	$(2\varphi_j + (1+\beta)\gamma_{j,t-1} + \eta_{j,t} + \varepsilon_{j,t} + \varepsilon_{j,t-1})/2$	$(3\varphi_j + (1+\beta+\beta^2)\gamma_{j,t-2} + (1+\beta)\eta_{j,t-1} + \eta_{j,t} + \varepsilon_{j,t} + \varepsilon_{j,t-1} + \varepsilon_{j,t-2})/3$
$Var(\hat{\tau}_t) =$	$Var(\varphi_j) + Var(\gamma_j) + Var(\varepsilon_j)$	$Var(\varphi_j) + \frac{2+2\beta}{4} Var(\gamma_j) + \frac{1}{2} Var(\varepsilon_j)$	$Var(\varphi_j) + \frac{2+4\beta+3\beta^2}{9} Var(\gamma_j) + \frac{1}{3} Var(\varepsilon_j)$
$Cov(\hat{\tau}_t, \hat{\tau}_{t+n}) =$	$Var(\varphi_j) + \beta^n Var(\gamma_j)$	$Var(\varphi_j) + \frac{\beta^n + \beta^{n+1}}{2} Var(\gamma_j)$	$Var(\varphi_j) + \frac{\beta^n + \beta^{n+1} + \beta^{n+2}}{3} Var(\gamma_j)$

Note: Presented calculations are based on the result that $Var(\eta_j) = (1-\beta^2)Var(\gamma_j)$ across all periods in a stationary time series.

TABLE III
VAM Reliability and Effect on Average Teacher Quality

Table III VAM Reliability and Effect on Average Teacher Quality								
Panel A. Conservative Estimate of Persistent Component								
		Total TQ Reliability	Persistent TQ Reliability	Increase in Average Teacher Quality Over Time				
				1 year	2 years	3 years	5 years	10 years
1-year VAMs	Reading	0.597	0.175	0.016	0.013	0.013	0.012	0.012
	Math	0.784	0.331	0.044	0.039	0.037	0.035	0.035
2-year VAMs	Reading	0.691	0.268	0.018	0.016	0.015	0.015	0.015
	Math	0.858	0.435	0.047	0.043	0.041	0.040	0.040
3-year VAMs	Reading	0.717	0.369	0.020	0.019	0.018	0.018	0.018
	Math	0.883	0.538	0.051	0.047	0.046	0.045	0.044
Panel B. Liberal Estimate of Persistent Component								
		Total TQ Reliability	Persistent TQ Reliability	Increase in Average Teacher Quality Over Time				
				1 year	2 years	3 years	5 years	10 years
1-year VAMs	Reading	0.597	0.215	0.018	0.016	0.015	0.015	0.015
	Math	0.784	0.354	0.045	0.040	0.038	0.037	0.037
2-year VAMs	Reading	0.691	0.330	0.021	0.019	0.019	0.018	0.018
	Math	0.858	0.464	0.049	0.045	0.044	0.043	0.043
3-year VAMs	Reading	0.721	0.447	0.023	0.022	0.022	0.021	0.021
	Math	0.884	0.571	0.053	0.049	0.048	0.047	0.047

Note: All calculated values presented are based on observed variance in VAM estimates (0.027 in reading and 0.062 in math) and estimated variance of the sampling error (0.011 and 0.013). Panel A uses a conservative estimate of the persistent component (0.005 and 0.021) to impute a value of beta (0.237 and 0.395). Panel B uses a liberal estimate of the persistent component (0.006 and 0.022) to impute a value of beta (0.310 and 0.424).

TABLE IV
Reading and Math Student Achievement Models

	Student Reading Achievement			Student Math Achievement		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Models with 1-Year Lagged VAM Effects (Number of Teachers=9678, Number of Students=649,650)						
Observable Teacher Characteristics						
2-3 Years Experience	-0.016		-0.015	-0.002		0.005
	(0.016)		(0.014)	(0.019)		(0.017)
4-5 Years Experience	0.003		-0.012	0.023		0.002
	(0.016)		(0.014)	(0.019)		(0.017)
6-9 yrs experience	0.010		-0.008	0.027		0.002
	(0.016)		(0.014)	(0.019)		(0.017)
>9 yrs experience	0.027*		0.004	0.035		0.005
	(0.016)		(0.014)	(0.019)		(0.017)
Holds master's degree	-0.005		-0.005	0.006		0.003
	(0.003)		(0.002)	(0.005)		(0.003)
Average Licensure Test Score	0.010**		0.005**	0.019**		0.009**
	(0.002)		(0.001)	(0.003)		(0.002)
College Selectivity	-0.001		0.000	-0.005*		-0.003*
	(0.001)		(0.001)	(0.002)		(0.001)
Fully Licensed	0.014		0.007	0.023*		0.008
	(0.008)		(0.007)	(0.012)		(0.008)
VAM Teacher Effects						
1-Yr Lagged Reading Effect		0.041**	0.040**		0.003*	0.003*
		(0.001)	(0.001)		(0.002)	(0.002)
1-Yr Lagged Math Effect		0.030**	0.030**		0.130**	0.130**
		(0.001)	(0.001)		(0.002)	(0.002)
R squared	0.69	0.69	0.69	0.72	0.73	0.73
Panel B. Tenured Teacher Models with 2-Year VAM Effects (Number of Teachers=609, Number of Students=26,280)						
Observable Teacher Characteristics						
6-9 Years Experience	0.012		0.010	0.016		0.005
	(0.010)		(0.010)	(0.013)		(0.012)
>9 Years Experience	0.031		0.026	0.083		0.037
	(0.051)		(0.050)	(0.048)		(0.050)
Holds Master's Degree	-0.010		-0.013	0.004		-0.002
	(0.014)		(0.012)	(0.021)		(0.019)
Average Licensure Test Score	0.002		0.002	0.015		0.016
	(0.008)		(0.007)	(0.012)		(0.010)
College Selectivity	0.002		0.000	-0.005		-0.004
	(0.006)		(0.006)	(0.008)		(0.007)
Fully Licensed	0.011		0.011	0.078		0.089
	(0.048)		(0.048)	(0.051)		(0.052)
VAM Teacher Effects						
2-Yr Lagged Pre-Tenure Reading Effect		0.038**	0.038**		0.000	0.000
		(0.008)	(0.008)		(0.009)	(0.009)
2-Yr Lagged Pre- Tenure Math		0.012	0.013		0.092**	0.092**
		(0.008)	(0.008)		(0.009)	(0.009)
R squared	0.67	0.67	0.67	0.72	0.73	0.73

**, *: Significant at 1% and 5% confidence level, respectively. Note: All models include the following controls: a student's pre-test score, race/ethnicity, gender, free- or reduced-price lunch status, and parental education. For models in Panel B, the omitted teacher experience category is 1 year, as classified for pay purposes by North Carolina. For models in Panel B, the omitted teacher experience category is 5.

TABLE V
Robustness Checks

	Student Reading Achievement				Student Math Achievement			
	VAM Based on Vector of Prior Achievement	New Principal	Observed Random Student-Teacher Match	School Fixed Effects	VAM Based on Vector of Prior Achievement	New Principal	Observed Random Student-Teacher Match	School Fixed Effects
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher Observables								
2-3 yrs experience	-0.056**	-0.010	0.011	-0.020	-0.012	-0.001	0.021	0.005
	(0.021)	(0.029)	(0.022)	(0.012)	(0.032)	(0.051)	(0.030)	(0.011)
4-5 years experience	-0.047**	0.010	0.015	-0.016	-0.015	-0.013	0.021	0.004
	(0.021)	(0.028)	(0.022)	(0.012)	(0.032)	(0.051)	(0.030)	(0.011)
6-9 yrs experience	-0.040	0.007	0.018	-0.012	-0.011	-0.012	-0.017	0.009
	(0.021)	(0.028)	(0.022)	(0.012)	(0.031)	(0.051)	(0.030)	(0.011)
>9 yrs experience	-0.027	0.026	0.032	-0.001	-0.006	0.001	0.023	0.013
	(0.021)	(0.028)	(0.022)	(0.012)	(0.031)	(0.050)	(0.030)	(0.011)
Holds master's degree	-0.006	-0.004	-0.008**	-0.003	-0.002	0.000	-0.001	0.001
	(0.003)	(0.005)	(0.003)	(0.002)	(0.004)	(0.006)	(0.004)	(0.002)
Average Licensure Test Score	0.005**	0.006*	0.004*	0.001	0.009**	0.014**	0.009**	0.009**
	(0.002)	(0.003)	(0.002)	(0.001)	(0.002)	(0.003)	(0.002)	(0.001)
College Selectivity	-0.001	-0.002	-0.001	0.002	-0.002	-0.001	-0.002	-0.002*
	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.003)	(0.002)	(0.001)
Licensure	0.002	0.021	0.009	0.005	0.015	0.033	0.013	0.018**
	(0.011)	(0.018)	(0.010)	(0.005)	(0.012)	(0.020)	(0.012)	(0.005)
VAM Teacher Effects								
Lagged Teacher Reading Effect	0.036**	0.035	0.043**	0.026**	0.002	0.008*	0.003	0.005**
	(0.002)	(0.003)	(0.002)	(0.001)	(0.002)	(0.004)	(0.002)	(0.001)
Lagged Teacher Math Effect	0.029**	0.035	0.027**	0.032**	0.116	0.122**	0.124**	0.111**
	(0.002)	(0.003)	(0.002)	0.001	(0.002)	(0.003)	(0.002)	(0.001)
R squared	0.54	0.67	0.69	0.69	0.7	0.72	0.73	0.73
No Teachers	4845	4,664	6,766	9,678	4,845	4,664	6,766	9,678
No Students	306,942	120,001	287,561	649,650	306,942	120,001	287,561	649,650

FIGURE I(A)
Overall Teacher Performance

FIGURE I(B)
Teacher Experience in District

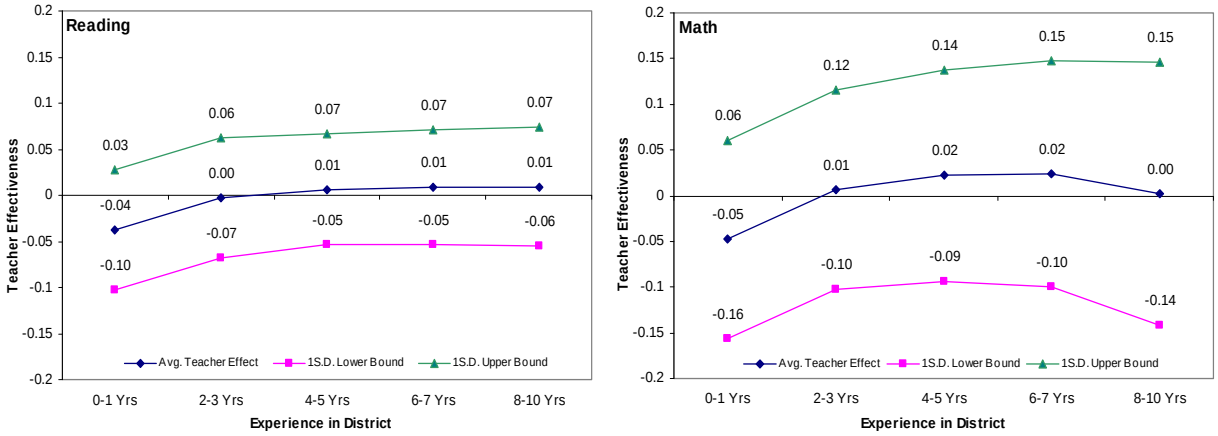


FIGURE I(C)
Teacher Experience in School

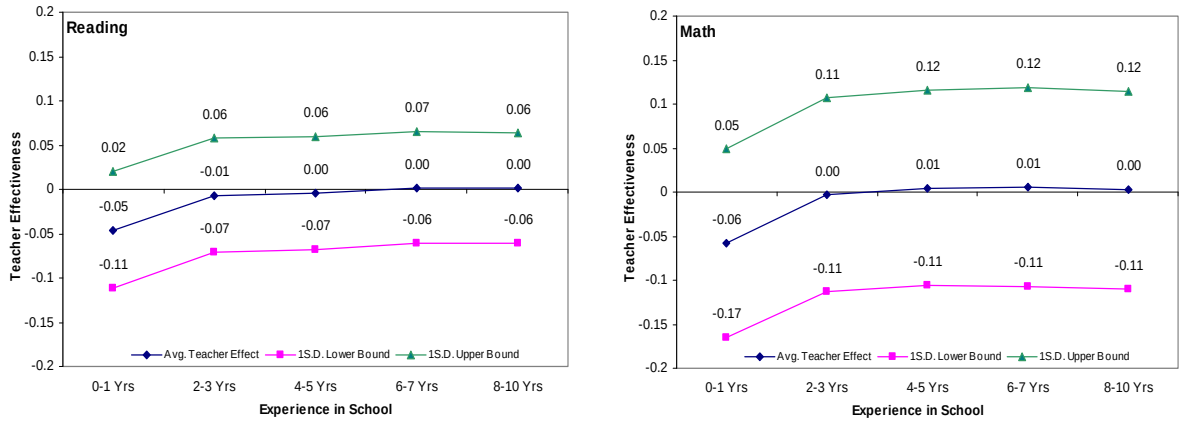


FIGURE II(A)
Correlation of Reading Effects at Increasing Intervals

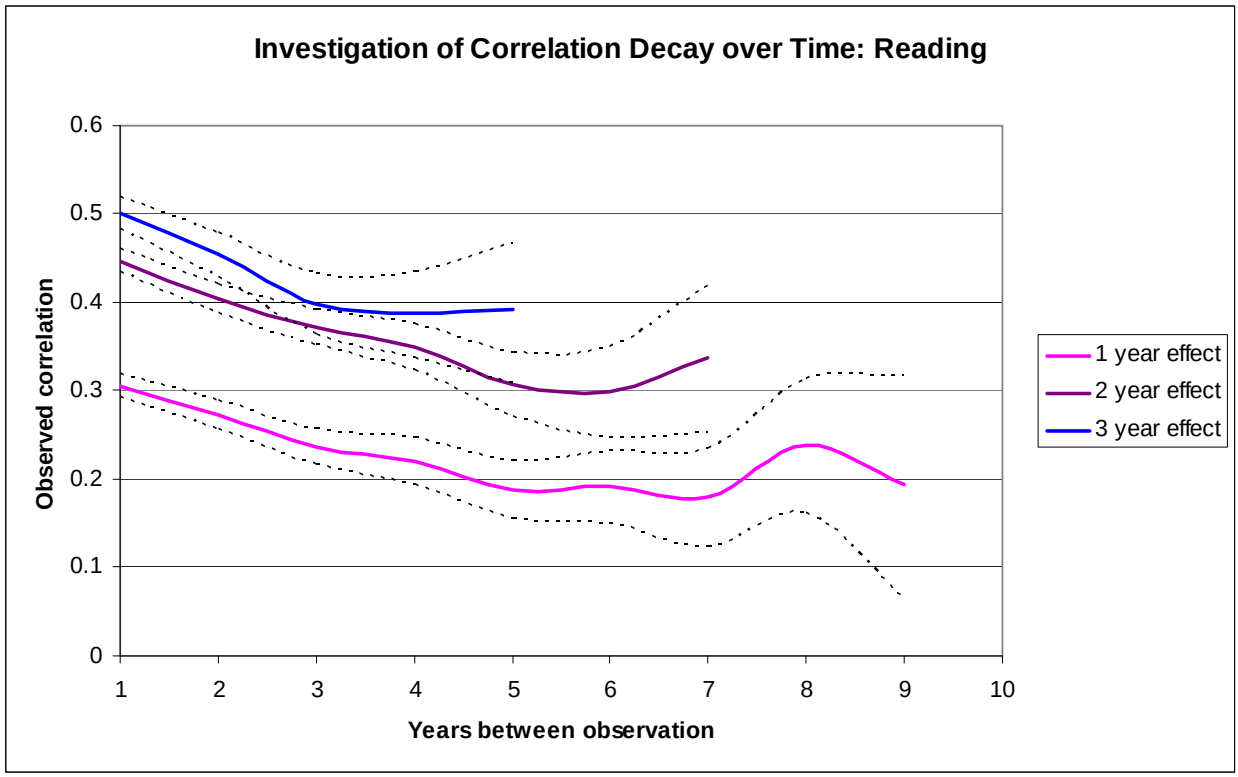


FIGURE II(B)
Correlation of Math Effects at Increasing Interval

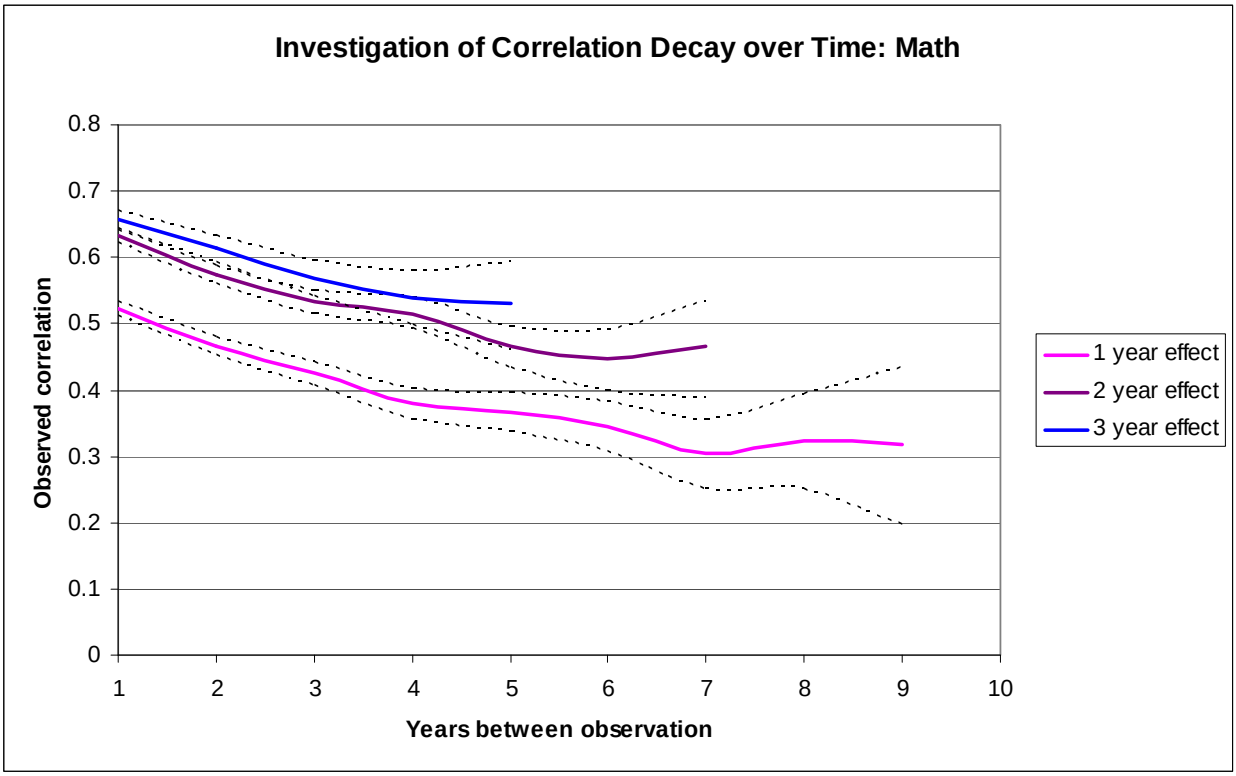
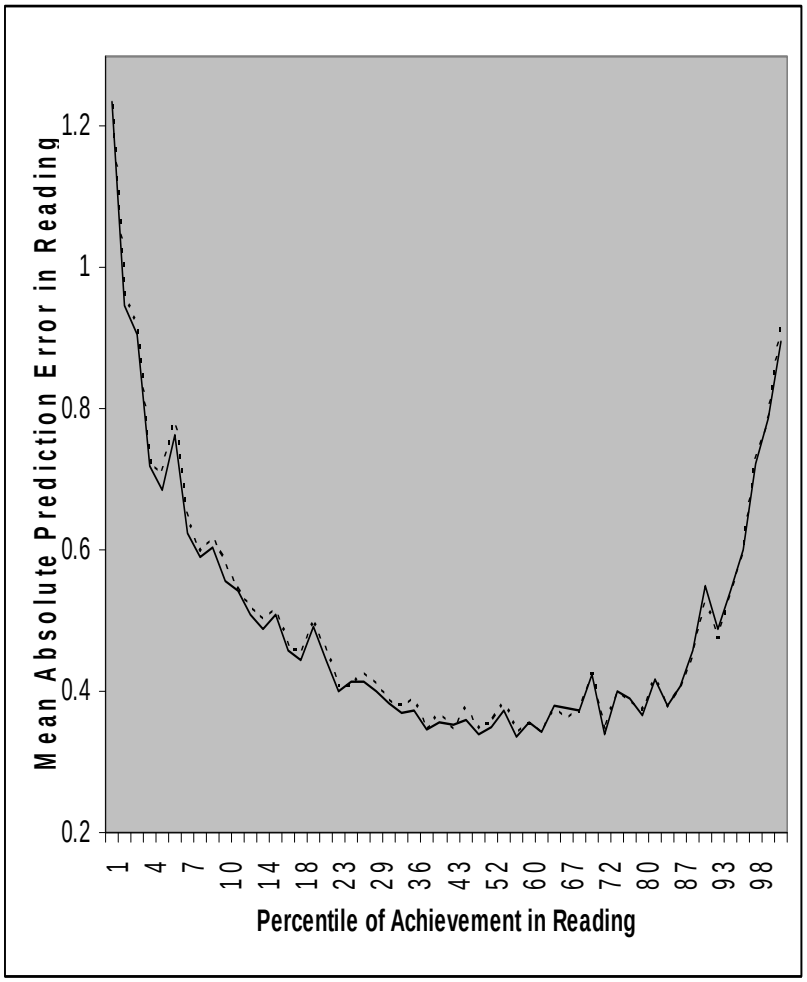


FIGURE III
Prediction Error as a Function of Achievement

Panel A. Reading



Panel B. Math

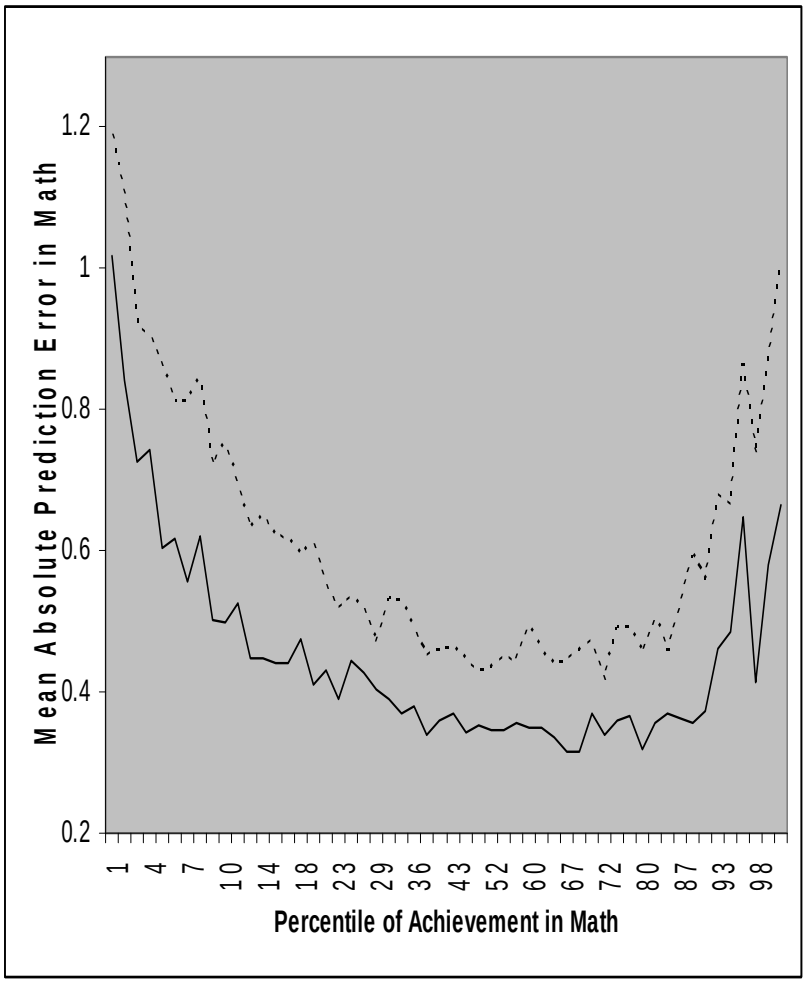
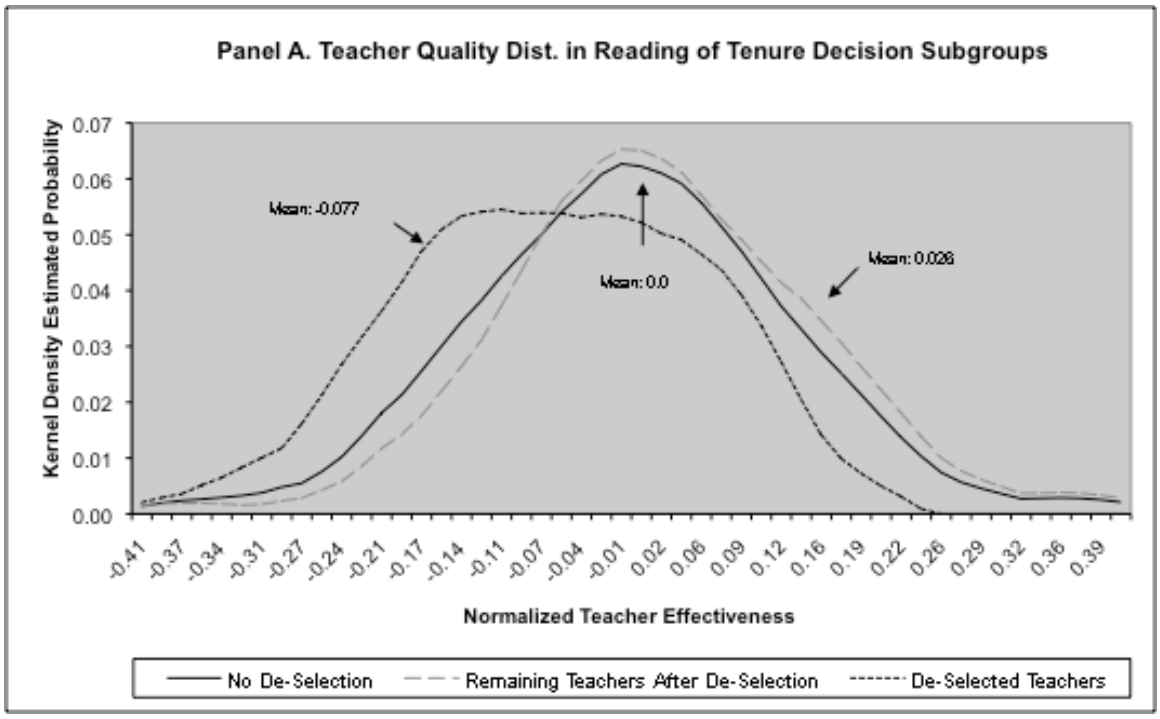
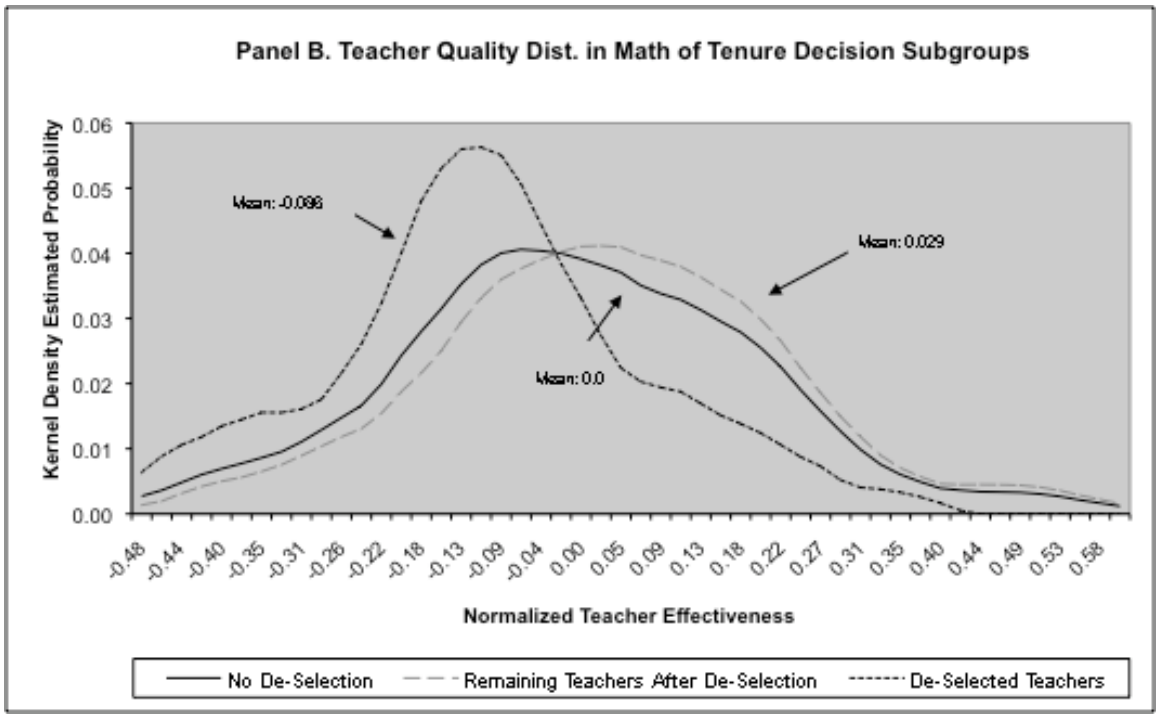

Mean Absolute Prediction Error in Math
1.2
1
0.8
0.6
0.4
0.2
1
4
7
10
14
18
23
29
36
43
52
60
67
72
80
87
93
98
Percentile of Achievement in Math

FIGURE IV
The Effects of De-Selection on Teacher Quality






CALDER