Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement

Jesse Rothstein*

Princeton University and NBER

This version: March 2009

First version: September 2007

Abstract

Growing concerns over the inadequate achievement of U.S. students have led to proposals to reward good teachers and penalize (or fire) bad ones. The leading method for assessing teacher quality is "value added" modeling (VAM), which decomposes students' test scores into components attributed to student heterogeneity and to teacher quality. Implicit in the VAM approach are strong assumptions about the nature of the educational production function and the assignment of students to classrooms. In this paper, I develop falsification tests for three widely used VAM specifications, based on the idea that future teachers cannot influence students' past achievement. In data from North Carolina, each of the VAMs' exclusion restrictions are dramatically violated. In particular, these models indicate large "effects" of 5th grade teachers on 4th grade test score gains. I also find that conventional measures of individual teachers' value added fade out very quickly and are at best weakly related to long-run effects. I discuss implications for the use of VAMs as personnel tools.

*Industrial Relations Section, Firestone Library, Princeton, NJ 08544. E-mail: jrothst@princeton.edu. Earlier versions of this paper circulated under the title "Do Value Added Models Add Value?" I am grateful to Nathan Wozny and Enkeleda Gjeci for exceptional research assistance. I thank Orley Ashenfelter, Henry Braun, David Card, Henry Farber, Bo Honoré, Brian Jacob, Tom Kane, Larry Katz, Alan Krueger, Sunny Ladd, David Lee, Lars Lefgren, Austin Nichols, Amine Ouazad, Mike Rothschild, Cecilia Rouse, Diane Schanzenbach, Eric Verhoogen, Tristan Zajonc, anonymous referees, and conference and seminar participants for helpful conversations and suggestions. I also thank the North Carolina Education Data Research Center at Duke University for assembling, cleaning, and making available the confidential data used in this study. Financial support was generously provided by the Princeton Industrial Relations Section and Center for Economic Policy Studies and the U.S. Department of Education (under grant R305A080560).

1 Introduction

Parallel literatures in labor economics and education adopt similar econometric strategies for identifying the effects of firms on wages and of teachers on student test scores. Outcomes are modeled as the sum of the firm or teacher effect, individual heterogeneity, and transitory, orthogonal error. The resulting estimates of firm effects are used to gauge the relative importance of firm and worker heterogeneity in the determination of wages. In education, so-called "value added models" (hereafter, VAMs) have been used to measure the importance of teacher quality to educational production, to assess teacher preparation and certification programs, and as important inputs to personnel evaluations and merit pay programs.[1]

All of these applications suppose that the estimates can be interpreted causally. But observational analyses can identify causal effects only under unverifiable assumptions about the correlation between treatment assignment – the assignment of students to teachers, or the matching of workers to firms – and other determinants of test scores and wages. If these assumptions do not hold, the resulting estimates of teacher and firm effects are likely to be quite misleading.

Anecdotally, assignments of students to teachers incorporate matching to take advantage of teachers' particular specialties, intentional separation of children who are known to interact badly, efforts on the principal's part to reward favored teachers through the allocation of easy-to-teach students, and parental requests (see, e.g., Jacob and Lefgren, 2007; Monk, 1987). These are difficult to model statistically. Instead, VAMs typically assume that teacher assignments are random conditional on a single (observed or latent) factor.

In this paper, I develop and implement tests of the exclusion restrictions of commonly-used value added specifications. My strategy exploits the fact that *future* teachers cannot have causal effects on *past* outcomes, while violations of model assumptions may lead to apparent counterfactual "effects" of this form. Test scores, like wages, are serially correlated, and as a result an association between the current teacher and the lagged score is strong evidence against

[1]On firm effects, see, e.g., Abowd and Kramarz (1999). For recent examinations of teacher effects modeling, see Braun (2005a,b); Harris and Sass (2006); McCaffrey et al. (2003); and Wainer (2004).

exogeneity with respect to the current score.

I examine three commonly used VAMs, two of which have direct parallels in the firm effects literature. In the simplest, most widely used VAM – which resembles the most common specification for firm effects – the necessary exclusion restriction is that teacher assignments are orthogonal to all other determinants of the so-called "gain" score, the change in a student's test score over the course of the year. If this restriction holds, 5th grade teacher assignments should not be correlated with students' gains in 4th grade. Using a large micro-data set describing North Carolina elementary students, I find that there is in fact substantial dispersion of students' 4th grade gains across 5th grade teachers. Students are particularly strongly sorted on the basis of past reading gains, though there is clear evidence of sorting on math gains as well. Because test scores exhibit strong mean reversion – and thus gains are negatively autocorrelated – sorting on past gains produces bias in the simple VAM's estimates.

The other VAMs that I consider rely on different exclusion restrictions, namely that classroom assignments are as good as random conditional on either the lagged test score or the student's (unobserved, but permanent) ability. I discuss how past gains can be used to test these restrictions as well. I find strong evidence in the data against each.

Evidently, classroom assignments respond dynamically to annual achievement in ways that are not captured by the controls typically included in VAM specifications. To evaluate the magnitude of the biases that assignments produce, I compare common VAMs to a richer model that conditions on the complete achievement history. Estimated teacher effects from the rich model diverge importantly from those obtained from the VAMs in common use. I discuss how selection on *unobservables* is likely to produce substantial additional biases. I use a simple simulation to explore the sensitivity of teacher rankings to these biases. Under plausible assumptions, simple VAMs can be quite misleading. The rich VAM that controls for all observables does better, but still yields rankings that diverge meaningfully from the truth.

My estimates also point to an important substantive result. To the extent that any of the VAMs that I consider identify causal effects, they indicate that teachers' long-run effects are at best weakly proxied by their immediate impacts. A teacher's effect in the year of exposure – the universal focus of value added analyses – is correlated only 0.3 to 0.5 with her cumulative effect over two years, and even less with her effect over three years. Accountability policies that rely on measures of short-term value added would do an extremely poor job of rewarding the teachers who are best for students' longer run outcomes.

An important caveat to the empirical results is that they may be specific to North Carolina. Students in other states or in individual school districts might be assigned to classrooms in ways that satisfy the assumptions required for common VAMs. And the results may not generalize to models of firm effects on worker wages. But at the least, VAM-style analyses should attempt to evaluate the model assumptions, perhaps with methods like those used here. Models that rely on incorrect assumptions are likely to yield misleading estimates, and policies that use these estimates in hiring, firing, and compensation decisions may reward and punish teachers for the students they are assigned as much as for their actual effectiveness in the classroom.

Section 2 reviews the use of pre-assignment variables to test exogeneity assumptions. Section 3 introduces the three VAMs, discusses their implicit assumptions, and describes my proposed tests. Section 4 describes the data. Section 5 presents results. Section 6 attempts to quantify the biases that non-random classroom assignments produce in VAM-based analyses. Section 7 presents evidence on teachers' long-run effects. I conclude, in Section 8, by discussing some implications for the design of incentive pay systems in education.

2 Using Panel Data To Test Exclusion Restrictions

A central assumption in all econometric studies of treatment effects is that the treatment is uncorrelated with other determinants of the outcome, conditional on covariates. Although the assumption is ultimately untestable – the "fundamental problem of causal inference" (Holland, 1986) – the data can provide indications that it is unlikely to hold. In experiments, for example, significant correlations between treatment and pre-assignment variables are interpreted as evidence that randomization was unsuccessful. [2] Panel data can be particularly useful. A correlation between treatment and some pre-assignment variable X need not indicate bias in the estimated treatment effect if X is uncorrelated with the outcome variable of interest. But outcomes are typically correlated within individuals over time, so an association between treatment and the lagged outcome strongly suggests that the treatment is not exogenous with respect to post-treatment outcomes.

This insight has been most fully explored in the literature on the effect of job training on wages and employment. Today's wage or employment status is quite informative about tomorrow's, even after controlling for all observables. Evidence that assignment to job training is correlated with lagged wage dynamics indicates that simple specifications for the effect of training on outcomes are likely to yield biased estimates (Ashenfelter, 1978). Richer models of the training assignment process may absorb this correlation while permitting identification (Heckman et al., 1987). But even these models may impose testable restrictions on the relationship between treatment and the outcome history (Ashenfelter and Card, 1985; Card and Sullivan, 1988; Jacobson et al., 1993). Of course, these sorts of tests cannot diagnose all model violations. If treatment assignments depend on unobserved determinants of future outcomes that are uncorrelated with the outcome history, the treatment effect estimator may be biased even though treatment is uncorrelated with past outcomes.

In value added studies, the multiplicity of teacher "treatments" can blur the connection to program evaluation methods. But the utility of past outcomes for specification diagnostics carries over directly. Identification of a teacher's effect rests on assumptions about the relationship between the teacher assignment and the other determinants of future achievement, and the relationship with past achievement can be informative about the plausibility of these assumptions.

Only a few studies have attempted to validate VAMs. Jacob and Lefgren (2008) and Harris and Sass (2007) show that value added coefficients are weakly but significantly correlated with

[2] Similar tests are often used in non-experimental analyses: Researchers conducting propensity score matching studies frequently check for "balance" of covariates conditional on the propensity score (Rosenbaum and Rubin, 1984), and Imbens and Lemieux (2008) recommend analogous tests for regression discontinuity analyses.

principals' ratings of teacher performance. Of course, if principal decisions about classroom assignments create bias in the VAMs, causality could run from principal opinions to estimated value added rather than the reverse. More relevant to the current analysis, Kane and Staiger (2008) demonstrate that VAM estimates from observational data are approximately unbiased predictors of teachers' effects when students are randomly assigned. While I examine a closely related question to that considered by Kane and Staiger, my larger and more representative sample permits me to extend their analysis in two ways. First, I have much more statistical power. This enables me to identify biases that are substantively important but that lie well within Kane and Staiger's confidence intervals. Second, my sample resembles the sort that would be used for any VAM intended as a teacher compensation or retention tool. In particular, it includes teachers specializing in students (e.g., late readers) who cannot be readily identified and excluded from large scale analyses. The likely exclusion of such teachers from Kane and Staiger's sample quite plausibly avoids the most severe biases in observational VAM estimates.[3]

3 Statistical Model and Methods

3.1 Defining the Problem

I take the parameter of interest in value added modeling to be the effect on a student's test score at the end of grade g of being assigned to a particular grade-g classroom rather than another classroom at the same school. Later, I extend this to look at dynamic treatment effects (that is, the effect of the grade-g classroom on the $g+s$ score). I do not distinguish between *classroom* and *teacher* effects, and use the terms interchangably. In the Appendix, I consider this distinction, defining a teacher's effect as the time-invariant component of the effects of the classrooms taught by the teacher over several years.

[3]In the Kane and Staiger experiment, principals were given the name of one teacher and asked to identify a comparison teacher such that it would be appropriate to randomly assign students within the pair. One imagines that principals generally chose a comparison who was assigned similar students as the focal teacher in the pre-experimental data. Moreover, a substantial majority of principals declined to participate, perhaps because the initial teacher was a specialist for whom no similar comparison could be found.

I am interested in whether common VAMs identify classroom effects with arbitrarily large samples. I therefore sidestep small sample issues by considering the properties of VAM estimates as the number of students grows with the number of teachers (and classrooms) fixed.[4] If classroom effects are identified under these unrealistic asymptotics, VAMs may be usable in compensation and retention policy with appropriate allowances for the sampling errors that arise with finite class sizes;[5] if not, these corrections are likely to go awry.

A final important distinction is between identification of the variance of teacher quality and the identification of individual teachers' effects. I focus exclusively on the latter. It is impractical to report each of several thousand teachers' estimated effects, however. I therefore report only the implied standard deviations (across teachers) of teachers' actual and counterfactual effects, along with tests of the hypothesis that the teacher effects are all zero.[6]

3.2 Data Generating Process and the Three VAMs

I consider a relatively general specification of the educational production function, modeled on those used by Todd and Wolpin (2003) and Harris and Sass (2006), that allows student achievement to depend on the full history of inputs received to date plus the student's innate ability. Separating classroom effects from other inputs, I assume that the test score of student i at the end of grade g, A_{ig}, can be written as

$$A_{ig} = \alpha_g + \sum_{h=1}^{g} \beta_{hgc(i,h)} + \mu_i \tau_g + \sum_{h=1}^{g} \varepsilon_{ih} \phi_{hg} + \nu_{ig}. \tag{1}$$

Here, β_{hgc} is the effect of being in classroom c in grade h on the grade-g test score, and $c(i,h) \in \{1,\ldots,J_h\}$ indexes the classroom to which student i is assigned in grade h. μ_i is individual

[4]Under realistic asymptotics, the number of classrooms should rise in proportion to the number of students. If so, classroom effects are not identified under any exogeneity restrictions: Even in the asymptotic limit, the number of students per teacher remains finite and the sampling error in an individual teacher's effect remains non-trivial.

[5]A typical approach shrinks a teacher's estimated effect toward the population mean in proportion to the degree of imprecision in the estimate. The resulting empirical Bayes estimate is the best linear predictor of the teacher's true effect, given the noisy estimate. See McCaffrey et al. (2003), pp. 63-68.

[6]Rivkin et al. (2005) develop a strategy for identifying the variance of teachers' effects, but not the effect of individual teachers, under weaker assumptions than are required by the VAMs described below.

ability. We might expect the achievement gap between high-ability and low-ability students to grow over time; this would correspond to $\tau_g > \tau_h > 0$ for each $h < g$. ε_{ih} captures all other inputs in grade h, including those received from the family, non-classroom peers, and the community. It might also include developmental factors: A precocious child might have positive εs in early grades and negative εs in later grades as her classmates catch up. As this example shows, ε is quite likely to be serially correlated within students across grades. Finally, v_{ig} represents measurement error in the grade-g test relative to the student's "true" grade-g achievement. This is independent across grades within students.[7]

A convenient restriction on the time pattern of classroom effects is uniform geometric decay, $\beta_{hg'c} = \beta_{hgc}\lambda^{g'-g}$ for some $0 \leq \lambda \leq 1$ and all $h \leq g \leq g'$. A special case is $\lambda = 1$, corresponding to perfect persistence. Although my results do not depend on these restrictions, I impose them as needed for notational simplicity. I consider non-uniform decay in Section 7. Note that there is no theoretical basis for restrictions on the decay of non-classroom effects (i.e. on ϕ_{hg}).

It will be useful to adopt some simplifying notation. Let $\omega_{ig} \equiv \sum_{h=1}^{g} \varepsilon_{ih}\phi_{hg}$ be the composite grade-g residual achievement, and let Δ indicate first differences across student grades: $\Delta\beta_{hgc} \equiv \beta_{hgc} - \beta_{h,g-1,c}$, $\Delta\tau_g \equiv \tau_g - \tau_{g-1}$, $\Delta\omega_{ig} \equiv \omega_{ig} - \omega_{ig-1}$, and so on.

Tractable VAMs amount to decompositions of A_{ig} (or of $\Delta A_{ig} \equiv A_{ig} - A_{ig-1}$) into the current teacher's effect $\beta_{ggc(i,g)}$, a student heterogeneity component, and an error assumed to be orthogonal to the classroom assignment. Models differ in the form of this decomposition. In this paper I consider three specifications: A simple regression of gain scores on grade and contemporaneous classroom indicators,

VAM1: $\Delta A_{ig} = \alpha_g + \beta_{ggc(i,g)} + e_{1ig}$;

a regression of score levels on classroom indicators and the lagged score,

VAM2: $A_{ig} = \alpha_g + A_{ig-1}\lambda + \beta_{ggc(i,g)} + e_{2ig}$;

and a regression that stacks gain scores from several grades and adds student fixed effects,

[7] I define the β parameters to include any classroom-level component of v_{ig} and assume that v_{ig} is independent across students in the same classroom.

VAM3: $\Delta A_{ig} = \alpha_g + \beta_{ggc(i,g)} + \mu_i + e_{3ig}$.

All three are widely used.[8] VAM2 and VAM3 can both be seen as generalizations of VAM1: Constraining $\lambda = 1$ converts VAM2 to VAM1, while constraining $\mu_i \equiv 0$ converts VAM3.

Despite their similarity, the three VAMs rely on quite distinct restrictions on the process by which students are assigned to classrooms. I discuss the three in turn.

3.2.1 The gain score model (VAM1)

Differencing the production function (1), we can write the grade-g gain score as

$$\Delta A_{ig} = \Delta\alpha_g + \sum_{h=1}^{g-1} \Delta\beta_{hgc(i,h)} + \beta_{ggc(i,g)} + \mu_i \Delta\tau_g + \Delta\omega_{ig} + \Delta\nu_{ig}. \tag{2}$$

If we assume that teacher effects do not decay, $\Delta\beta_{hgc} = 0$ for all $h < g$. The error term e_{1ig} from VAM1 then has three components:

$$e_{1ig} = \mu_i \Delta\tau_g + \Delta\omega_{ig} + \Delta\nu_{ig}. \tag{3}$$

VAM1 will yield consistent estimates of the grade-g classroom effects if and only if, for each c,

$$E\left[e_{1ig} \,|\, c(i,g) = c\right] = 0. \tag{4}$$

Differences in last year's gains across this year's classrooms are informative about this restriction. Using (2), the average $g-1$ gain in classroom c is:

$$E\left[\Delta A_{ig-1} \,|\, c(i,g) = c\right] = \Delta\alpha_{g-1} + E\left[\beta_{g-1,g-1,c(i,g-1)} \,|\, c(i,g) = c\right] + E\left[e_{1ig-1} \,|\, c(i,g) = c\right]. \tag{5}$$

[8] The most widely used VAM, the Tennessee Value Added Assessment System (TVAAS; see Sanders et al., 1997), is specified as a mixed model for level scores that depend on the full history of classroom assignments, but this model implies an equation for annual gain scores of the form used in VAM1. VAM2 is more widely used in the recent economics literature. See, for example, Aaronson et al. (2007); Kane et al. (2006); Jacob and Lefgren (2008); and Goldhaber (2007). VAM3 was proposed by Boardman and Murnane (1979), and has been used recently by Rivkin et al. (2005); Harris and Sass (2006); Jacob and Lefgren (2008); and Boyd et al. (2007).

The first term is constant across c and can be neglected. The second term might vary with c if (for example) a principal compensates for a bad teacher assignment in grade $g-1$ with assignment to a better-than-average teacher in grade g. This can be absorbed by examining the across-$c(i,g)$ variation in ΔA_{ig-1} *controlling for* $c(i,g-1)$. I estimate specifications of this form below.[9] Any remaining variation across grade-g classrooms in $g-1$ gains, after controlling for $g-1$ classroom assignments, must indicate that students are sorted into grade-g classrooms on the basis of e_{1ig-1}.

Whether this would indicate a problem with assumption (4) depends on whether e_{1ig} is serially correlated. Equation (2) indicates four sources of potential serial correlation. First, ability appears in both e_{1ig} and e_{1ig-1} (unless $\Delta\tau_g = 0$). Second, the ε_{ig} process may be serially correlated. Third, even if ε is white noise, $\Delta\omega_{ig}$ is a moving average of order $g-1$ (absent strong restrictions on the ϕ coefficients). Finally, Δv_{ig} is an MA(1), degenerate only if $\text{var}(v) = 0$. Rothstein (2008) concludes that Δv_{ig} accounts for as much as 80% of the variance of ΔA_{ig}.

The discussion of serial correlation in e_{1ig} helps clarify the conditions in which (4) might hold. The most natural model that is consistent with (4) is for assignments to depend only on student ability, μ_i, and for ability to have the same effect on achievement in grades g and $g-1$ (i.e., $\Delta\tau_g = 0$). With these restrictions, VAM1 can be seen as the first-difference estimator for a fixed effects model, with strict exogeneity of classroom assignments conditional on μ_i. By contrast, (4) is not likely to hold if $c(i,g)$ depends, even in part, on ω_{ig-1}, v_{ig-1}, or A_{ig-1}.

[9] This is a test of the hypothesis that students are randomly assigned to grade-g classrooms *conditional on the* $g-1$ *classroom.* It has has zero power unless there is independent variation in $c(i,g-1)$ and $c(i,g)$. In the Tennessee STAR experiment, "streaming" was quite common, and in many schools there is zero independent variation in 3rd grade classroom assignments controlling for 2nd grade assignments. In this case, identification of teacher effects is possible only with strong maintained assumptions about decay (e.g., that $\beta_{ghc} = \lambda^{h-g}\beta_{ggc}$ for all $h \geq g$ with known λ; Nye et al. (2004) implicitly assume $\lambda = 1$).

3.2.2 The lagged score model (VAM2)

VAM2 frees up the coefficient on the lagged test score. If teacher effects decay geometrically at uniform rate $1-\lambda$, the grade-g score can be written in terms of the $g-1$ score:

$$A_{ig} = \breve{\alpha}_g + A_{ig-1}\lambda + \beta_{ggc(i,g)} + e_{2ig}, \tag{6}$$

where $\breve{\alpha}_g = \alpha_g - \alpha_{g-1}\lambda$, and

$$e_{2ig} = \mu_i \left(\tau_g - \tau_{g-1}\lambda\right) + \sum_{h=1}^{g-1} \varepsilon_{ih} \left(\phi_{hg} - \phi_{hg-1}\lambda\right) + \varepsilon_{ig} + \left(\nu_{ig} - \nu_{ig-1}\lambda\right). \tag{7}$$

As before, each of the terms in (7) is likely to be serially correlated. The VAM2 exclusion restriction, $E\left[e_{2ig} \,|\, c(i,g) = c\right] = 0$, would hold if grade-$g$ classroom assignments were random conditional on A_{ig-1}. It is unlikely to hold if assignments depend directly on e_{2ig-1} or on any of its components. In particular, $c(i,g)$ cannot depend on μ_i except through A_{ig-1}.[10]

The VAM2 exclusion restriction can again be evaluated by replacing the dependent variable, ΔA_{ig}, with its lag, ΔA_{ig-1}. By (6), the lagged score equals

$$A_{ig-1} = \breve{\alpha}_{ig-1} + A_{ig-2}\lambda + \beta_{g-1,g-1,c(i,g-1)} + e_{2ig-1}. \tag{8}$$

This can be rearranged to express the $g-2$ score in terms of the $g-1$ score and classroom:

$$A_{ig-2} = \frac{-1}{\lambda}\left[\breve{\alpha}_{ig-1} - A_{ig-1} + \beta_{g-1,g-1,c(i,g-1)} + e_{2ig-1}\right]. \tag{9}$$

Thus, the grade-g classroom assignment will have predictive power for the score in grade $g-2$, controlling for $g-1$ achievement,[11] if grade-g classrooms are correlated either with the $g-1$

[10] If $\tau_g - \tau_{g-1}\lambda$ is constant across g, (6) can be seen as a fixed effects model with a lagged dependent variable. λ and β_{gg} can be identified via IV or GMM (instrumenting for ΔA_{ig-1} in a model for ΔA_{ig}) if $c(i,g)$ depends on μ_i but is strictly exogenous conditional on this (Anderson and Hsiao, 1981; Arellano and Bond, 1991). See, e.g., Koedel and Betts (2007). Value added researchers typically apply OLS to (6). This is inconsistent for λ and identifies β_{ggc} only if $c(i,g)$ is random conditional on A_{ig-1}.

[11] The test can alternatively be expressed in terms of the "effect" of grade-g teachers on gains in $g-1$, controlling

teacher's effect (i.e. with $\beta_{g-1,g-1,c(i,g-1)}$) or with e_{2ig-1}. As in VAM1, the former can be ruled out by controlling for $g-1$ classroom assignments; the latter would indicate a violation of the VAM2 exclusion restriction if e_2 is serially correlated.

3.2.3 The fixed effects in gains model (VAM3)

The final VAM returns to the earlier assumption of zero decay of teachers' effects.[12] It incorporates the ability term in (2) into the estimating equation,

$$\Delta A_{ig} = \Delta\alpha_g + \beta_{ggc(i,g)} + \mu_i \Delta\tau_g + e_{3ig}, \tag{10}$$

leaving only two components in the error term, $e_{3ig} = \Delta\omega_{ig} + \Delta\nu_{ig}$.

Assuming that $\Delta\tau_g = 1$ for each g, (10) is a fixed effects model. The presence of the student fixed effect, combined with the small time dimension of student data sets, means that VAM3 requires stronger assumptions than the earlier models. An OLS regression with fixed effects is numerically equivalent to a regression of the de-meaned outcome on the de-meaned explanatory variables. The de-meaned grade-g gain is:

$$\Delta A_{ig} - \frac{1}{G}\sum_{h=1}^{G}\Delta A_{ih} = \left(\Delta\alpha_g - \frac{1}{G}\sum_{h=1}^{G}\Delta\alpha_h\right) + \left(\beta_{ggc(i,g)} - \frac{1}{G}\sum_{h=1}^{G}\beta_{hhc(i,h)}\right) + \left(e_{3ig} - \frac{1}{G}\sum_{h=1}^{G}e_{3ih}\right). \tag{11}$$

The equation for the de-meaned gain score thus has a grade-specific intercept and coefficients for all classroom assignments in grades 1 through G. Importantly, the error terms from all grades enter into (11). Thus, correlation between the classroom assignment in one grade and the error term in that or any other grade would bias the estimated β coefficients, even in large samples. To avoid bias, teacher assignments must be strictly exogenous conditional on μ_i.[13]

for the score in $g-1$.

[12]While VAM1 and VAM2 can easily be generalized to allow for non-uniform decay, VAM3 cannot.

[13]For practical value added implementations, G is rarely larger than three or four, so large-G asymptotics are infeasible. Without strict exogeneity, one small-G approach is to focus on the first difference of (10). When $G > 2$, OLS estimation of the first-differenced equation requires only that $c(i,g)$ be uncorrelated with e_{3ig-1}, e_{3ig}, and

Conditional strict exogeneity means that the same information, μ_i or some function of it, is used to make teacher assignments in each grade. This requires, in effect, that principals decide on classroom assignments for the remainder of a child's career before she starts kindergarten. If teacher assignments are updated each year in response to the student's performance during the previous year, strict exogeneity is violated.

The extension of my test to the strict exogeneity assumption in VAM3 is a direct application of Chamberlain's (1984) correlated random effects model. If ability enter's each year's gain score, this could explain an apparent effect of (for example) 5th grade teachers on 4th grade gains in VAM1: Both 5th grade teacher assignments and 4th grade gains depend on μ. My VAM3 test exploits the fact that 3rd grade gains also depend on the scalar μ_i. So 5th grade teachers who appear to have positive effects on 4th grade gains – because they are assigned high-μ students – should also appear to have positive effects on 3rd grade gains. An indication that a 5th grade teacher has different effects on 3rd and 4th grade gains would indicate that omitted time-varying determinants of gains are correlated with teacher assignments, and therefore that assignments are not strictly exogenous.

Formally, consider a projection of μ onto the full sequence of classroom assignments:

$$\mu_i = \xi_{1c(i,1)} + \ldots + \xi_{Gc(i,G)} + \eta_i. \tag{12}$$

ξ_{hc} is the incremental information about μ_i provided by the knowledge that the student was in classroom c in grade h, conditional on classroom assignments in all other grades. Substituting (12) into (10), we obtain

$$\Delta A_{ig} = \Delta\alpha_g + \sum_{h=1}^{G} \pi_{hgc(i,h)} + \eta_i + e_{3ig}, \tag{13}$$

where $\pi_{ggc} = \xi_{gc}\Delta\tau_g + \beta_{ggc}$ and $\pi_{hgc} = \xi_{hc}\Delta\tau_g$ for $h \neq g$. Under conditional strict exogeneity,

e_{3ig+1}. Though this is weaker than strict exogeneity, it is difficult to imagine an assignment process that would satisfy one but not the other. Another option is IV/GMM (see note 10), instrumenting for both the g and $g-1$ classroom assignments. Satisfactory instruments are not apparent.

$E[e_{3ih} \,|\, c(i,1), \ldots, c(i,G)] = 0$ for each h, and the fact that (12) is a linear projection ensures that η_i is uncorrelated with the regressors as well. An OLS regression of grade-g gains onto classroom indicators in grades 1 through G thus estimates the π_{hgc} coefficients without bias. When $G \geq 3$, the underlying parameters are overidentified. To see this, note that

$$\pi_{31} = \xi_3 \Delta\tau_1 = \xi_3 \Delta\tau_2 \frac{\Delta\tau_1}{\Delta\tau_2} = \pi_{32} \frac{\Delta\tau_1}{\Delta\tau_2}. \tag{14}$$

$\Delta\tau_1$ and $\Delta\tau_2$ are scalars, so (14) represents $J_3 - 1$ overidentifying restrictions on the $2J_3$ elements of the π_{31} and π_{32} vectors.[14]

Equation (14) implies that the elements of π_{31} should be perfectly correlated with the corresponding elements of π_{32} (or, if $\Delta\tau_1/\Delta\tau_2 < 0$, perfectly negatively correlated), so the correlation between elements of the estimated coefficient vectors $\hat{\pi}_{31}$ and $\hat{\pi}_{32}$ should be close to 1 (or -1). A formal test uses optimal minimum distance (OMD), minimizing

$$D = \left(\begin{pmatrix} \hat{\pi}_{31} \\ \hat{\pi}_{32} \end{pmatrix} - \begin{pmatrix} \xi_3 \Delta\tau_1 \\ \xi_3 \Delta\tau_2 \end{pmatrix} \right)' W^{-1} \left(\begin{pmatrix} \hat{\pi}_{31} \\ \hat{\pi}_{32} \end{pmatrix} - \begin{pmatrix} \xi_3 \Delta\tau_1 \\ \xi_3 \Delta\tau_2 \end{pmatrix} \right) \tag{15}$$

over the vector ξ_3 and the scalars $\Delta\tau_1$ and $\Delta\tau_2$. When W is the sampling variance of $\left(\hat{\pi}'_{31}\, \hat{\pi}'_{32}\right)'$, D is distributed χ^2 with $J_3 - 1$ degrees of freedom under the null hypothesis of strict exogeneity.[15] If D is above the 95% critical value from this distribution, the null is rejected.

3.3 Implementation

To put the three VAMs in the best possible light, I focus on estimation of within-school differences in classroom effects. For many purposes, one might want to make across-school com-

[14] There are $J_1 - 1$ additional overidentifying restrictions created by a similar proportionality relationship between π_{12} and π_{13}: Past teachers should have similar effects on all future grades' gains. These restrictions might fail either because strict exogeneity is violated or because teachers' effects decay (that is, $\beta_{12} \neq \beta_{13}$). I therefore focus on restrictions on the *future* teacher coefficients, as these provide sharper tests of strict exogeneity.

[15] Although there are $J_3 - 2$ parameters to be estimated, they are underidentified: Multiplying ξ_3 by a constant and dividing $\Delta\tau_1$ and $\Delta\tau_2$ by the same constant does not change the fit. In the implementation, I normalize $\Delta\tau_1 = 1$.

parisons. But students are not randomly assigned to schools, and those at one school may gain systematically faster than those at another for reasons unrelated to teacher quality. Random assignment to classrooms within schools is at least somewhat plausible. To isolate within-school variation, I augment each of the estimating equations discussed above with a set of indicators for the school attended.[16] The tests for VAM1 and VAM2 then amount to tests of whether students are (conditionally) randomly assigned to classrooms *within schools*. They resemble tests of successful randomization in stratified experiments, treating schools as strata.

Intuitively, I will reject random assignment if replacing a set of school indicators with 5th grade classroom indicators adds more explanatory power for 4th grade gains than would be expected by chance alone. Let S_5 and T_5 be matrices of indicators for 5th grade classrooms and schools. These are collinear, so to eliminate this I define $\tilde{T}_5$ as the submatrix of T_5 that results from excluding the columns corresponding to one classroom per school. The VAM1 test is based on a simple regression:

$$\Delta A_4 = \alpha + S_5\delta + \tilde{T}_5\beta + \varepsilon. \tag{16}$$

The identifying assumption of VAM1 is rejected if $\beta \neq 0$. I use a heteroskedasticity-robust score test (Wooldridge, 2002, p. 60) to evaluate this. I also estimate versions of (16) that include controls for 4th grade classroom assignments. To test VAM2, I replace the dependent variable in (16) with A_3 and add a control for A_4 on the right-hand-side.

It is clear from the definition of $\tilde{T}_5$ that only schools with multiple classrooms per grade can contribute to the analysis. One might be concerned that schools with only two or three classrooms will be misleading. The Appendix presents a Monte Carlo analysis of the VAM1 and VAM2 tests in schools of varying sizes. The VAM1 test has appropriate size even with just two classrooms per school, so long as the number of students per classroom is large. (Recall that I focus on large-class asymptotics.) With small classes, the asymptotic distribution of the test

[16] This makes W singular in (15). For the OMD analysis of VAM3, I drop the elements of π_{gh} that correspond to the largest class at each school.

statistic is an imperfect approximation, and as a result the test over-rejects slightly. When there are 20 students per class, the test of VAM1 has size around 10%. With empirically reasonable parameter values, the VAM2 test performs similarly.[17,18]

I also report the standard deviation of the teacher coefficients (the βs in (16)) themselves. The standard deviation of these coefficients necessarily exceeds that which would be obtained were the coefficients estimated without error. Aaronson et al. (2007) propose a simple estimator for the variance of the true coefficients – those that would be identified with large samples of students per teacher – across teachers. Let γ be a mean-zero J-vector of true projection coefficients and let $\hat{\gamma}$ be an unbiased finite-sample estimate of γ, with $E\left[\gamma'\left(\hat{\gamma}-\gamma\right)\right]=0$. The variance (across elements) of γ can be written as:

$$E\left[\gamma'\gamma\right]=E\left[\hat{\gamma}'\hat{\gamma}\right]-E\left[\left(\hat{\gamma}-\gamma\right)'\left(\hat{\gamma}-\gamma\right)\right]. \tag{17}$$

$E\left[\hat{\gamma}'\hat{\gamma}\right]$ is simply the variance across teachers of the coefficient estimates.[19] $E\left[\left(\hat{\gamma}-\gamma\right)'\left(\hat{\gamma}-\gamma\right)\right]$ is the average heteroskedasticity-robust sampling variance. Both are weighted by the number of students taught.

Specifications that include indicators for classroom assignments in several grades simultaneously – like that used for the test of VAM3 – introduce two complications. I discuss them briefly here, then in more detail in the Appendix. First, school indicators in several grades are identified only from students who switch schools between grades. School switching is likely to be endogenous to a variety of unobserved student characteristics. In specifications containing classroom assignments from multiple grades, I restrict my sample to students who do not switch

[17] When students are assigned to classrooms based on the lagged score and when this score incorporates implausibly high degrees of clustering at the 4th grade classrom level, the VAM2 test rejects at high rates even with large classes. This reflects my use of a test that assumes independence of residuals within schools. Unfortunately, it is not possible to allow for dependence, as clustered variance-covariance matrices are consistent only if the number of clusters grows with the number of parameters fixed (Kezdi, 2004) and in my application, the number of parameters grows with the number of clusters.

[18] Kinsler (2008) claims that the VAM3 test also over-rejects in simulations. In personal communication, he reports that the problem disappears with large classes.

[19] $\hat{\gamma}$ is normalized to have mean zero across teachers at the same school, and its variance is adjusted for the degrees of freedom that this consumes.

schools, and include only a single set of school indicators.

Second, the coefficients for teachers in different grades can only be separately identified when there is sufficient shuffling of students between classrooms. If students are perfectly streamed – if a student's classmates in 4th grade were also her classmates in 3rd grade – the 3rd and 4th grade classroom indicators are collinear. I exclude from my samples a few schools where inadequate shuffling leads to perfect collinearity.

4 Data and Sample Construction

The specifications described in Section 3 require longitudinal data that track students' outcomes across several grades, linked to classroom assignments in each grade. I use administrative data on elementary students in North Carolina public schools, assembled and distributed by the North Carolina Education Research Data Center. These data have been used for several previous value added analyses (see, e.g., Clotfelter et al., 2006; Goldhaber, 2007).

I examine end-of-grade math and reading tests from grades 3 through 5. To construct the 3rd grade gain, I use "pre-tests" given at the beginning of 3rd grade in place of 2nd grade scores, which were not given. I standardize the scale scores separately for each subject-grade-year combination.[20]

The North Carolina data identify the school staff member who administered the end-of-grade tests. In the elementary grades, this was usually the regular teacher. Following Clotfelter et al. (2006), I count a student-teacher match as valid if the test administrator taught a "self-contained" (i.e. all day, all subject) class for the relevant grade in the relevant year, if that class was not designated as special education or honors, and if at least half of the tests that the teacher administered were to students in the correct grade. Using this definition, 73% of 5th graders can be matched to teachers. In each of my analyses, I restrict the sample to students with valid

[20]The original score scale is meant to ensure that one point corresponds to an equal amount of learning at each grade and at each point in the within-grade distribution. Rothstein (2008) and Ballou (2008) emphasize the importance of this property for value added modeling. All of the results here are robust to using the original scale.

teacher matches in all grades for which teacher assignments are controlled.

I focus on the cohort of students who were in 5th grade in 2000-2001. Beginning with the population (N=99,071), I exclude students who have inconsistent longitudinal records (e.g. gender changes between years); who were not in 4th grade in 1999-2000; who are missing 4th or 5th grade test scores; or who cannot be matched to a 5th grade teacher. I additionally exclude 5th grade classrooms that contain fewer than 12 sample students or are the only included classroom at the school. This leaves my base sample, consisting of 60,740 students from 3,040 5th grade classrooms and 868 schools.

My analyses all use subsets of this sample that provide sufficient longitudinal data. In analyses of 4th grade gains, for example, I exclude students who have missing 3rd grade scores or who were not in 3rd grade in 1998-1999. In specifications that include identifiers for teachers in multiple grades, I further exclude students who changed schools between grades, plus a few schools where streaming produces perfect collinearity.

Table 1 presents summary statistics. I show statistics for the population, for the base sample, and for my most restricted sample (used for estimation of equation (13)). The last is much smaller than the others, largely because I require students to have attended the same school in grades 3 through 5 and to have valid teacher matches in each grade. Table 1 indicates that the restricted sample has higher mean 5th grade scores than the full population. This primarily reflects the lower scores of students who switch schools frequently.[21] Average 5th grade gains are similar across samples. The Appendix describes each sample in more detail.

As discussed above, my tests can be applied only if there is sufficient re-shuffling of classrooms between grades. An Appendix table shows the fraction of students' 5th grade classmates who were also in the same 4th grade classes, by the number of 4th grade classes at the school. Complete reshuffling (combined with equally-sized classes) would produce 0.5 with two classes, 0.33 with three, and so on. The actual fractions are larger than this, but only

[21]Table 1 shows that average 3rd and 4th grade scores in the "population" are well above zero. The norming sample that I use to standardize scores in each grade consists of all students in that grade in the relevant year (i.e. of all 3rd graders in 1999), while only those who make normal progress to 5th grade in 2001 are included in the sample for Columns 1-2. The low scores of students who repeat grades account for the discrepancy.

slightly. In schools with exactly three 5th grade teachers, for example, 35% of students' 5th grade classmates were also their classmates in 4th grade. In only 7% of multiple-classroom schools do the 4th and 5th grade classroom indicators have deficient rank.

Table 2 presents the correlation of test scores and gains across grades and subjects. The table indicates that 5th grade scores are correlated above 0.8 with 4th grade scores in the same subject, while correlations with scores in earlier grades or other subjects are somewhat lower. 5th grade gains are strongly negatively correlated with 4th grade levels and gains in the same subject and weakly negatively with those in the other subject. The correlations between 5th and 3rd grade gains are small but significant both within and across subjects.

VAM3 is predicated on the notion that student ability is an important component of annual gains. Assuming that high-ability students gain faster, this would imply positive correlations between gains in different years. There is no indication of this in Table 2. One potential explanation is that noise in the annual tests introduces negative autocorrelation in gains, but Rothstein (2008) concludes that even true gains are negatively autocorrelated. This strongly suggests that VAM3 is poorly suited to the test score data generating process.

5 Results

Tables 3, 4, and 5 present results for the three VAMs in turn. I begin with VAM1, in Table 3. I regress 5th grade math and reading gains (in Columns 1 and 2, respectively) on indicators for 5th grade schools and classrooms, excluding one classroom per school. In each case, the hypothesis that all of the classroom coefficients are zero (i.e. that classroom indicators have no explanatory power beyond that provided by school indicators) is decisively rejected. The VAM indicates that the within-school standard deviations of 5th grade teachers' effects on math and reading are 0.15 and 0.11, respectively. This is similar to what has been found in other studies (e.g., Aaronson et al., 2007; Rivkin et al., 2005).

Columns 3 and 4 present falsification tests in which 4th grade gains are substituted for the 5th grade gains as dependent variables, with the specification otherwise unchanged. The standard deviation of 5th grade teachers' "effects" on 4th grade gains is 0.08 in each subject, and the hypothesis of zero association is rejected in each specification. In both the standard deviation and statistical significance senses, 5th grade classroom assignments are slightly more strongly associated with 4th grade reading gains than with math gains.

One potential explanation for these counterfactual effects is that they represent omitted variables bias deriving from my failure to control for 4th grade teachers. Columns 5-8 present estimates that do control for 4th grade classroom assignments, using a sample of students who attended the same school in 4th and 5th grades and can be matched to teachers in each grade. Two aspects of the results are of interest. First, 4th grade teachers have strong independent predictive power for 5th grade gains. This is at least suggestive that the "zero decay" assumption is violated. I return to this in Section 7. Second, the coefficients on 5th grade classroom indicators in models for 4th grade gains remain quite variable – even more so than in the sparse specifications in Columns 3 and 4 – and are significantly different from zero. Evidently, the correlation between 5th grade teachers and 4th grade gains derives from sorting on the basis of the 4th grade *residual*, not merely from between-grade correlation of teacher assignments.

These results strongly suggest that the exclusion restrictions for VAM1 are violated. To demonstrate this conclusively, however, we need to show that the residual in VAM1, e_{1ig}, is serially correlated. To examine this, I re-estimated VAM1 for 4th grade teachers' effects on 4th grade gains. The correlation between $\hat{e}_{1i4}$ and $\hat{e}_{1i5}$ is -0.38 in math and -0.37 in reading.

The negative serial correlation of e_1 implies that students with high gains in 4th grade will tend to have low gains in 5th grade, and vice versa. Because VAM1 evidently does not adequately control for classroom assignments, it gives unearned credit to teachers who are assigned students who did poorly in 4th grade, as these students will predictably post unusually high 5th grade gains when they revert toward their long-run means. Similarly, teachers whose students did unusually well in 4th grade will be penalized by the students' fall back toward their long-run means in 5th grade. Indeed, an examination of the VAM1 coefficients indicates that 5th grade teachers whose students have above-average 4th grade gains have systematically lower estimated value added than teachers whose students underperformed in the prior year. Importantly, this pattern is stronger than can be explained by sampling error in the estimated teacher effects; it reflect true mean reversion and not merely measurement error.

Table 4 repeats the falsification exercise for VAM2. The structure is identical to that of Table 3. Columns 1 and 2 present estimates of the basic VAM for 5th grade teachers' effects on 5th grade scores, controlling for 4th grade math and reading scores. The standard deviations of 5th grade teachers' effects are nearly identical to those in Table 3. Columns 3 and 4 substitute 3rd grade scores as the dependent variable. Once again, we see that 5th grade teachers are strongly predictive, more so in reading than in math. Columns 5-8 augment the specification with controls for 4th grade teachers. The 5th grade teacher coefficients are no longer jointly significant in the 3rd grade math score specification, though they remain quite large in magnitude. They are still highly significant in the specification for 3rd grade reading scores.[22]

The VAM2 residuals, like those from VAM1, are non-trivially correlated between 4th and 5th grades, -0.21 in math and -0.19 in reading. They are also correlated across subjects: -0.14 between 4th grade reading and 5th grade math. Thus, the evidence that 5th grade teacher assignments are correlated with the 4th grade residuals indicates that the VAM2 exclusion restriction is violated, regardless of whether the dependent variable is the math or the reading score. As before, 5th grade teachers' effects on 5th grade scores are negatively correlated with their counterfactual "effects" on 4th grade gains, suggesting that mean reversion in student achievement – combined with non-random classroom assignments – is an important source of bias in VAM2.

To implement the VAM3 falsification test, I begin by selecting the subsample with non-missing 3rd and 4th grade gains; valid teacher assignments in grades 3, 4, and 5; and continuous enrollment at the same school in all three grades. I exclude 26 schools where the three sets of indicators for teachers in grades 3, 4, and 5 (dropping one teacher in each grade from each

[22] As noted earlier, Monte Carlo analyses indicate that my tests over-reject slightly in small samples. In samples resembling those used here, less than 5% of the simulations yield p-values below 0.025. That the observed p-values are generally well below this suggests that the results are unlikely to be spurious.

school) are collinear. I then regress both the 3rd and 4th grade gains on school indicators and on each of the three sets of teacher indicators.[23]

Table 5 reports estimates for math gains, in Columns 1 and 2, and for reading gains, in Columns 4 and 5. The first panel shows the standard deviations (adjusted for sampling error) of the coefficients for each grade's teachers. Gains in each subject and in each grade are substantially correlated with classroom assignments in all three grades. Although p-values are not shown, in all 12 cases the hypothesis of zero effects is rejected. Columns 3 and 6 report the across-teacher correlations between the coefficients in the models for 3rd and 4th grade gains (i.e., between π_{g3} and π_{g4}). The most important correlation is that for 5th grade teachers, -0.04 for math and -0.06 for reading. Recall that strict exogeneity implies that the 5th grade teacher coefficients in the model for 4th grade gains should be proportional to the corresponding coefficients in the model for 3rd grade gains, $\pi_{54} = \left({}^{\Delta\tau_4}\!/\!_{\Delta\tau_3}\right)\pi_{53}$, implying a correlation of ± 1. The near-zero correlations strongly suggest that a single ability factor is unable to account for the apparent "effects" of 5th grade teachers on gains in earlier grades.

Indeed, these correlations are direct evidence against the VAM3 identifying assumption of conditional strict exogeneity. The lower panel of Table 5 presents OMD estimates of the restricted model.[24] For math scores, the estimated ratio ${}^{\Delta\tau_4}\!/\!_{\Delta\tau_3}$ is 0.14, implying that student ability is much more important to 3rd grade than to 4th grade gains. Thus, the constrained estimates imply negligible coefficients for 5th grade teachers in the equation for 4th grade gains, and do a very poor job of fitting the unconstrained estimate of the standard deviation of these coefficients, 0.099. The test statistic D is 2,136, and the overidentifying restrictions are overwhelmingly rejected. In the reading specification, the ${}^{\Delta\tau_4}\!/\!_{\Delta\tau_3}$ ratio is close to one, and the restricted model allows for meaningful coefficients on 5th grade teachers in both the 3rd and 4th grade gain equations, albeit much less variability than is seen in the unconstrained model.

[23]It is not essential to the correlated random effects test that the full sequence of teacher assignments back to grade 1 be observed, but the test may over-reject if classroom assignments in grades 3-5 are correlated with those in 1st and 2nd grade and if the latter have continuing effects on 3rd and 4th grade gains. Recall, however, that VAM3 assumes such lagged effects away.

[24]The OMD analysis uses a variance-covariance matrix W that is robust to arbitrary heteroskedasticity and within-student, between-grade clustering. See the Appendix.

But the test statistic is even larger here, and the restricted model is again rejected. We can thus conclude that 5th grade teacher assignments are not strictly exogenous with respect to either math or reading gains, even conditional on single-dimensional (subject-specific) student heterogeneity. The identifying assumption for VAM3 is thus violated.

The results in Tables 3, 4, and 5 indicate that all three of the VAMs considered here rely on incorrect exclusion restrictions – teacher assignments evidently depend on the past learning trajectory even after controlling for student ability or the prior year's test score. It is possible, however, that slight modifications of the VAMs could eliminate the endogeneity. I have explored several alternative specifications to gauge the robustness of the results. I have re-estimated VAM1 and VAM2 with controls for student race, gender, free lunch status, 4th grade absences, and 4th grade TV viewing; these have no effect on the tests. The three VAMs also continue to fail falsification tests when I use the original score scales or score percentiles in place of standardized-by-grade scores, or when I used data from other cohorts. As a final investigation, I have extended the tests to evaluate VAM analyses that use data from multiple cohorts of students to distinguish between permanent and transitory components of a teacher's "effect." As discussed in the Appendix, the implicit assumptions under which this can avoid the biases identified here do not appear to hold in the data.

6 How Much Does This Matter?

The results in Section 5 indicate that the identifying assumptions for all three VAMs are violated in the North Carolina data. However, if classroom assignments nearly satisfy the assumptions underlying the VAMs, the models might yield almost unbiased estimates of teachers' causal effects. In this Section, I use the degree of sorting on prior outcomes to quantify the magnitude of the biases resulting from non-random assignments. I focus on VAM1 and VAM2, as the lack of correlation between 3rd and 5th grade gains (Table 2) strongly suggests that the additional complexity and strong maintained assumptions of VAM3 are unnecessary.

In general, classroom assignments may depend both on variables observed by the econometrician and on unobserved factors. The former can in principle be incorporated into VAM specifications. Accordingly, the first part of my investigation focuses on the role of observable characteristics that are omitted from VAM1 and VAM2. I compare VAM1 and VAM2 to a richer specification, VAM4, that controls for teacher assignments in grades 3 and 4, end-of-grade scores in both subjects in both grades, and scores from the tests given at the beginning of 3rd grade. This would identify 5th grade teachers' effects if assignments were random conditional on the test score and teacher assignment history. It is thus more general than VAM2. It does not strictly nest VAM1, however: Assignment of teachers based purely on student ability (μ_i) would satisfy the VAM1 exclusion restriction but not that for VAM4. If assignments depend on both ability and lagged scores, VAM1, VAM2, and VAM4 are all misspecified.

Table 6 presents the comparisons. The first rows show the estimated standard deviations of teachers' effects obtained from VAM1 and VAM2, as applied to the subset of students with complete test score histories and valid teacher assignments in each prior grade. The unadjusted estimates are somewhat higher than those in Tables 3 and 4, as the smaller sample yields noisier estimates, but the sampling-adjusted estimates are quite similar to those seen earlier. The next two rows of the Table show estimates from the richer specification. Standard deviations are somewhat larger, but not dramatically so.

The final two rows describe the bias in the simpler VAMs relative to VAM4 (that is, $\beta_{55}^{VAM1} - \beta_{55}^{VAM4}$ and $\beta_{55}^{VAM2} - \beta_{55}^{VAM4}$). I again show both the raw standard deviation of the point estimates and an adjusted standard deviation that removes the portion due to sampling error. For VAM1, the bias has a standard deviation over a third as large as that of the VAM4 effects. For VAM2, which already includes a subset of the controls in VAM4, the bias is somewhat smaller. For both VAMs, the bias is more important in estimates of teachers' value added for math scores than for reading scores.

Of course, the exercise carried out here can only diagnose bias in VAM1 and VAM2 from selection on *observables* – variables that can easily be included in the VAM specification. In a companion paper (Rothstein, forthcoming), I attempt to quantify the bias that is likely to result from selection on unobservables. Following the intuition of Altonji et al. (2005) that the weight of observable (to the econometrician) and unobservable variables in classroom assignments is likely to mirror their relative weights in predicting achievement, one can use the degree of sorting on observables to estimate the importance of unobservables and therefore the magnitude of the bias in estimated teacher effects. Under varying assumptions about the amount of information that parents and principals have, I find that the bias from non-random assignments is quite plausibly 75% as large (in standard deviation terms) as the estimates of teachers' effects in VAM1, and perhaps half this large in VAM2.[25]

To provide a better sense of the import of non-random classroom assignments for the value of VAMs in teacher compensation and retention decisions, I simulate true and estimated teacher effects with joint distributions resembling those reported in Table 6 and in Rothstein (forthcoming). For each of several scenarios characterizing the assignment of students to classrooms, I generate 10,000 teachers' true effects and coefficients from VAMs 1, 2, and 4.[26] I assume that true effects and biases are both normally distributed, and that the VAM coefficients are free of sampling error. I then compute three statistics to summarize the relationship of the VAM estimates to teachers' true effects: the correlation between teachers' true effects and the VAM coefficients, the rank correlation, and the fraction of teachers with true effects in the top quintile who are indicated to be in the top quintile by the VAMs.

Results are presented in Table 7. Each panel corresponds to a distinct assumption about the classroom assignment process. In the first panel, I assume that selection is solely on the basis of the observed test score history. Using the model for reading scores from Table 6, the standard deviation of teachers' true effects is 0.148, and the standard deviations of the biases in VAM1 and VAM2 are 0.054 and 0.028, respectively. Columns 4-6 show the reliability of teacher quality under different metrics. True effects and ranks are very highly correlated with

[25] Kane and Staiger's (2008) comparison of experimental and non-experimental value added estimates would be unlikely to detect biases of this magnitude.

[26] It is not possible to use the estimates from Table 6 directly because I wish to abstract from the role of sampling error. The simulation is described in greater detail in the Appendix.

the effects and ranks indicated by VAMs 1 and 2. 79 to 90% of teachers who are in the top quintile of the actual quality distribution are judged to be so by the simple VAMs.

But this analysis assumes, implausibly, that selection is solely on observables. Panels B-E present alternative estimates that allow variables that are not controlled even in VAM4 to play a role in classroom assignments, as in Rothstein (forthcoming). In Panel B, I assume that classroom assignments depend both on the test score history that is reported in my data and a on second, unobserved history (e.g., student grades) that provides an independent, equally noisy measure of the student's trajectory through grades 2-4. Allowing for this moderate degree of selection-on-unobservables notably degrades the performance of VAM1, but VAM2 and VAM4 continue to perform reasonably well. In Panel C, I assume that there are two separate unobserved achievement measures. Performance degrades still further; while the correlations between true effects and the VAM2 and VAM4 estimates remain large, only about four fifths of top-quintile teachers are judged to be so by the two VAMs.

Panel D allows for even more unobserved information to be used in classroom assignments: I assume that the principal knows the student's true achievement in grades 2-4. Now, even VAM4 is correlated less than 0.9 with teachers' true effects, and less than three quarters of true top-quintile teachers get top-quintile ratings from any the VAMs. Finally, Panel E presents an extreme scenario corresponding to Altonji et al.'s (2005) assumption that selection on unobservables is like selection on observables. This is not realistic, as principals cannot perfectly predict student achievement, but it provides a useful bound for the degree of bias that non-random classroom assignments might produce in VAM-based estimates. This bound is tight enough to be informative: Even in this worst case, the VAMs retain some signal, and VAM2 and VAM4 continue to correctly classify over half of top-quintile teachers.

It is difficult to know which of the scenarios is the most accurate. Panel E likely assumes too much sorting on unobservables, while Panel A almost certainly assumes too little. The truth almost certainly lies in between, perhaps resembling the scenarios depicted in Panels B and C. These suggest that VAMs that control only for past test scores – typically the only available variables – have substantial signal but nevertheless introduce important misclassification into any assessment of teacher quality. Only 60-80% of the highest quality teachers will receive rewards given on the basis of high VAM scores.

Moreover, Table 7 omits three major sources of error in VAM-based quality measures that would magnify the misclassification rates seen there. First, I have suppressed the role of sampling error that would inevitably arise in VAM-based estimates. It is well documented (Lockwood et al., 2002; McCaffrey et al., 2008) that this alone produces high misclassification rates. Second, all of the analyses in this paper are based on comparisons of teachers within schools. Like most other value added studies, I make no effort to measure across-school differences in teacher quality. But most policy applications of value added would require comparisons across as well as within schools. Because students are not even approximately randomly assigned to schools, these comparisons are likely to be less informative about causal effects than are the within-school comparisons considered here.

Finally, I have assumed that teachers' effects on their students' end-of-grade scores are the sole outcome of interest. This may be incorrect. In particular, if teachers can allocate effort between teaching-to-the-test and raising students' long-run learning trajectories (e.g., by working to instill a love of reading), one would like to reward the second rather than the first. This suggests that the effects that matter may be those on students' long-run outcomes rather than on their end-of-grade scores. I consider this issue in the next Section.

7 Short-Run vs. Long-Run Effects

Recall from Columns 5-6 of Tables 3 and 4 that 4th grade teachers appear to have large effects on students' 5th grade gains. Given the results for 4th grade gains, these "effects" cannot be treated as causal. But setting this issue aside, we can use the lagged teacher coefficients to evaluate restrictions on time pattern of teachers' effects (that is, on the relationship between β_{gg} and $\beta_{g,g+s}$ in the production function (1)) that are universally imposed in value added analyses.

When only a single grade's teacher assignment is included, VAM2 implicitly assumes that teachers' effects decay at a uniform, geometric rate ($\beta_{g,g+s} = \beta_{gg}\lambda^s$ for $\lambda \in [0, 1]$), while VAM1 assumes zero decay ($\lambda = 0$). It is not clear that either restriction is reasonable.[27] While several studies have estimated λ,[28] all have done so under the restriction that decay is uniform. As a final investigation, I analyze the validity of this restriction by comparing a grade-g teacher's initial effect in grade g with her longer-run effect on scores in grade $g+1$ or $g+2$, without restricting the relationships among them.[29] If in fact teachers' effects decay uniformly, the initial and longer-run effects should be perfectly correlated (except for sampling error).

I begin by estimating VAM1 and VAM2 for 3rd, 4th, and 5th grade scores or gains, augmenting each specification with controls for past teachers back to 3rd grade. I then compute 3rd and 4th grade teachers' cumulative effects over one, two, and (for 3rd grade teachers) three years. Table 7 presents summary statistics for these cumulative effects. I show their standard deviation and their correlation with the initial effects β_{ggc}, both adjusted for sampling error. Two aspects of the results are of note. First, there is much more variation in 4th grade teachers' effects on 4th grade scores than in those same teachers' effects on 5th grade scores. With uniform decay at rate $(1-\lambda)$, $\text{var}(\beta_{g,g+s}) = \lambda^{2s}\,\text{var}(\beta_{gg})$, so this is consistent with the mounting recent evidence that teachers' effects decay importantly in the year after contact (Andrabi et al., 2009; Kane and Staiger, 2008; Jacob et al., 2008). Second, the correlation between teachers' first year effects and their two year cumulative effects is much less than one, ranging between 0.33 and 0.51 depending on the model and subject. Correlations with three-year cumulative effects are (mostly) lower, centered around 0.4. This is not even approximately consistent with uniform decay. Even if we assume that the VAM-based estimates can be treated as causal, a

[27] Although a full discussion is beyond the scope of this paper, assumptions about "decay" are closely related to issues of test scaling and content coverage (Rothstein, 2008; Ballou, 2008; Martineau, 2006). To illustrate, consider a 3rd grade teacher who focuses on addition and subtraction. This will raise her students' 3rd grade scores but may do little for their performance on a 5th grade multiplication test.

[28] See, e.g., Andrabi et al. (2009), Sanders and Rivers (1996), and Konstantopoulos (2007).

[29] For VAM1, the effect of being in classroom c in grade g on achievement in grade $g+s$ is simply $\sum_{t=0}^{s} \beta_{g,g+t,c}$. In VAM2, the presence of a lagged dependent variable complicates the calculation of cumulative effects. If only the same-subject score is controlled, the effect of 3rd grade teacher c on 5th grade achievement is $(\beta_{33c}\lambda + \beta_{34c})\,\lambda + \beta_{35c}$. A similar but more complex expression characterizes the effects when lagged scores in both math and reading are controlled, as in my estimates.

teacher's first year effect is a poor proxy for her longer-run impact.

The final panel of Table 7 explores the implications of this analysis for teacher quality measurement. I use the estimates in Table 8 as parameters for my simulation to compare traditional end-of-year VAM coefficients to teachers' longer-run (two year) effects, treating the latter as the "truth." The results are not encouraging. Correlations are well below 0.5, and only about a third of teachers in the top quintile of the distribution of two-year cumulative effects are also in the top quintile of the one-year effect distribution. It is apparent that misspecification of the outcome variable produces extreme amounts of misclassification. Note, moreover, that this analysis assumes that the VAM1 and VAM2 exclusion restrictions are valid. A full account of the utility of VAMs for identifying good teachers would need to combine the analyses of lagged effects and endogenous classroom assignments. This would imply even higher rates of misclassification than are produced by either on its own.

8 Discussion

Panel data allows flexible controls for individual heterogeneity, but even panel data models can identify treatment effects only if assignment to treatment satisfies strong exclusion restrictions. This has long been recognized in the literature on program evaluation, but has received relatively little attention in the literature on the estimation of teachers' effects on student achievement. In this paper, I have shown how the availability of lagged outcome measures can be used to evaluate common value added specifications.

The results presented here show that the assumptions underlying common VAMs are substantially incorrect, at least in North Carolina. Classroom assignments are not exogenous conditional on the typical controls, and estimates of teachers' effects based on these models cannot be interpreted as causal. Clear evidence of this is that each VAM indicates that 5th grade teachers have quantitatively important "effects" on students' 4th grade learning. These results have important implications for educational research, for research in a variety of related areas, and for education policy. I discuss these in turn.

First, it is clear that an important priority in educational research should be to build richer VAMs that can accommodate dynamic sorting of students to classrooms. By contrast, there is little apparent need to allow for permanent heterogeneity in students' rates of growth. One approach to modeling classrom assignments might be to assume that classroom assignments depend on the principal's best prediction of students' unobserved ability, with predictions updated each year based on student grades and test scores. None of the VAMs considered here can accommodate assignments of this form, which on its face seems quite plausible, but approaches like those taken by Altonji et al. (2005) and Rothstein (forthcoming) may be useful.

I am skeptical, however, that purely econometric solutions will be adequate. There is likely to be important heterogeneity across schools in both information structures and principal objectives. Thus, there would be large returns to incorporating information about the actual school-level assignment process – perhaps gathered from surveys of principals, as in Monk (1987) – into the value added specification. In addition, more attention to the specification of the outcome variable is needed. Are we interested in measuring a teacher's short-run effect or her impact on test scores in later grades? The former is evidently a poor proxy for the latter.

Any proposed VAM should be subjected to thorough validation and falsification analyses. The tests implemented here suggest a starting point, and may be adaptable to richer models. Failure to reject the exclusion restrictions need not indicate that the restrictions are correct, as my tests can identify only sorting based on past observables. But rejection does indicate that the VAM-based estimates are likely to be misleading about teachers' causal effects.

The present analysis also has implications beyond the specific application to measuring teacher productivity. Estimates of the quality of schools and of the effects of firms on workers' wages use identical econometric models, and rely on similar exclusion restrictions. Evidence about the "effects" of future schools and employers on current outcomes would be informative about the validity of both sets of estimates.

Finally, the results here have important implications for the use of existing VAMs in education policy. My results indicate that policies based on these VAMs will reward or punish teachers who do not deserve it and fail to reward or punish teachers who do. The literature on pay-for-performance suggests some implications of this result. First, and most clearly, the stakes attached to VAM-based measures should be relatively small. Baker (1992, 2002) considers a performance measure that is less than perfectly correlated with the worker's contribution to firm output. He notes that high-stakes compensation will create incentives for workers to direct excess effort to the unproductive component of the performance measure. In education, this might take the form of teachers lobbying their principals to be assigned the "right" students who will yield predictably high value added scores. In Baker's model, misallocation of effort can be kept to a tolerable level by keeping the variable component of compensation small.[30] Another argument for low stakes in VAM-based compensation is provided by Hölmstrom and Milgrom (1991), who discuss implications of the results presented in Section 7 above: If short-term test scores are poor proxies for the dimensions of achievement that really matter, it may be better to forgo or limit incentive pay rather than encourage excessive teaching to the test.

A second and more speculative suggestion is that VAM-based estimates should be used as only one among several inputs into an accountability system that also incorporates principals' subjective ratings (see, e.g., Baker et al., 1994). There are two reasons for this. First, principals may have information about the direction of the bias in a particular teacher's VAM-based estimate that is not otherwise available to the econometrician, so incorporation of their opinions might lead to better-targeted incentives (Hölmstrom, 1979). Second, use of the VAM as the sole basis for teacher compensation and/or retention would permit principals to reward or punish teachers only through the assignment of desirable or undesirable students. Anecdotally, this is an important management tool for principals, who may induce disfavored teachers to resign by assigning them difficult students. But there is evidence that teacher-student matching is an important determinant of student learning (Clotfelter et al., 2006; Dee, 2005), so manipulation of

[30]See also Milgrom (1988), who argues that an important goal of organizational design should be to limit the incentive for workers to devote their time to "influence activities," and Lazear (1989), who argues that tournament stakes should be kept small to limit the incentive for "sabotage."

matches can have real efficiency consequences. If the principal's subjective judgment is incorporated directly into the incentive scheme, he or she will be able to allocate students to teachers to maximize output without sacrificing his or her ability to influence rewards and sanctions. Of course, this suggestion presumes high quality principals who have enough time to observe teachers' classrooms and enough training to distinguish good from bad teachers. Without this, neither subjective evaluations nor VAM-based estimates that depend importantly on classroom assignments are likely to provide much useful information.

References

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

Abowd, John M. and Francis Kramarz, "The Analysis of Labor Markets using Matched Employer-Employee Data," in Orley C. Ashenfelter and David Card, eds., *Handbook of Labor Economics*, Vol. 3B, Amsterdam: North-Holland, 1999, pp. 2629–2710.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, February 2005, *113* (1), 151–184.

Anderson, T.W. and Cheng Hsiao, "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, September 1981, *76* (375), 598–609.

Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc, "Here Today, Gone Tomorrow? Examining the Extent and Implications of Low Persistence in Child Learning," January 2009. Unpublished manuscript, Harvard.

Arellano, Manuel and Stephen Bond, "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *The Review of Economic Studies*, April 1991, *58* (2), 277–297.

Ashenfelter, Orley, "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, February 1978, *60* (1), 47–57.

_ **and David Card**, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *The Review of Economics and Statistics*, 1985, *67* (4), 648–660.

Baker, George, "Distortion and Risk in Optimal Incentive Contracts," *Journal of Human Resources*, Autumn 2002, *37* (4), 728–751.

Baker, George P., "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, June 1992, *100* (3), 598–614.

_ , Robert Gibbons, and Kevin J. Murphy, "Subjective Performance Measures in Optimal Incentive Contracts," *The Quarterly Journal of Economics*, November 1994, *109* (4), 1125–1156.

Ballou, Dale, "Test Scaling and Value-Added Measurement," June 2008. Unpublished manuscript.

Boardman, Anthony E. and Richard J. Murnane, "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement," *Sociology of Education*, April 1979, *52* (2), 113–121.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools," Working Paper 10, Center for Analysis of Longitudinal Data in Education Research. September 2007.

Braun, Henry I., "Using Student Progress To Evaluate Teachers: A Primer on Value-Added Models," Technical Report, ETS Policy Information Center. September 2005.

_ , "Value-Added Modeling: What Does Due Diligence Require?," in Robert W. Lissitz, ed., *Value Added Models in Education: Theory and Applications*, Maple Grove, Minn.: JAM Press, 2005, pp. 19–39.

Card, David and Daniel Sullivan, "Measuring the Effect of Subsidized Training Programs on Movements In and Out of Employment," *Econometrica*, May 1988, *56* (3), 497–530.

Chamberlain, Gary, "Panel Data," in Z. Griliches and M.D. Intriligator, eds., *Handbook of Econometrics*, Vol. II, Amsterdam: Elsevier North-Holland, 1984, pp. 1248–1318.

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

Dee, Thomas S., "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?," *American Economic Review*, May 2005, *95* (2), 158–165.

Goldhaber, Dan, "Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?," Working Paper 9, Center for Analysis of Longitudinal Data in Education Research. April 2007.

Harris, Douglas N. and Tim R. Sass, "Value-Added Models and the Measurement of Teacher Quality," April 2006. Unpublished manuscript.

_ and _ , "What Makes for a Good Teacher and Who Can Tell?," July 2007. Unpublished manuscript.

Heckman, James J., V. Joseph Hotz, and Marcelo Dabos, "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?," *Evaluation Review*, 1987, *11* (4), 395–427.

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, December 1986, *81* (396), 945–960.

Hölmstrom, Bengt, "Moral Hazard and Observability," *The Bell Journal of Economics*, Spring 1979, *10* (1), 74–91.

_ and Paul Milgrom, "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, 1991, pp. 24–52.

Imbens, Guido W. and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, February 2008, *142* (2), 615–635.

Jacob, Brian A. and Lars Lefgren, "What Do Parents Value in Education? An Empirical Examination of Parents' Revealed Preferences for Teachers," *Quarterly Journal of Economics*, November 2007, *122* (4), 1603–1637.

_ and _ , "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, January 2008, *25* (1), 101–136.

_ , _ , and David Sims, "The Persistence of Teacher-Induced Learning Gains," Working Paper 14065, National Bureau of Economic Research. June 2008.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan, "Earnings Losses of Displaced Workers," *The American Economic Review*, September 1993, *83* (4), 685–709.

Kane, Thomas J. and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," July 21, 2008. Unpublished manuscript.

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

Kezdi, Gabor, "Robust Standard Error Estimation in Fixed-Effects Panel Models," *Hungarian Statistical Review*, 2004, (Special Issue No. 9), 95–116.

Kinsler, Josh, "Estimating Teacher Value-Added in a Cumulative Production Function," October 2008. Unpublished manuscript, University of Rochester.

Koedel, Cory and Julian R. Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function," Working paper 07-08, University of Missouri Department of Economics. April 2007.

Konstantopoulos, Spyros, "How Long Do Teacher Effects Persist?," Discussion Paper No. 2893, IZA. June 2007.

Lazear, Edward P., “Pay Equality and Industrial Politics,” *Journal of Political Economy*, June 1989, *97* (3), 561–580.

Lockwood, J.R., Thomas A. Louis, and Daniel F. McCaffrey, “Uncertainty in rank estimation: Implications for value-added modeling accountability systems,” *Journal of Educational and Behavioral Statistics*, 2002, *27* (3), 255.

Martineau, Joseph A., “Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability,” *Journal of Educational and Behavioral Statistics*, Spring 2006, *31* (1), 35–62.

McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton, “Evaluating Value-Added Models for Teacher Accountability,” Report, RAND. 2003.

_ , Tim R. Sass, and J.R. Lockwood, “The Intertemporal Stability of Teacher Effect Estimates,” June 2008. Unpublished manuscript.

Milgrom, Paul R., “Employment Contracts, Influence Activities, and Efficient Organization Design,” *The Journal of Political Economy*, February 1988, *96* (1), 42–60.

Monk, David H., “Assigning Elementary Pupils to Their Teachers,” *Elementary School Journal*, November 1987, *88* (2), 167–187.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, “How Large Are Teacher Effects?,” *Educational Evaluation and Policy Analysis*, Fall 2004, *26* (3), 237–257.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, “Teachers, Schools, and Academic Achievement,” *Econometrica*, March 2005, *73* (2), 417–458.

Rosenbaum, Paul R. and Donald B. Rubin, “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, September 1984, *79* (387), 516–524.

Rothstein, Jesse, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” Working Paper 25, Princeton University Education Research Section, 2008.

_ , “Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables,” *Education Finance and Policy*, forthcoming.

Sanders, William L. and June C. Rivers, “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement,” Research Progress Report, University of Tennessee Value-Added Research and Assessment Center, November 1996.

_ , Arnold M. Saxton, and Sandra P. Horn, “The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment,” in Jason Millman, ed., *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?*, Thousand Oaks, CA: Corwin, 1997, pp. 137–162.

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

Wainer, Howard, "Introduction to a Special Issue of the Journal of Educational and Behavioral Statistics on Value-Added Assessment," *Journal of Educational and Behavioral Statistics*, Spring 2004, *29* (1), 1–3.

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2002.

Table 1. Summary statistics

	Population		Base sample		Most restricted sample	
	Mean	SD	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)
# of students	99,071		60,740		23,415	
# of schools	1,269		868		598	
1 5th grade teacher	122		0		0	
2 5th grade teacher	168		207		122	
3-5 5th grade teachers	776		602		440	
>5 5th grade teacher	203		59		36	
# of 5th grade classrooms	4,876		3,040		2,116	
# of 5th grade classrooms w/ valid teacher mat	3,315		3,040		2,116	
Female	49%		50%		51%	
Black	29%		28%		23%	
Other non-white	8%		7%		6%	
Consistent student record	99%		100%		100%	
Complete test score record, G4-5	88%		99%		100%	
G3-5	81%		91%		100%	
G2-5	72%		80%		100%	
Changed schools between G3 and G5	30%		27%		0%	
Valid teacher assignment in grade 3	68%		78%		100%	
grade 4	70%		86%		100%	
grade 5	72%		100%		100%	
Fr. of students in G5 class in same G4 class	0.22	[0.19]	0.22	[0.17]	0.30	[0.19]
Fr. of students in G5 class in same G3 class	0.15	[0.15]	0.15	[0.13]	0.28	[0.18]
Math scores 3rd grade (beginning of year)	0.11	[0.97]	0.14	[0.96]	0.20	[0.96]
3rd grade (end of year)	0.09	[0.94]	0.11	[0.94]	0.19	[0.91]
4th grade (end of year)	0.04	[0.97]	0.07	[0.97]	0.20	[0.93]
5th grade (end of year)	0.00	[1.00]	0.09	[0.98]	0.20	[0.94]
3rd grade gain	-0.02	[0.70]	-0.02	[0.69]	0.00	[0.69]
4th grade gain	-0.02	[0.58]	-0.01	[0.58]	0.01	[0.56]
5th grade gain	-0.01	[0.55]	0.01	[0.55]	-0.01	[0.53]
Reading scores 3rd grade (beginning of year)	0.08	[0.98]	0.12	[0.98]	0.17	[0.98]
3rd grade (end of year)	0.08	[0.95]	0.11	[0.94]	0.19	[0.91]
4th grade (end of year)	0.04	[0.98]	0.07	[0.97]	0.18	[0.93]
5th grade (end of year)	0.00	[1.00]	0.07	[0.97]	0.17	[0.94]
3rd grade gain	0.01	[0.76]	0.00	[0.75]	0.01	[0.75]
4th grade gain	-0.02	[0.59]	-0.02	[0.59]	0.00	[0.57]
5th grade gain	-0.01	[0.59]	0.00	[0.58]	-0.02	[0.57]

Notes: Summary statistics are computed over all available observations. Test scores are standardized using all 3rd graders in 1999, 4th graders in 2000, and 5th graders in 2001, respectively, regardless of grade progress. "Population" in Columns 1-2 is students enrolled in 5th grade in 2001, merged to 3rd and 4th grade records (if present) for the same students in 1999 and 2000, respectively. Columns 3-4 describe the base sample discussed in the text; it excludes students with missing 4th and 5th grade test scores, students without valid 5th grade teacher matches, 5th grade classes with fewer than 12 sample students, and schools with only one 5th grade class. Columns 5-6 further restrict the sample to students with non-missing scores in grades 3-5 (plus the 3rd grade beginning-of-year tests) and valid teacher assignments in each grade, at schools with multiple classes in each school in each grade and without perfect collinearity of classroom assignments in different grades.

Table 2. Correlations of test scores and score gains across grades

	Summary statistics		Correlations				N
			5th grade score		5th grade gain		
	Mean	SD	Math	Reading	Math	Reading	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math scores							
G5	0.02	1.00	1	0.78	0.29	0.08	70,740
G4	0.07	0.97	0.84	0.73	-0.27	-0.07	61,535
G3	0.09	0.95	0.80	0.70	-0.02	-0.03	57,382
G3 pretest	0.08	0.97	0.71	0.64	*0.00*	-0.03	50,661
Reading scores							
G5	0.01	1.00	0.78	1	0.10	0.31	70,078
G4	0.06	0.97	0.73	0.82	-0.05	-0.29	61,535
G3	0.09	0.95	0.70	0.78	-0.01	-0.05	57,344
G3 pretest	0.08	0.99	0.59	0.65	*0.00*	-0.05	50,629
Math gains							
G4-G5	0.01	0.55	0.29	0.10	1	0.25	61,349
G3-G4	-0.01	0.58	0.11	0.07	-0.41	-0.07	56,171
G2-G3	0.02	0.70	0.08	0.05	-0.02	0.01	50,615
Reading gains							
G4-G5	0.00	0.58	0.08	0.31	0.25	1	60,987
G3-G4	-0.02	0.59	0.08	0.10	-0.08	-0.41	56,159
G2-G3	0.02	0.75	0.09	0.10	-0.01	0.02	50,558

Notes: Each statistic is calculated using the maximal possible sample of valid student records with observations on all necessary scores and normal grade progress between the relevant grades. Column 7 lists the sample size for each row variable; correlations use smaller samples for which the column variable is also available. Italicized correlations are not different from zero at the 5% level.

Table 3. Evaluation of VAM1: Regression of gain scores on teacher indicators

	5th grade gain		4th grade gain		5th grade gain		4th grade gain	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher coefficients								
5th grade teachers								
Unadjusted SD	0.179	0.160	0.134	0.142	0.197	0.181	0.151	0.168
Adjusted SD	0.149	0.113	0.077	0.084	0.163	0.126	0.090	0.105
p-value	<0.001	<0.001	0.016	0.002	<0.001	<0.001	0.035	<0.001
4th grade teachers								
Unadjusted SD					0.188	0.181	0.220	0.193
Adjusted SD					0.150	0.125	0.182	0.140
p-value					<0.001	<0.001	<0.001	<0.001
Exclude invalid 4th grade teacher assignments & 5th grade movers?	n	n	n	n	y	y	y	y
# of students	55,142	55,142	55,142	55,142	40,661	40,661	40,661	40,661
# of 5th grade teachers	3,038	3,038	3,038	3,038	2,761	2,761	2,761	2,761
# of schools	868	868	868	868	783	783	783	783
R2	0.195	0.100	0.132	0.086	0.297	0.176	0.254	0.174
Adjusted R2	0.148	0.047	0.081	0.033	0.203	0.066	0.154	0.064

Notes: Dependent variables are as indicated at the top of each column. Regressions include school indicators, 5th grade teacher indicators, and (in columns 5-8) 4th grade teacher indicators, with one teacher per school per grade excluded. P-values are for test of hypothesis that all teacher coefficients equal zero, using the heteroskedasticity-robust score test proposed by Wooldridge (2002, p. 60). Standard deviations are of teacher coefficients, normalized to have mean zero at each school and weighted by the number of students taught. Adjusted standard deviations are computed as described in Appendix B2. Sample for Columns 1-4 includes students from the base sample (see text) with non-missing scores in each subject in grades 3-5. Columns 5-8 exclude students without valid 4th grade teacher matches and those who switched schools between 4th and 5th grade.

Table 4. Evaluation of VAM2: Regression of score levels on teacher indicators and lagged levels

	5th grade score		3rd grade score		5th grade score		3rd grade score	
	Math	**Reading**	**Math**	**Reading**	**Math**	**Reading**	**Math**	**Reading**
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Teacher coefficients								
5th grade teachers								
Unadjusted SD	0.176	0.150	0.120	0.129	0.191	0.169	0.138	0.150
Adjusted SD	0.150	0.109	0.067	0.076	0.161	0.121	0.079	0.091
p-value	<0.001	<0.001	0.040	0.007	<0.001	<0.001	0.162	0.001
4th grade teachers								
Unadjusted SD					0.160	0.162	0.182	0.175
Adjusted SD					0.121	0.109	0.142	0.126
p-value					<0.001	<0.001	<0.001	<0.001
Continuous controls								
4th grade math score	0.683	0.239	0.632	0.213	0.708	0.255	0.668	0.229
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
4th grade reading score	0.195	0.617	0.218	0.620	0.189	0.613	0.206	0.621
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Exclude invalid 4th grade teacher assignments & 5th grade movers?	n	n	n	n	y	y	y	y
# of students	55,142	55,142	55,142	55,142	40,661	40,661	40,661	40,661
# of 5th grade teachers	3,038	3,038	3,038	3,038	2,761	2,761	2,761	2,761
# of schools	868	868	868	868	783	783	783	783
R2	0.313	0.249	0.274	0.237	0.385	0.315	0.354	0.307
Adjusted R2	0.273	0.206	0.231	0.193	0.302	0.224	0.268	0.215

Notes: Dependent variables are as indicated at the top of each column. Regressions include school indicators, 4th grade math and reading scores, 5th grade teacher indicators, and (in columns 5-8) 4th grade teacher indicators, with one teacher per school per grade excluded. P-values are for test of hypothesis that all teacher coefficients equal zero, using the heteroskedasticity-robust score test proposed by Wooldridge (2002, p. 60). Standard deviations are of teacher coefficients, normalized to have mean zero at each school and weighted by the number of students taught. Adjusted standard deviations are computed as described in Appendix B2. Samples correspond to those in Table 3.

Table 5. Correlated random effects evaluation of VAM3: Gain score specification with student fixed effects

	Math			Reading		
	3rd grade	4th grade	Corr((1),(2))	3rd grade	4th grade	Corr((4),(5))
	(1)	(2)	(3)	(4)	(5)	(6)
Unrestricted model						
Standard deviation of teacher effects, adjusted						
5th grade teacher	0.135	0.099	-0.04	0.144	0.123	-0.06
4th grade teacher	0.136	0.193	-0.07	0.160	0.163	-0.08
3rd grade teacher	0.228	0.166	-0.36	0.183	0.145	-0.24
Fit statistics						
R2	0.314	0.376		0.245	0.284	
Adjusted R2	0.129	0.209		0.042	0.092	
Restricted model (Optimal Minimum Distance)						
Ratio, effect on G4 / effect on G3	0.14			1.17		
SD of G5 teacher effects	0.126	0.018		0.088	0.103	
Objective function	2,136			2,174		
95% critical value	1,684			1,684		
p value	<0.001			<0.001		

Notes: N=25,974. Students who switched schools between 3rd and 5th grade, who are missing test scores in 3rd or 4th grade (or on the 3rd grade beginning-of-year tests), or who lack valid teacher assignments in any grade 3-5 are excluded. Schools with only one included teacher per grade or where teacher indicators are collinear across grades are also excluded. "Unrestricted model" reports estimates from a specification with school indicators and indicators for classrooms in grades 3, 4, and 5. Restricted model reports optimal minimum distance estimates obtained from the coefficients from the unrestricted models for 3rd and 4th grade gains, excluding the largest class in each grade in each school. Restriction is that the 4th grade effects are a scalar multiple of the 3rd grade effects. Weighting matrix is the inverse of the robust sampling variance-covariance matrix for the unrestricted estimates, allowing for cross-grade covariances.

Table 6. Magnitude of bias in VAM1 and VAM2 relative to a richer specification that controls for all past observables

	VAM1		VAM2	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Standard deviation of 5th grade teachers' estimated effects from traditional VAM				
Unadjusted for sampling error	0.203	0.189	0.197	0.176
Adjusted for sampling error	0.162	0.127	0.162	0.121
SD of 5th grade teachers' estimated effects from rich specification (VAM4)				
Unadjusted for sampling error	0.206	0.200	0.206	0.200
Adjusted for sampling error	0.172	0.148	0.172	0.148
SD of bias in traditional VAMs relative to the rich specification				
Unadjusted for sampling error	0.118	0.130	0.097	0.106
Adjusted for sampling error	0.060	0.054	0.037	0.028

Notes: N=23,415. Sample is that used in Table 5, less observations with missing 5th grade scores and those in schools rendered unusable (i.e. only one valid classroom or collinearity between 3rd, 4th, and 5th grade classroom indicators) by this exclusion. "Rich" specification controls for classroom assignments in grades 3 and 4 and for scores in math and reading in grades 2, 3, and 4. "Bias" is the difference between the VAM1/VAM2 estimates and those from the rich specification. Unadjusted estimates summarize the estimated coefficients. Adjustments for sampling error are described in Appendix B.

Table 7. Simulations of the effects of student selection and heterogeneous decay on teacher quality estimates

	Data generating process			Simulation: Comparisons between true effects and those indicated by VAM		
	SD of truth	SD of bias	(2) as % of (1)	Correlation	Rank correlation	Reliability of top quintile
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Selection is on observables						
VAM1	0.148	0.054	36%	0.93	0.93	0.79
VAM2	0.148	0.028	19%	0.98	0.98	0.90
VAM4	0.148	0	0%	1.00	1.00	1.00
Panel B: Selection is on history of two tests, one observed						
VAM1	0.148	0.124	84%	0.77	0.75	0.62
VAM2	0.148	0.049	33%	0.95	0.94	0.82
VAM4	0.148	0.028	19%	0.98	0.98	0.89
Panel C: Selection is on history of three tests, one observed						
VAM1	0.148	0.137	92%	0.74	0.73	0.60
VAM2	0.148	0.060	40%	0.93	0.92	0.78
VAM4	0.148	0.041	28%	0.96	0.96	0.85
Panel D: Selection is on true and observed achievement history						
VAM1	0.148	0.166	112%	0.64	0.63	0.52
VAM2	0.148	0.089	60%	0.86	0.85	0.70
VAM4	0.148	0.078	53%	0.89	0.88	0.73
Panel E: Selection on unobservables is like selection on observables						
VAM1	0.148	0.212	143%	0.57	0.56	0.49
VAM2	0.148	0.140	95%	0.73	0.71	0.59
VAM4	0.148	0.147	99%	0.71	0.70	0.58
Panel F: Selection conforms to VAM assumptions, but effects of interest are those on the following year's score						
VAM1	0.118	0.148	125%	0.42	0.40	0.38
VAM2	0.110	0.147	133%	0.33	0.32	0.34

Notes: Estimates in Column 1 are taken from the rich specification for reading in Table 6 (Panels A-E) and from Columns 2 and 4 of Table 8 (Panel F). Column 2 is from Table 6, Columns 2 and 4 in Panel A and is computed from the models reported in Table 8 in Panel F. In Panels B-E, estimates from Table 10 of Rothstein (2008) are used, with an adjustment for the different test scale used here. See Appendix for details. Columns 4-6 are computed by drawing 10,000 teachers from normal distributions with the standard deviations described in Columns 1-2. Estimates of the correlation between teachers' true effects and the bias in their estimated effects (-0.33 for VAM 1 and -0.43 for VAM2) are used in Panel A. In Panels B-E, this correlation is constrained to zero. In Panel F, the estimated correlation is used again; this is -0.38 for VAM1 and -0.43 for VAM2. "Reliability of top quintile" in Column 6 is the fraction of teachers whose true effects are in the top quintile who are estimated to be in the top quintile by the indicated VAM.

Table 8. Persistence of teacher effects in VAMs with lagged teachers

	VAM1		VAM2	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Cumulative effect of 4th grade teachers over two years				
Standard deviation of 4th grade teacher effects, adjusted				
on 4th grade scores	0.184	0.150	0.188	0.140
on 5th grade scores	0.108	0.118	0.118	0.110
Correlation(effect on 4th grade, effect on 5th grade), adjusted	0.455	0.413	0.511	0.334
Cumulative effect of 3rd grade teachers over three years				
Standard deviation of 3rd grade teacher effects, adjusted				
on 3rd grade scores	0.218	0.172	0.209	0.167
on 4th grade scores	0.136	0.126	0.120	0.130
on 5th grade scores	0.185	0.199	0.129	0.147
Correlation(effect on 3rd grade, effect on 5th grade), adjusted	0.395	0.341	0.450	0.447

Notes: N=23,415. Sample is identical to that used in Table 6. Effects of 4th grade teachers on 5th grade scores and of 3rd grade teachers on 4th and 5th grade scores are cumulative effects. For VAM1, the specification for gains in grade g includes controls for teachers in grades 3 through g, and the cumulative effect of the grade h teacher on the grade g gain is the sum of the effects in h, h+1, ..., g. For VAM2, the specification is augmented with controls for math and reading scores in grade g-1. The calculation of cumulative effects is described in footnote 29. All standard deviations and correlations are adjusted for the influence of sampling error, as described in Appendix B.