NBER WORKING PAPER SERIES

TEACHER QUALITY IN EDUCATIONAL PRODUCTION: TRACKING, DECAY, AND STUDENT ACHIEVEMENT

Jesse Rothstein

Working Paper 14442
http://www.nber.org/papers/w14442

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2008

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 Honor, 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). The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement
Jesse Rothstein
NBER Working Paper No. 14442
October 2008
JEL No. I21,J24,J33

ABSTRACT

Growing concerns over the 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.

Jesse Rothstein
Industrial Relations Section
Firestone Library
Princeton University
Princeton, NJ 08544
and NBER
jrothst@princeton.edu

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 impose an assumption 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 and firms cannot have causal effects on *past* outcomes, while violations of model assumptions may lead to apparent counterfactual "effects" of this form. Both test scores and wages are serially correlated, and as a result an association between the current teacher or firm and the lagged outcome is

[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).

strong evidence against exogeneity with respect to the current outcome.

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 saturated model that conditions on the complete achievement history. Estimated teacher effects from the saturated 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.

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 are likely to do a 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 about the assignment of students to teachers and the matching of workers to firms cannot support their intended uses. Policies that use VAM-based 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 prior work that uses 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. Section 8 concludes.

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. 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

analogous tests are used in regression discontinuity analyses (Imbens and Lemieux, 2008).

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 studies of teacher productivity, the multiplicity of teacher "treatments" can blur the connection between value added modeling and 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.[2]

[2]Only a few studies have attempted to validate VAMs. Jacob and Lefgren (2008) and Harris and Sass (2007) show that value added estimates are correlated with principals' ratings of teacher performance. And Kane and Staiger (2008) demonstrate that VAMs estimated on observational data predict teachers' experimental effects. These studies are extremely valuable, but they cannot rule out quantitatively important biases coming from non-random classroom assignments. The estimated correlations between principal ratings and value added are relatively weak, and the Kane and Staiger experimental sample is too small (and potentially non-representative) to rule out

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.

I am interested in whether common VAMs identify classroom effects with arbitrarily large samples. I therefore sidestep small sample issues. 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. I instead consider the properties of VAM estimates as the number of students grows with the number of teachers (and classrooms) fixed. 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;[3] 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. As it is impractical to report each of several thousand teachers' estimated effects, I report only summaries of their distribution across teachers. I select statistics, like the estimated standard deviation of 5th grade teachers' effects on students' 4th grade achievement, that are informative about

any but the most extreme alternatives regarding classroom assignments.

[3] 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.

whether the VAMs can identify individual teacher effects without bias.[4]

3.2 Data Generating Process and the Three VAMs

Following Todd and Wolpin (2003) and Harris and Sass (2006), student achievement can be modeled as a linear, additive function of the full history of inputs received to date plus the student's innate ability. Separating classroom effects from other inputs, we can write the test score of student i at the end of grade g, A_{ig}, as

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

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 ability. We might expect the achievement gap between high-ability and low-ability students to grow over time; this would correspond to $\tau_{g+s} > \tau_g$ for each g and each $s > 0$. 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, ν_{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.[5]

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 $h \leq g < g'$. 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

[4]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.

[5]I define the β parameters to include any classroom-level component of ν_{ig} and assume that ν_{ig} is independent across students in the same classroom. The Appendix discusses the use of repeated observations on teachers to distinguish correlated errors from teachers' true causal effects.

is no theoretical basis for restrictions on the time pattern 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 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: $\Delta A_{ig} = \alpha_g + \beta_{ggc(i,g)} + e_{1ig}$;

an augmented regression that controls for the prior year's score,

VAM2: $\Delta A_{ig} = \alpha_g + A_{ig-1}\psi + \beta_{ggc(i,g)} + e_{2ig}$;[6]

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

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

All three VAMs are widely used.[7] VAM2 and VAM3 can both be seen as generalizations of VAM1: VAM2 is equivalent to VAM1 when $\psi = 0$, while VAM3 reduces to VAM1 when $\mu_i \equiv 0$.

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.

[6] VAM2 is more commonly specified as an equivalent model for the end-of-year score, $A_{ig} = \alpha_g + A_{ig-1}(\psi + 1) + \beta_{ggc(i,g)} + e_{2ig}$. Relative to this, the expression in the text merely subtracts A_{ig-1} from each side. I focus on the gain score version to maintain the parallel with VAM1 and VAM3.

[7] 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).

3.3 The gain score model (VAM1)

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

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

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:

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

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:

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

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.[8] 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

[8]This strategy has zero power unless there is independent variation in $c(i,g-1)$ and $c(i,g)$. If students are "streamed," moving together with the same classmates from grade to grade, controls for $c(i,g-1)$ will absorb all across-$c(i,g)$ variation. In the Tennessee STAR experiment (see Nye et al., 2004), streaming was quite common, and in many schools there is zero independent variation in 3rd grade classroom assignments controlling for 2nd grade assignments. This makes it impossible to distinguish the effects of 2nd and 3rd grade teachers, and prevents the use of my test. In the observational data examined below, students are substantially reshuffled between grades.

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 process of order $g-1$ (absent strong restrictions on the ϕ coefficients). Finally, Δv_{ig} is an MA(1), degenerate only if $\text{var}(v) = 0$.[9]

The discussion of serial correlation in e_{1ig} helps clarify the conditions in which (4) will likely 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}.

3.4 The lagged score model (VAM2)

VAM2 augments VAM1 with a control for 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:

$$A_{ig} = \left(\alpha_g - \alpha_{g-1}\lambda\right) + A_{ig-1}\lambda + \beta_{ggc(i,g)} + \mu_i\left(\tau_g - \tau_{g-1}\lambda\right) + \omega_{ig} - \omega_{ig-1}\lambda + v_{ig} - v_{ig-1}\lambda, \quad (6)$$

and the grade-g gain is thus

$$\Delta A_{ig} = \breve{\alpha}_g + A_{ig-1}\psi + \beta_{ggc(i,g)} + e_{2ig} \quad (7)$$

[9] Rothstein (2008b) concludes that Δv_{ig} accounts for as much as 80% of the variance of ΔA_{ig}.

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

$$e_{2ig} = \mu_i \left(\tau_g - \tau_{g-1}\lambda\right) + \sum_{h=1}^{g-1} \varepsilon_{ih} \left(\phi_{hg} - \phi_{hg-1}\lambda\right) + \varepsilon_{ig} + \left(\nu_{ig} - \nu_{ig-1}\lambda\right). \tag{8}$$

As before, each of the terms in (8) is likely to be serially correlated. The VAM2 exclusion restriction, $E\left[e_{2ig} \,|\, c\left(i,g\right) = c\right] = 0$, 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, $c\left(i,g\right)$ 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

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

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

$$\Delta A_{ig-1} = \frac{1}{1+\psi}\left[\breve{\alpha}_{ig-1} - A_{ig-1}\psi + \beta_{g-1,g-1,c(i,g-1)} + e_{2ig-1}\right]. \tag{10}$$

Thus, the grade-g classroom assignment will have predictive power for the gain score in grade $g-1$, controlling for $g-1$ achievement, if grade-g classrooms are correlated either with grade-$g-1$ teacher effects (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.

[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. IV and GMM estimates of the first-difference of (6), treating ΔA_{ig-1} as an endogenous variable, can identify λ and β_{gg} if $c\left(i,g\right)$ depends on μ_i but is strictly exogenous conditional on this (Anderson and Hsiao, 1981; Arellano and Bond, 1991). Koedel and Betts (2007) is the only teacher value added study of which I am aware that takes account of the issues raised by lagged dependent variables. Value added researchers typically apply OLS to (7). This is inconsistent for ψ, and identifies β_{ggc} only if $c\left(i,g\right)$ is random conditional on A_{ig-1}.

3.5 The fixed effects in gains model (VAM3)

The final VAM returns to the earlier assumption of zero decay of teachers' effects.[11] It incorporates the ability term in (2) into the estimating equation,

$$\Delta A_{ig} = \Delta\alpha_g + \beta_{ggc(i,g)} + \mu_i \Delta\tau_g + e_{3ig}, \tag{11}$$

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

The presence of the student fixed effect in VAM3, combined with the small time dimension of student data sets, means that VAM3 requires stronger assumptions than the earlier models. Assuming that $\Delta\tau_g = 1$ for each g, (11) is a fixed effects model. 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:

$$\Delta A_{ig} - \frac{1}{G}\sum_{h=1}^{G}\Delta A_{ih} = \left(\Delta\alpha_g - \frac{1}{G}\sum_{h=1}^{G}\Delta\alpha_h\right) + \left(\beta_{ggc(i,g)} - \frac{1}{G}\sum_{h=1}^{G}\beta_{hhc(i,h)}\right) + \left(e_{3ig} - \frac{1}{G}\sum_{h=1}^{G}e_{3ih}\right). \tag{12}$$

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 (12). 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. [12]

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

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

[12] As G gets large, $\frac{1}{G}$ shrinks toward zero, and e_{3ih} disappears from the equation for the de-meaned grade-g gain, $g \neq h$. For practical value added implementations, however, G is rarely larger than three or four. Without strict exogeneity, one small-G approach is to focus on the first difference of (11). 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 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.

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. Under strict exogeneity, any apparent effect of (for example) 5th grade teachers on 4th grade gains in VAM1 appears only because both 5th grade teacher assignments and 4th grade gains depend on μ. 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 thus imply 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:

$$\mu_i = \xi_{1c(i,1)} + \ldots + \xi_{Gc(i,G)} + \eta_i. \tag{13}$$

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 (13) into (11), we obtain

$$\Delta A_{ig} = \Delta\alpha_g + \sum_{h=1}^{G} \pi_{hgc(i,h)} + \eta_i + e_{3ig}, \tag{14}$$

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, $E\left[e_{3ih} \,|\, c\left(i,1\right), \ldots, c\left(i,G\right)\right] = 0$ for each h, and the fact that (13) 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

$$\pi_{31} = \xi_3 \Delta\tau_1 = \xi_3 \Delta\tau_2 \frac{\Delta\tau_1}{\Delta\tau_2} = \pi_{32} \frac{\Delta\tau_1}{\Delta\tau_2}. \tag{15}$$

$\Delta\tau_1$ and $\Delta\tau_2$ are scalars, so (15) represents $J_3 - 1$ overidentifying restrictions on the $2J_3$ elements of the π_{31} and π_{32} vectors.[13]

Equation (15) 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

$$D = \left(\begin{pmatrix} \hat{\pi}_{31} \\ \hat{\pi}_{32} \end{pmatrix} - \begin{pmatrix} \xi_3 \Delta\tau_1 \\ \xi_3 \Delta\tau_2 \end{pmatrix} \right)' W^{-1} \left(\begin{pmatrix} \hat{\pi}_{31} \\ \hat{\pi}_{32} \end{pmatrix} - \begin{pmatrix} \xi_3 \Delta\tau_1 \\ \xi_3 \Delta\tau_2 \end{pmatrix} \right) \tag{16}$$

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.[14] If D is above the 95% critical value from this distribution, the null is rejected. In practice, implementations of VAM3 treat μ_i as a fixed effect, thus imposing the additional restriction that $\Delta\tau_2 = \Delta\tau_1$. Under the null that this model is correct, the restricted D has J_3 degrees of freedom.

3.6 Implementation and Computation

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 comparisons. But students are not randomly assigned to schools, and those at one school may gain

[13] There are J_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.

[14] 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 $\Delta\tau_1 = 1$.

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. The indicators for all of the classrooms at a school are collinear with the indicator for the school, and I normalize the classroom coefficients to have mean zero across classrooms in the same grade at the same school.[15]

The tables below report summary statistics for the teacher coefficients rather than the full coefficient vectors themselves. Due to sampling error, summary statistics computed from the estimated coefficients differ from those that would be obtained were the true coefficients known. Aaronson et al. (2007) propose a simple estimator for the variance of the true coefficients across teachers. Let γ be a mean-zero J-vector of true projection coefficients – those that would be obtained with an infinitely large sample – 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 the difference between the variance of the estimated coefficients and the sampling variance:

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

I compute $E\left[\hat{\gamma}'\hat{\gamma}\right]$ using a degrees-of-freedom adjustment for the school-level normalization of the estimated coefficients; $E\left[\left(\hat{\gamma}-\gamma\right)'\left(\hat{\gamma}-\gamma\right)\right]$ is merely the average sampling variance of the normalized coefficients. All calculations are weighted by the number of students taught.

Some of the specifications discussed above – particularly (14) – include indicators for classroom assignments in several grades simultaneously. This introduces several 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

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

do not switch schools, and include only a single set of school indicators.

Second, specifications with several sets of classroom indicators have design matrices of large dimension. Numerical inverses may be unstable. My focus on samples of non-movers eliminates this problem when the specification includes only school and teacher indicators, as it ensures that indicators for teachers at different schools are uncorrelated and that the design matrix is block diagonal. I treat these specifications as separate regressions for each school, each with only a few dozen regressors. Specifications that include continuous covariates (e.g., VAM2) cannot be decomposed in this way. For these, I begin with brute-force estimates, then verify the estimated coefficients using an iterative algorithm (described in the Appendix) that does not require inversion of large matrices.

A final complication is that 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 public school students in North Carolina. The data, assembled and distributed by the North Carolina Education Research Data Center, have been extensively cleaned to ensure accurate matches between the component administrative data systems, and 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.[16]

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 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 (14)). 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

[16]The test 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 (2008b) 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.

base and restricted samples have higher mean 5th grade scores than the full population. This primarily reflects the lower scores of students who switch schools frequently.[17] 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 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 less than full rank (after dropping one teacher per grade).

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 (i.e. that $\tau_{g+1} > \tau_g$ for each g), 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 (2008a,b) concludes that noise cannot account for the magnitude of the observed negative year-to-year correlation. This strongly suggests that VAM3 is poorly suited to the test score data generating process.

[17] 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.

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 classrooms, then normalize the resulting coefficients to have mean zero within each school. In each case, the hypothesis that all of the teacher 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 gains, 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 4th grade gains 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 4th grade math gain specification, though they remain quite large in magnitude. They are still highly significant in the specification for 4th grade reading gains.

The VAM2 residuals, like the VAM1 residuals, are strongly 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 earlier reading gains even after controlling for 4th grade scores

in both subjects indicates that the VAM2 exclusion restriction is violated, regardless of whether the dependent variable is the math or the reading gain. As before, 5th grade teachers' effects on 5th grade gains 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.

As discussed in Section 3.5, the falsification test for VAM3 takes a different form. 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 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.[18]

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, $\pi_{54} = (\Delta\tau_4/\Delta\tau_3)\,\pi_{53}$, 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, they are direct evidence against the VAM3 identifying assumption of conditional strict exogeneity.

[18] 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.

The lower panel of Table 5 presents OMD estimates of the restricted model.[19] I consider two versions, one that constrains $\Delta\tau_4/\Delta\tau_3 = 1$ (as would be needed in order to estimate VAM3 using conventional fixed effects methods) and another that does not. Neither model is able to fit the data. For math scores, the estimated ratio $\Delta\tau_4/\Delta\tau_3$ from the less restrictive model 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. 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, and free lunch status; this has no effect on the tests. Similarly, I have explored a variety of alternative test scalings. The three VAMs continue to fail falsification tests when I use the original score scales or percentiles in place of the standardized-by-grade scores used in Tables 3, 4, and 5.

The results are also not specific to the cohort examined here; I obtain similar results using data from other cohorts. As a final investigation, I have extended the tests to evaluate VAM

[19]The OMD analysis uses a variance-covariance matrix W that is robust to arbitrary heteroskedasiticity and within-student, between-grade clustering.

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 that are observable 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 saturated specification 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 specification 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 the saturated model. Of course, if assignments depend on both ability and lagged scores, VAM1, VAM2, and the saturated VAM are all misspecified.

Table 6 presents comparisons of the saturated VAM with VAM1 and VAM2. 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. The sampling-adjusted estimates are quite similar to those from the larger sample. The next two rows of the Table show estimates from the saturated specification. Standard deviations are somewhat larger, but not dramatically so.

The final two rows describe the bias in the simpler VAMs relative to the saturated model (that is, $\beta_{55}^{VAM1} - \beta_{55}^{saturated}$ and $\beta_{55}^{VAM2} - \beta_{55}^{saturated}$). 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 the standard deviation of the estimated effects. For VAM2, which already includes a subset of the controls in the saturated model, 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, 2008a), I attempt to quantify the bias that is likely to result from selection on unobservables. Classroom assignments likely also depend on characteristics – behavior, personality, parental intervention, etc. – that may be observed by the principal but are unobserved by the econometrician. These characteristics may be predictive of future outcomes. Following the intuition (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 plausibly 50-75% as large (in standard deviation terms) as the estimates of teachers' effects in VAM1, and perhaps half

this large in VAM2.[20] These estimates imply that VAM2 and especially VAM1 seriously misidentify teachers' true causal effects, crediting teachers for the students they are assigned. One cannot be confident that a teacher identified as good by these models is in fact a good teacher, rather than simply a teacher who was given students predicted (by principals, if not by the econometrician) to gain quickly.

7 Short-Run vs. Long-Run Effects

Although classroom assignments are the focus of this paper, it is worth returning to another implication of the results in Section 5. 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. One can certainly imagine that some teaching styles (e.g., "teaching to the test") would produce large short-run effects that decay quickly while other styles (emphasizing independent exploration) might yield smaller short-run effects that persist and even grow in later years.[21] As this example shows, it is far from clear that accountability policy should focus exclusively on short-run effects rather than long-run effects if the two in fact differ.

While several studies have attempted to estimate the decay parameter ψ,[22] this is the first

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

[21] 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, 2008b; Ballou, 2008; Martineau, 2006).

[22] Studies predating this one include Andrabi et al. (2008), Sanders and Rivers (1996), and Konstantopoulos (2007).

value added study of which I am aware that estimates teachers' immediate and lagged effects without imposing a restriction of uniform decay. 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 $g+2$.[23] Under the uniform decay restriction, these should be perfectly correlated (except for sampling error).

I begin by estimating VAM1 and VAM2 for 3rd, 4th, and 5th grade 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, the standard deviation of teachers' estimated "effects" falls in the year after contact – 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 $(1-\lambda)$, $\text{var}(\beta_{g,g+s}) = \lambda^s \text{var}(\beta_{gg})$, so this is consistent with the mounting evidence that teachers' effects decay importantly in the year after contact (Andrabi et al., 2008; 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 teacher's first year effect is a poor proxy for her longer-run impact.

As a final exercise, I bring together the analyses of endogeneity bias and decay to investigate whether estimates of short-run effects from VAM1 and VAM2 are reasonably accurate proxies from those that would be obtained from a superior model for longer-run effects. I estimate

[23] 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 $(\beta_{33c}(1+\psi_4)+\beta_{34c})(1+\psi_5)+\beta_{35c}$. A similar but more complex expression characterizes the effects when lagged scores in both math and reading are controlled, as in my estimates.

the saturated VAM from Section 6 for both 4th and 5th grade gains, controlling for all past observables, and compute the implied cumulative effect of 4th grade teachers on students' 5th grade outcomes. Figure 1 shows the scatterplot of VAM1 and VAM2 estimates of 4th grade teacher effects against those from the cumulative saturated specification. Both VAM1- and VAM2-based estimates of effects on math scores correlate just over 0.4 with those from the richer model, while correlations for reading achievement are below 0.35.

Many teacher accountability policies focus only on the very best and very worst teachers. Figure 1 shows the 20th and 80th percentiles of the distribution of estimated effects from each model. For each contrast, I compute the fraction of teachers in the top and bottom quintile according to the cumulative, saturated specification who are assigned to the same quintile by VAM1 or VAM2. These are similar to the correlations, around 0.43 for math and 0.35 for reading. Even ignoring the impact of sampling error, which would tend to exacerbate these results but is not accounted for here, it is clear that model misspecification produces extreme amounts of misclassification. Policies that use VAM1 or VAM2 to attempt to identify the best and worst teachers will both reward and punish teachers who do not deserve it and fail to reward and punish teachers who do.

8 Discussion

Access to panel data allows the econometrician to control for individual heterogeneity much more flexibly than can be accomplished in cross-sectional data, but even panel data models can identify treatment effects only if assignment to treatment satisfies strong ignorability assumptions. 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 sub-

stantially 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.

This result casts serious doubt on the value of simple VAMs for accountability and incentive policies, which will clearly be sensitive to the assignment of students to teachers. Teachers operating under high-stakes VAM-based accountability and incentive systems can be expected to lobby their principals to be assigned the "right" students who will predictably yield high value added scores, and principals will presumably alter their assignment rules to direct these students toward favored teachers. As teacher-student matching is a potentially important determinant of student learning (Clotfelter et al., 2006; Dee, 2005), distortion of these matches due to efforts to manipulate teachers' value added scores can have real efficiency consequences.

It is clear that richer VAMs are needed. These will need to accommodate dynamic classroom assignments and will probably require behavioral assumptions about the principal's objective function and information set. For example, one might assume that classroom assignments depend on the principal's best prediction of students' unobserved ability, and that this prediction is after receipt with each year's test results. None of the VAMs considered here can accommodate assignments of this form, which on its face seems more plausible than the identifying assumptions for VAM1, VAM2, or VAM3.

Attempts to infer causal effects even from rich, dynamic VAMs call for a great deal of caution and attention to the required assumptions. Any VAM proposed for policy use should be subjected both to thorough validation and to falsification exercises. 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.

Even with a valid model, it will also be important to measure teachers' effects on student

achievement over several years, not merely at the end of the year of exposure. Estimates of teacher quality are evidently quite sensitive to this aspect of the model. By contrast, there is little apparent need to allow for permanent heterogeneity in students' rates of growth, as the data provide no indication of such heterogeneity.

The questions investigated and methods used here have applications beyond the estimation of individual teacher quality. The Appendix shows that conclusions about the relationship between teachers' observed characteristics and their value added also rest on unsteady ground. 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.

A Data Appendix

This appendix describes the construction of the samples used in the paper. I begin with records on all students who were enrolled in 5th grade in North Carolina public schools in 2000-2001. From this universe, I exclude students with inconsistent longitudinal records (i.e. "male" in some years and "female" in others, amounting to less than 1% of the population); those who cannot be matched to 4th grade records from 1999-2000, perhaps because they skipped a grade or attended private school (10%); those who cannot be matched to a 5th grade teacher or for whom the 5th grade test administrator is not a valid teacher as defined in the text (24%); those whose 5th grade class has fewer than 12 included students (1%); and those whose elementary school contains only a single included 5th grade class (3%). This leaves me with a sample of 60,740, 61.3% of the initial population. I refer to this sample as the "base" sample.

Each of my analyses uses subsets of this sample that have complete data on test scores and teacher assignments for enough years to permit the analysis. A student might be excluded from the analytical subsample for a particular analysis because there is no record in one of the necessary grades; because there is a record but no test score; because the student changed schools between grades; because she could not be matched to a valid teacher in each of the required grades; because she was the only otherwise-usable student from her class in one or more grades; because there was only one included class at her school in one or more grades; or because the school did not shuffle students adequately between grades, leading to collinearity between the classroom assignments in one year and those in other years. Appendix Table A1 describes the samples used in Columns 1-4 of Tables 3 and 4 (requiring complete test histories from grades 3-5 and teacher assignments in grade 5); in Columns 5-8 of those Tables (also requiring valid teacher assignments in 4th grade); and in Table 5 (also requiring 3rd grade teacher assignments and scores from the beginning-of-third-grade tests).

Appendix Table A2 reports statistics on shuffling of classrooms between 4th and 5th grades. This uses a somewhat different sample than other tables, consisting of all students with valid records and valid teacher matches in both grades 4 and 5 who did not switch schools or make abnormal progress between grades. Using this sample, I count the number of 4th grade classes at the school, and I compute for each student the fraction of her 5th grade classmates who were also in her 4th grade class. I average this over the full sample and over subsamples defined by the number of 4th grade teachers at the school. I also identify schools where dummies for the J_4 4th grade teachers and J_5 5th grade teachers have rank less than $J_4 + J_5 - 2$, indicating perfect collinearity of at least one teacher assignment with the others, and re-compute the statistic excluding observations from those schools.

B Technical Appendix

This appendix provides more detail on some of the computations undertaken in the paper.

B.1 School-level normalizations

As discussed in the text, each of my regressions includes fixed effects for the school at-

tended, and coefficients on teacher indicators are normalized to have mean zero at the school level. This normalization is easiest to describe if the sample consists of only a single school. Let T be an N-by-J matrix of indicators for having been taught by each of the J teachers in a particular grade at that school. Many of my regressions take the form

$$y = \alpha + T\beta + \varepsilon. \tag{18}$$

Let $S = [1\,T]$ be the data matrix formed by augmenting the T matrix with a constant. Because each student has exactly one teacher, $S'S$ has rank J, so not all of the $J+1$ coefficients in α and β can be separately identified. Suppose, without loss of generality, that the last element of T is dropped. Let $\hat{b}$ be the estimates of the remaining elements of β, and let V_b be the estimated sampling variance-covariance matrix for $\hat{b}$. Form $\hat{\beta} = \left(\hat{b}'\,0\right)'$, and let V be the corresponding variance matrix,

$$V \equiv \begin{pmatrix} V_b & 0_J \\ 0_J' & 0 \end{pmatrix}, \tag{19}$$

where 0_J is a column vector of J zeros.

Let n be a J-vector with elements n_j, where n_j is the number of students taught by teacher j. Then the weighted average element of $\hat{\beta}$, weighting each teacher by the number of students taught, can be written as $\bar{\hat{\beta}} = (n'1_J)^{-1} n'\hat{\beta}$ (where 1_J is a J-vector of ones), and the vector $\hat{\beta} - \bar{\hat{\beta}} = \left(I_J - 1_J (n'1_J)^{-1} n'\right)\hat{\beta} \equiv D\hat{\beta}$ has weighted mean zero across teachers. The sampling variance matrix for the normalized coefficients $\hat{\beta} - \bar{\hat{\beta}}$ is simply DVD'. This has rank $J-1$.

The extension of this procedure to samples spanning many schools is straightforward. Suppose that the teacher indicators are ordered, so that the first J_1 come from school 1, the next J_2 from school 2, and so on. Let $\hat{\beta}$ be the full vector of estimated coefficients with the coefficient for the final teacher at each school set to zero (i.e the J_1, $(J_1 + J_2)$, etc., elements of $\hat{\beta}$), and let V be the sampling variance matrix (with rows and columns of zeros corresponding to the zero elements of $\hat{\beta}$). Finally, let $D_s$be the J_s-by-J_s demeaning matrix for school s, computed as described above. Then the demeaning matrix for the full sample is block diagonal:

$$D = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_S \end{pmatrix}. \tag{20}$$

As before, the demeaned vector of coefficients is $D\hat{\beta}$ and the variance-covariance matrix is DVD'. This variance-covariance matrix has rank equal to $\sum_s J_s - S$.

B.2 Sampling-adjusted standard deviations

For many of the models considered in the paper, I report the standard deviation across teachers of the teacher coefficients. Let $\hat{\theta}$be a J-vector of coefficients, normalized as described above within each of S schools, let V be the variance-covariance matrix, and let n be a vector of student counts.

The (weighted) variance of teachers' effects is

$$\hat{\text{var}}\left(\hat{\theta}\right) = \frac{1}{J-S}\hat{\theta}' \,\text{diag}\left\{\tilde{n}\right\}\hat{\theta}, \tag{21}$$

where $\text{diag}\{\tilde{n}\}$ is the J-by-J matrix with diagonal element j equal to $n_j/\bar{n}$ (where $\bar{n} = (1'1)^{-1}1'n$) and zeros off the diagonal. Note that this incorporates a degrees-of-freedom adjustment for the school-level normalization.

The standard deviation of teachers' estimated effects is merely the square root of the above expression. This overstates the standard deviation that would be obtained in an infinitely large sample. Let θ be the plim of $\hat{\theta}$, under the fixed-J asymptotics described in the text, and let $\hat{\theta} = \theta + u$, where u is sampling error and $E[uu'] = V$. This suggests that we can write the variance of the "true" (net of sampling error) effects as $\text{var}(\theta) = \text{var}\left(\hat{\theta}\right) - \text{var}(u)$, where these variances are computed across the elements of θ and weighted by n. The $\text{var}\left(\hat{\theta}\right)$ term is estimated as described above. $\text{var}(u)$ is estimated as $\frac{1}{J}\sum_j \bar{n}^{-1} n_j v_{jj}$, where $\bar{n} \equiv (1'1)^{-1}1'n$, as above, and v_{jj} is the jth diagonal element of V.

B.3 Computation of regressions with teacher indicators for multiple grades when there are no covariates

Several of the specifications used here include indicators for teachers in several grades simultaneously. The correlated random efffects analysis is the most involved, with indicators for 3rd, 4th, and 5th grade teachers in the same regression (equation (14)):

$$\tilde{A}_{i3} = T_{i3}\pi_{33} + T_{i4}\pi_{43} + T_{i5}\pi_{53} + e_{3i3} \tag{22}$$

$$\tilde{A}_{i4} = T_{i3}\pi_{34} + T_{i4}\pi_{44} + T_{i5}\pi_{54} + e_{3i4}. \tag{23}$$

Two computational challenges arise. First, not all of the π coefficients can be separately computed. The particular problem arises because I restrict the sample to students who do not change schools. The fitted values of the regressions would be unchanged were we to add a constant c to each element of the $\pi_{g,h}$ corresponding to a teacher at a particular school j and subtract the same constant from the similarly-defined elements of $\pi_{k,h}$ for some $k \neq g$. As a result, the mean of $\pi_{g,h}$ across all teachers in grade g at school j cannot be separately identified. I augment (22) and (23) with school indicators, then select one teacher in each grade at each school to exclude from the regressions.[24] I treat the excluded π coefficient as zero, with sampling variance zero. After estimating the regression, I normalize the coefficients of (22) and (23) to have mean zero across teachers in each grade at each school, using the procedure described above.

The second issue derives from the sheer size of the regression. Even after excluding the over-identified coefficients, each of the T_{ig} vectors has over 2,200 elements, and the full regression (after dropping redundant indicators) has 5,501 regressors. Numerical inversion of a matrix of this dimension may introduce inaccuracies. My focus on samples of students who do not

[24] The sample used for these regressions excludes schools where, due to insufficient mixing, the $[T_{i3}\ T_{i4}\ T_{i5}]$ submatrix corresponding to teachers at the school has rank less than $J_{s3} + J_{s4} + J_{s5} - 2$.

switch schools permits a simpler computation. Re-order the independent variables in equations (22) and (23) as $X = \left[X_{(1)}, X_{(2)}, \ldots, X_{(J)}\right]$, where $X_{(j)}$ contains the indicator for school j and the indicators for all teachers (in all three grades) at school j. Any sample student who ever appears in school j never appears in any other school, so $X'_{(j)}X_{(k)} = 0$ for all $j \neq k$. This ensures that $X'X$ is block-diagonal:

$$X'X = \begin{pmatrix} X'_{(1)}X_{(1)} & 0 & \cdots & 0 \\ 0 & X'_{(2)}X_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X'_{(J)}X_{(J)} \end{pmatrix}. \tag{24}$$

$(X'X)^{-1}$ is also block-diagonal, with blocks consisting of the inverse of the school-level design matrices:

$$\left(X'X\right)^{-1} = \begin{pmatrix} \left(X'_{(1)}X_{(1)}\right)^{-1} & 0 & \cdots & 0 \\ 0 & \left(X'_{(2)}X_{(2)}\right)^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(X'_{(J)}X_{(J)}\right)^{-1} \end{pmatrix}. \tag{25}$$

Each block has dimension of only a few dozen, so inversion is straightforward. The π coefficients (before the within-school normalization) and robust sampling variances are readily computed from $(X'X)^{-1}$. The covariances between the coefficients of equations (22) and (23) can be computed with

$$\operatorname{cov}\left(\Pi^4, \Pi^5\right) = \left(X'X\right)^{-1} X' \operatorname{diag}\left(\hat{e}_{i4}\hat{e}_{i5}\right) X \left(X'X\right)^{-1}. \tag{26}$$

This implicitly clusters on the individual student, and is equivalent to applying system OLS to the simultaneous equations (22) and (23).

B.4 Computation of regressions with teacher indicators for multiple grades when there are continuous covariates

In a few cases (e.g. the "saturated" model discussed in Section 6), I include continuous regressors Z along with the school and teacher indicators from several grades. These regressions have the form

$$y = X\Pi + Z\psi + \varepsilon. \tag{27}$$

Letting $W = [X\ Z]$ and $\Lambda = [\Pi'\ \psi']'$, we have $y = W\Lambda + \varepsilon$. Because the ψ coefficients are common across schools, $W'W$ is no longer block-diagonal, and the school-by-school strategy described above cannot be used directly here. In these models, I use a brute-force OLS regression estimator (implemented in Matlab) to compute the regression of the school de-meaned y on the de-meaned W. This may introduce numerical inaccuracy in the estimated coefficients,

$\hat{\Lambda}_0$. To avoid this, I use an iterative algorithm to obtain improved coefficient estimates. At each iteration t (beginning with t=1), there are two steps:

1. Treat the ψ parameters as known, using values from the previous iteration, $\hat{\psi}_{t-1}$. Regress $y - Z\hat{\psi}_{t-1}$ on X. The methods used in the previous section can be applied here, as $X'X$ is block diagonal with blocks corresponding to schools. Label the resulting coefficients $\hat{\Pi}_t$

2. Treating the $\hat{\Pi}_t$ coefficients as known, regress $y - X\hat{\Pi}_t$ on Z. Z typically contains only a handful of variables, so this is simple to calculate. Label the resulting coefficients $\hat{\psi}_t$, and use these as inputs to step 1 on the next iteration.

These steps are repeated until the coefficient vector converges. Convergence is considered to have been achieved when the maximum change in the regression residuals $e_t \equiv y - X\hat{\Pi}_t - Z\hat{\psi}_t$ from the previous iteration – that is, $\|e_t - e_{t-1}\|_{sup}$ – is less than $10^{-6}\sigma_y$.

This is essentially the Gauss-Seidel method, though the structure of the problem makes it possible to use only two sub-vectors of the full parameter vector Λ rather than stepping through each element of Λ separately as in typical implementations. It can be shown to be a contraction mapping on the sum of squared errors, so the coefficients necessarily converge to the OLS coefficients. Abowd et al. (2002) use a similar (in spirit, though not in detail) computational strategy.

In practice, the initial brute-force estimates are quite accurate, and only one or two iterations are required before convergence is achieved. As the iterative algorithm does not yield standard errors, I use a brute-force estimate of $(W'W)^{-1}$ to compute these.

C Additional Specifications

C.1 Teachers' observable characteristics

VAMs are used not only to estimate individual teachers' effects, but also to assess the relationship of teacher quality with teachers' observed characteristics (see, e.g., Clotfelter et al., 2006, 2007; Goldhaber and Brewer, 1997; Hanushek and Rivkin, 2006). These analyses replace the teacher indicators in VAM1, VAM2, or VAM3 with vectors of teacher observables – education, experience, etc. The tests developed in the main text can be applied to these models as well. Appendix Table C1 presents results for mathematics. (Results for reading are similar and are available from the author.) I focus on a short vector of teacher characteristics: An indicator for whether the teacher has a master's degree, a linear experience measure, an indicator for whether the teacher has less than two years of experience, and the teacher's score on the Praxis tests required to obtain elementary certification in North Carolina.[25] As in the other analyses,

[25]Each test is standardized among North Carolina teachers who took it in the same year, then (when multiple scores are available) scores are averaged across tests.

I restrict attention to students who can be assigned to valid teachers in each grade for which teacher characteristics will be controlled and who do not switch schools between grades. I further exclude students for whom I am unable to assemble complete characteristics for each of the relevant teachers.

Column 1 presents estimates from VAM1 of the effects of 4th and 5th grade teachers on 5th grade gains, controlling for school fixed effects and clustering the standard errors on the school. The 5th grade teacher coefficients echo those in the literature: A master's degree appears to make little difference, but inexperienced teachers have quite negative effects on student gains. Interestingly, inexperienced 4th grade teachers seem to have large *positive* effects on 5th grade gains, perhaps indicating that students quickly make up for time lost during 4th grade. See the discussion in Section 7.

Column 2 repeats the VAM1 specification, this time using the 4th grade gain as the dependent variable. The 4th grade teacher coefficients are consistent with those seen for 5th grade teachers in Column 1. But Column 2 also indicates that the 5th grade teacher's Praxis score is positively associated with the 4th grade gain score, while the coefficient on the dummy for an inexperienced 5th grade teacher is negative and nearly significant ($t = -1.85$). The hypothesis that all 5th grade teacher characteristics have zero coefficients is rejected ($p = 0.02$). This is clear evidence that the VAM1 exclusion restriction is violated by student sorting.

Columns 3 and 4 present the analysis of VAM2, modeling 5th grade scores in Column 3 and 3rd grade scores in Column 4. Results in Column 3 are similar to those in Column 2. In Column 4, none of the 5th grade coefficients are individually significant, but the test that all are zero is marginally significant (p = 0.11). Given the low power of my tests for analyses of teacher characteristics, which are only weakly correlated with student achievement in any grade, I interpret this as only mildly encouraging.

Columns 5 and 6 present the correlated random effects analysis that I use to evaluate VAM3, modeling 3rd and 4th grade gains, respectively, as functions of the characteristics of teachers in grades 3 through 5. I again consider two restricted models, one that constrains student ability to enter identically into each grade's gain score equation and another that allows different ability coefficients in different grades. The former model – corresponding to the version of VAM3 that is uniformly used in the literature – implies that the 5th grade teacher coefficients in columns 5 and 6 of Table 6 should be equal. Iin fact, we see a significant negative coefficient for the no experience indicator in the model for 4th grade gains and a marginally significant ($t = 1.67$) positive coefficient in the model for 3rd grade gains. The hypothesis of equal effects is decisively rejected (p=0.02). The less restrictive model requires only that the coefficients in columns 5 and 6 be proportional to one another. This restriction is consistent with the data (p=0.81). However, the OMD estimates indicate a factor of proportionality of -0.92. If we normalize $\tilde{\tau}_3 = 1$, defining "ability" to have a positive effect on 3rd grade gains, the model indicates that high ability students gain much *less* during 4th grade than their low ability peers. An alternative interpretation of this extremely counterintuitive result is that the test is unable to detect violations of strict exogeneity in this context. The correlated random effects test has power against violations of strict exogeneity only if classroom assignments depend on factors that are correlated with the included variables. As all of the coefficients except those for the inexperienced teacher indicator are small and far from statistically significant, and as even the inexperienced teacher coefficients are consistent with the model only with implausible coefficient estimates, the simplest interpretation is that VAM3 is poorly suited to identifying the

effects of teacher characteristics on student achievement. Indeed, when I extend the analysis to use the characteristics of 6th grade teachers – students are typically in middle school in 6th grade, and ability tracking is more pronounced – to strengthen the overidentification test (see Rothstein, 2008b), I reject proportionality of the 6th grade teacher coefficients.

C.2 Distinguishing between teacher and classroom effects using cross-cohort comparisons

In the main paper, I use the terms "classroom effects" and "teacher effects" interchangeably to describe the effects of being in a single classroom. Under certain circumstances a distinction between the two – between a teacher's effect that is the same every year and a classroom effect that may vary from year to year as the teacher is assigned new cohorts of students – may make it possible to obtain unbiased estimates of teachers' causal effects under weaker conditions than are considered in the text.

Let β_{tyc} be the effect of being in classroom c taught by teacher t in year y. (I suppress grade subscripts for notational simplicity.) We can decompose this into a permanent component associated with the teacher and a time-varying component associated with transitory aspects of the classroom in year y. Let $c(t,y)$ be the classroom taught by teacher t in year y, and assume that $\beta_{tyc(t,y)} = \theta_t + \upsilon_{ty}$. Here, θ_t is the teacher's effect, and υ_{ty} is the additional portion of the classroom effect. If we assume that the non-random assignments of students to classrooms are completely transitory – that the pre-assignment characteristics of students in classroom $c(t,y)$ are uncorrelated both with the characteristics of students in $c(t,y+1)$ and with the teacher's true effect θ_t – then the bias in $\hat{\beta}_{tyc(t,y)}$ will be uncorrelated from one year to the next. A decomposition of $\hat{\beta}$ into permanent teacher components and transitory components – a regression of $\hat{\beta}_{tyc}$ onto teacher indicators – would yield unbiased estimates of the permanent teacher components θ_t. Alternatively, the variance of θ_t across teachers can be estimated from the between-year covariance of β:

$$E\left[\beta_{tyc(t,y)}\beta_{t,y+1,c(t,y+1)}\right] = E\left[\theta_t^2\right] + E\left[\nu_{ty}\nu_{t,y+1}\right] + E\left[\theta_t\nu_{ty}\right] + E\left[\theta_t\nu_{t,y+1}\right]. \tag{28}$$

By the assumptions above, the final three terms are all zero. This sort of decomposition has been used by Hanushek et al. (2005) and Kane and Staiger (2008), among others.

This strategy relies crucially on the assumption that the assignments are uncorrelated across years. If some teachers are repeatedly assigned students with high expected gains that are not controlled in the VAM, this will create bias in the estimates of θ_t and $E\left[\theta_t^2\right]$. To evaluate whether assignments are in fact uncorrelated across years, I use students who were in 5th grade in 2000 to estimate a regression of 5th grade gains on all prior scores, absorbing 5th grade classroom indicators. This resembles the saturated VAM used above, but it excludes classroom indicators from prior grades. Using the coefficients from this regression, I form predicted 5th grade gains for each 5th grade student in both 2000 and 2001, then average these to the classroom level. These mean predicted gains represent bias in single-cohort estimates of VAM1. I also residualize the predicted gains against 4th grade scores to obtain the bias in VAM2. I then correlate the average predicted gains (or residual gains) of a teacher's students in 2000 with those for the same teacher's students in 2001.

In each VAM and in each subject, these cross-cohort correlations are positive and highly statistically significant. Evidently, teachers who are assigned good students in one year are typically assigned better-than-average students the next year as well. Thus, while data following teachers for several years may have some value for reducing bias from non-random assignments – the (observable) quality of a teacher's students is not perfectly correlated over time – the assumptions that would support simple corrections are not satisfied in the North Carolina data.

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.

_ , Robert H. Creecy, and Francis Kramarz, "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data," March 2002. Unpublished manuscript.

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, "Do Value-Added Estimate Add Value? Accounting for Learning Dynamics," March 2008. 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.

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.

_ , _ , and _ , "How and Why Do Teacher Credentials Matter for Student Achievement?," Working paper 12828, National Bureau of Economic Research January 2007.

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.

_ **and Dominic J. Brewer**, "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," *Journal of Human Resources*, Summer 1997, *32* (3), 505–523.

Hanushek, Eric A. and Steven G. Rivkin, "Teacher Quality," in Eric A. Hanushek and Finis Welch, eds., *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elsevier North-Holland, 2006, pp. 2–28.

_ , John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin, "The Market for Teacher Quality," February 2005.

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.

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.

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.

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.

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, "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables," Working Paper 26, Princeton University Education Research Section, 2008.

_ , "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Working Paper 25, Princeton University Education Research Section, 2008.

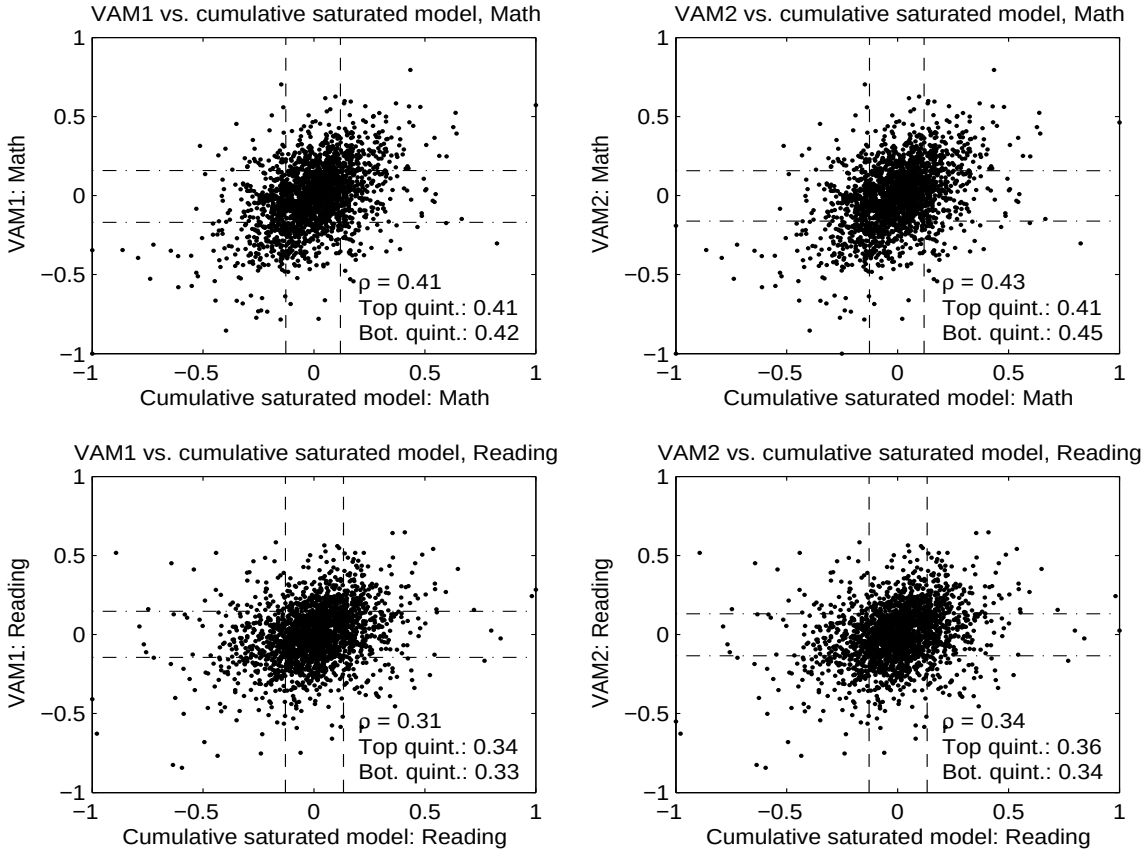
Sanders, William L. and June C. Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," Research Progress Report, University of Tennessee Value-Added Research and Assessment Center, November 1996.

_ **, 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.

Figure 1: Comparison of VAM1 and VAM2 for 4th grade teacher effects to estimates of 4th grade teachers' effects on 5th grade scores from the saturated VAM



Notes: The graphs show scatterplots of 4th grade teachers' estimated effects on 4th grade gains from VAM1 and VAM2 (vertical axes) against effects on 5th grade scores computed from a saturated VAM that controls for all past teachers and scores (horizontal axes). Teacher effects are normalized to mean zero within each school. Dashed lines show the 20th and 80th percentile of the estimated effects. Each panel shows the correlation between the two sets of estimates (weighted by the number of students taught, but not adjusted for sampling error), plus the fraction of teachers who are assigned to the top and bottom quintiles by the cumulative saturated model who are also assigned to these quintiles by VAM1 and VAM2.

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 match		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 the gain score VAM

	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)
Standard deviation of teacher coefficients								
5th grade teacher								
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	<0.01	<0.01	0.02	<0.01	<0.01	<0.01	0.03	<0.01
4th grade teacher								
Unadjusted SD					0.188	0.181	0.220	0.193
Adjusted SD					0.150	0.125	0.182	0.140
p-value					<0.01	<0.01	<0.01	<0.01
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: 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. Adjustments and p-values are based on heteroskedasticity-robust variances.

Table 4. Evaluation of the lagged score VAM

	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 teacher								
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	<0.01	<0.01	0.04	<0.01	<0.01	<0.01	0.16	<0.01
4th grade teacher								
Unadjusted SD					0.160	0.162	0.182	0.175
Adjusted SD					0.121	0.109	0.142	0.126
p-value					<0.01	<0.01	<0.01	<0.01
Continuous controls								
4th grade math score	-0.317	0.239	0.368	-0.213	-0.292	0.255	0.332	-0.229
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
4th grade reading score	0.195	-0.383	-0.218	0.380	0.189	-0.387	-0.206	0.379
	(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: Samples correspond to those in Table 3. Adjustments, p-values, and standard errors are robust to heteroskedasticity.

Table 5. Gain score VAM with student fixed effects: Correlated random effects estimates

	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	
Constant coefficients restricted model (OMD)						
Ratio, effect on G4 / effect on G3	1			1		
SD of G5 teacher effects	0.068	0.068		0.098	0.098	
Objective function	8,269			9,514		
95% critical value	1,685			1,685		
p value	<0.01			<0.01		
Scalar coefficients restricted model (OMD)						
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	<0.01			<0.01		

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.

Table 6. Magnitude of bias in VAM1 and VAM2 relative to a saturated 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				
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 saturated specification				
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 simple VAMs relative to the saturated 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.

Table 7. 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

Appendix Table A1. Construction of analytical samples

	Sample A		Sample B		Sample C	
	(1)		(2)		(3)	
Sample used in	Tables 3-4, Cols 1-4		Tables 3-4, Cols 5-8		Table 5	
Require student data in grades	3, 4, 5		3, 4, 5		2, 3, 4	
Require teacher links in grades	5		4, 5		3, 4, 5	
	N	**%**	**N**	**%**	**N**	**%**
Base sample	60,740	100%	60,740	100%	60,740	100%
Excluded for						
Missing record	3,772	6%	3,772	6%	3,772	6%
Missing test scores	1,825	3%	1,466	2%	5,226	9%
Changed schools	0	--	7,181	12%	15,083	25%
Missing/invalid teacher match	0	--	6,497	11%	9,400	15%
Only student in class	1	0%	10	0%	110	0%
Only class in school	0	--	384	1%	556	1%
Collinearity	0	--	769	1%	619	1%
Final sample	55,142		40,661		25,974	

Appendix Table A2. Average fraction of 5th grade classmates who were in the same 4th grade class

	Number of 4th grade classes at school						
	1	2	3	4	5+	2+	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Base sample							
# of students	1,515	6,032	12,508	12,441	14,717	45,698	47,213
# of schools	109	206	268	197	164	835	944
Fr. of 5th grade classmates who were in the same 4th grade class	1.00	0.52	0.35	0.27	0.21	0.31	0.33
Schools with perfect collinearity							
# of students	1,515	600	402	293	191	1,486	3,001
# of schools	109	35	16	7	4	62	171
Exclude schools with perfect collinearity							
# of students		5,432	12,106	12,148	14,526	44,212	44,212
# of schools		171	252	190	160	773	773
Fr. of 5th grade classmates who were in the same 4th grade class		0.51	0.35	0.27	0.20	0.30	0.30

Notes: A school has "perfect collinearity" if the J_4 indicators for 4th grade teachers and the J_5 indicators for 5th grade teachers together have rank less than $J_4 + J_5 - 1$.

Appendix Table C1. Models for the effects of teacher observable characteristics on math gains

	VAM1		VAM2		VAM3 (correlated random effects)	
	5th grade	4th grade	5th grade	4th grade	3rd grade	4th grade
	(1)	(2)	(3)	(4)	(5)	(6)
5th grade teacher						
MA degree	-0.05	-1.49	-0.75	-0.74	2.20	-1.12
	(1.30)	(0.99)	(1.30)	(0.90)	(1.43)	(1.04)
Experience	0.09	0.05	0.07	0.06	-0.04	-0.02
	(0.07)	(0.05)	(0.07)	(0.05)	(0.07)	(0.05)
1(Experience < 2)	**-5.35**	-2.87	**-5.95**	-2.02	3.65	**-4.13**
	(1.88)	(1.55)	(1.84)	(1.41)	(2.19)	(1.61)
Praxis score	1.50	**1.32**	**2.26**	0.41	-1.03	1.03
	(0.80)	(0.61)	(0.77)	(0.54)	(0.82)	(0.62)
4th grade teacher						
MA degree	-1.93	2.83	-1.19	1.92	0.67	**3.25**
	(1.30)	(1.53)	(1.12)	(1.23)	(1.33)	(1.62)
Experience	-0.10	0.07	-0.09	0.05	-0.07	0.13
	(0.07)	(0.08)	(0.06)	(0.06)	(0.07)	(0.08)
1(Experience < 2)	**5.21**	**-5.77**	**3.76**	**-3.96**	1.06	**-5.89**
	(1.75)	(2.00)	(1.57)	(1.66)	(2.09)	(2.16)
Praxis score	-1.48	**2.18**	-0.72	1.29	0.17	**2.53**
	(0.76)	(0.89)	(0.65)	(0.72)	(0.81)	(0.94)
3rd grade teacher						
MA degree					0.25	0.72
					(1.91)	(1.44)
Experience					0.18	**-0.16**
					(0.11)	(0.08)
1(Experience < 2)					-0.58	-1.04
					(3.05)	(2.24)
Praxis score					0.34	-0.05
					(1.07)	(0.80)
4th grade scores (*100)						
Math			**-0.31**	**0.36**		
			(0.01)	(0.01)		
Reading			**0.21**	**-0.22**		
			(0.01)	(0.01)		
N	20,251	20,251	20,251	20,251	18,239	18,239
R2	0.147	0.142	0.264	0.278	0.105	0.145
p-value, G5 teacher						
coeffs. = 0	<0.01	0.02	<0.01	0.11	0.13	0.04
Restricted specification, G5 teacher effects are equal in G3, G4 models						
p-value					0.02	
Restricted specification, G5 teacher effects are proportional in G3, G4 models						
Ratio, effect on G4 to effect on G3					-0.92	
p-value for overid. test					0.81	

Note: Dependent variables in each column are math gain scores in the relevant grade, multiplied by 100.