PRELIMINARY. COMMENTS WELCOME.

Do Value-Added Models Add Value?

Tracking, Fixed Effects, and Causal Inference

Jesse Rothstein*
Princeton University and NBER
November 20, 2007

Abstract:

Researchers conducting non-experimental studies of panel data often attempt to remove the potentially biasing effects of individual heterogeneity through the inclusion of fixed effects. I evaluate so-called "Value Added Models" (VAMs) that attempt to identify teachers' effects on student achievement. I develop falsification tests based on the assumption that teachers in later grades cannot have causal effects on students' test scores in earlier grades. A simple VAM like those used in the literature fails this test: 5th grade teachers have nearly as large effects on 4th grade gains as on 5th grade gains. This is direct evidence of non-random assignment. I use a correlated random effects model to generalize the test to more complex estimators that allow for tracking on the basis of students' permanent ability. The identifying restrictions of these estimators are again rejected. Teacher assignments evidently respond dynamically to year-to-year fluctuations in students' achievement. I propose models of this process that permit identification. Estimated teacher effects are quite sensitive to model specification; estimators that are consistent in the presence of (some forms of) dynamic tracking yield very different assessments of teacher quality than those obtained from common VAMs. VAMs need further development and validation before they can support causal interpretations or policy applications.

* Industrial Relations Section, Firestone Library, Princeton, NJ 08544. E-mail: jrothst@princeton.edu. I am grateful to Nathan Wozny and Enkeleda Gjeci for exceptional research assistance. I thank Orley Ashenfelter, Henry Braun, David Card, Henry Farber, Bo Honoré, Brian Jacob, Alan Krueger, David Lee, Austin Nichols, Amine Ouazad, Mike Rothschild, Cecilia Rouse, Diane Schanzenbach, Eric Verhoogen, and seminar participants at the Universities of Chicago, Houston, and Toronto for helpful conversations and suggestions. I also thank the North Carolina Education Data Research Center (especially Elizabeth Glennie and Kara Bonneau) and the North Carolina Department of Public Instruction for assembling, cleaning, and making available the confidential data used in this study, and the Princeton Industrial Relations Section and Center for Economic Policy Studies for financial support.

I. Introduction

Experiments are in short supply, but the demand for causal estimates is large. Non-experimental methods can identify causal effects, if their maintained assumptions about the assignment to treatment and the processes governing outcomes are correct. Some data structures allow tests of these assumptions, permitting an assessment of the validity of the treatment effect estimates (Ashenfelter and Card, 1985; Heckman, Hotz and Dabos, 1987).

A common approach to non-experimental inference is to condition on individual fixed effects. Advances in computing have made ever more complex models tractable. Models that decompose the variation in wages into worker and firm components, for example, have spawned a burgeoning literature on the determinants of firm wage effects.[1] Quite similar specifications have been used to distinguish the effects of student ability and teacher quality on students' test scores. On the basis of this sort of model, a consensus has formed that teacher quality is an extremely important determinant of student achievement, and that an important goal of educational policy should be the creation of incentives to attract, retain, and motivate high-quality teachers.[2]

Each teacher and each firm constitutes a distinct treatment and the effects of teachers and firms participating in any experiment[3] would therefore be uninformative about the effects of non-participants. An important advantage of non-experimental estimators is that they can be used to measure effects for the entire population. So-called "growth models" are increasingly important components of school accountability policy, and several states have incorporated econometric models for teacher quality into their teacher accountability programs. Some analysts (e.g., Gordon, et al., 2006) have gone so far as to suggest that schools be encouraged to fire as many as one quarter of new teachers who are found to make inadequate contributions to student learning.

[1] Abowd and Kramarz (1999) provide a somewhat outdated review. See also Abowd et al. (1999).

[2] See, e.g., Gordon et al. (2006), Hanushek (2002), Hanushek and Rivkin (2006), Koppich (2005), and Peterson (2006).

[3] Dee and Keys (2004) and Nye et al. (2004) use the Tennessee STAR experiment to study teachers' value added.

These applications require causal estimates. Workers and students are not randomly assigned to firms and teachers. The inclusion of controls for individual heterogeneity *may* permit causal identification, but only under rarely-specified assumptions for which there is little evidence.

This paper investigates the estimation of teachers' causal effects on student achievement through so-called "Value Added Models," or VAMs, applied to non-experimental data.[4] The focus on education is dictated by data availability; the theoretical analysis applies equally to other applications, in particular the analogous models for firm effects on worker wages, though of course the empirical results might not. I enumerate the identifying assumptions of the commonly-used VAMs and construct tests of these assumptions.

The most important assumption made in value added modeling concerns the assignment of students to teachers. Most VAMs assume that teacher assignments are uncorrelated with other determinants of student learning. (The most widely used VAM, the Tennessee Value Added Assessment System, or TVAAS, assumes that there is no heterogeneity in student growth rates.[5]) This is unlikely. In most schools, there is some degree of non-random assignment, either via formal "ability tracking" or informally, at the principal's discretion subject to the influence of parental requests. If the information used to form these assignments is predictive of future achievement growth, simple VAMs will not yield causal estimates.

My tests are based on a simple falsification exercise: Future treatments cannot have causal effects on current outcomes, and models that indicate such effects must be misspecified.[6] I

[4] Recent examinations of value added modeling include Ballou (2002), Braun (2005a, b), Harris and Sass (2006), McCaffrey et al. (2003), and a symposium in the *Journal of Educational and Behavioral Statistics* (Wainer, 2004). Clotfelter, Ladd, and Vigdor (2006) highlight the importance of non-random teacher assignments for value added analysis, the focus of the current study, but conclude that this is a relatively minor issue in North Carolina.

[5] See Ballou et al. (2004); Bock et al. (1996); Sanders and Horn (1994, 1998); Sanders and Rivers (1996); and Sanders et al. (1997). The TVAAS model allows error terms to be correlated across grades. But these error terms are in achievement *level* equations; in fact, TVAAS does not allow for observed or unobserved heterogeneity in *growth rates*.

[6] This is directly analogous to the strategy used by Heckman et al. (1987) to evaluate non-experimental estimators of the effect of job training on workers' wages. Heckman et al. argue that estimators which fail to eliminate differences in pre-

demonstrate that a simple VAM of the form typically used in the literature indicates large effects of current teachers on past achievement: A student's 5th grade teacher has nearly as large an "effect" on her 4th grade achievement growth as on her learning during 5th grade. This is direct evidence of non-random assignment.

Richer VAMs may be able to identify causal effects in the presence of sorting, but the appropriate model depends on the type of sorting used. It is useful to distinguish between two types: That based solely on students' permanent characteristics (e.g., IQ), and that based in part on the time-varying component of student performance. I refer to the former as "static tracking" – even if it is less formal and complete than an announced ability tracking policy – and the latter as "dynamic" tracking. If students entering a school are assigned permanently into advanced or remedial tracks, tracking is static. If students can be re-assigned after entrance when annual test scores indicate that the initial assignment was incorrect, however, the tracking is dynamic.

Some value added analyses have used enriched VAMs that absorb differences in student ability across classrooms via the inclusion of student fixed effects. These can identify teachers' causal effects in the presence of static tracking. If there is dynamic tracking, however, the available estimators are inconsistent for teachers' causal effects.[7]

The interpretation of student fixed effects VAMs thus depends crucially on the form of tracking that is used. To evaluate the static tracking assumption, I embed my falsification exercise in Chamberlain's (1982, 1984) correlated random effects model. Static tracking implies that estimators which fail to account for student heterogeneity will yield the same apparent effect of future teachers

training wages between the treatment and comparison groups must be relying on incorrect assumptions about the wage process or about the assignment of workers to training.

[7] A similar result applies to the separation of worker and firm heterogeneity in wages: Firm effects are identified by fixed effects regressions if worker-firm matches depend only on permanent worker characteristics, but not if mobility is related to the short-term innovation in a worker's productivity. I discuss below the infeasibility of estimators for models with dynamic assignment (Arellano, 2001) in the teacher or firm effects contexts.

in all grades, as each reflects the omitted student effect. That is, grade-6 teachers should appear to have similar effects on grade-4 and grade-5 scores. The data definitively reject this restriction.

I close by considering models of dynamic tracking that would permit the estimation of teachers' causal effects. I describe two sets of assumptions about the educational production function and about the teacher assignment process, appropriate for different settings, that permit identification. I present estimates using approximations to the implied estimators that are feasible given the available data. These indicate that tracking leads to important biases in teachers' estimated effects, as estimates that account for (a restricted type of) tracking are not very highly correlated with those from simpler VAMs. They also point to a second important, often overlooked factor: Students' past teachers appear to have continuing effects on their gains as they progress through school. Effects in the second year after contact (e.g., of the 4th grade teacher on the 5th grade gain) are negatively correlated ($\varrho = -0.5$) with effects in the first year, and the first year effects are poor proxies ($\varrho = 0.5$) for teachers' cumulative contributions. Correct models of tracking and of educational production are essential for the causal interpretation of value added estimates.

II. The Education Production Function and Value Added Modeling[8]

A. Educational Production

A student's achievement depends on the cumulative impact of inputs received to date from a variety of sources – family, school, peers, community, etc. – as well as on the student's permanent ability. School-based inputs are my focus here. A general educational production function is:

(1) $A_{ig} = f_g\left(S_i(g), \mu_i, \varepsilon_i(g)\right),$

where A_{ig} is the achievement score of student i in grade g, $S_i(g)$ contains the full history of school-based inputs from birth through grade g, μ_i is contains all non-school inputs (both family and

[8] I draw in this section on Harris and Sass (2006) and Todd and Wolpin (2003).

individual) that do not vary over time, and $\varepsilon_i(g)$ is a history of time-varying non-school inputs and random errors in each year.

If $f_g(\,)$ is linear, this yields an equation of the form

$$A_{ig} = S_{i1}\,\beta_1^{g} + \ldots + S_{ig}\,\beta_g^{g} + \mu_i\,\tau^{g} + \varepsilon_{i1}\,\varphi_1^{g} + \ldots + \varepsilon_{ig}\,\varphi_g^{g}, \tag{2}$$

where S_{ih} is a vector of school-based inputs in grade h, β_h^g ($h \leq g$) is the effect of grade-h school inputs on grade-g achievement, ε_{ih} and φ_h^g are the corresponding variables and coefficients for time-varying non-school inputs, and τ^g is the effect of permanent characteristics.[9]

B. Gain Scores

The difference between (2) and the corresponding equation for achievement in grade g-1,

$$A_{ig-1} = S_{i1}\,\beta_1^{g-1} + \ldots + S_{ig-1}\,\beta_{g-1}^{g-1} + \mu_i\,\tau^{g-1} + \varepsilon_{i1}\,\varphi_1^{g-1} + \ldots + \varepsilon_{ig-1}\,\varphi_{g-1}^{g-1}, \tag{3}$$

characterizes the "gain score," the change in achievement between grade g-1 and grade g:

$$\begin{aligned} \Delta A_{ig} \equiv A_{ig} - A_{ig-1} = {} & S_{i1}\,(\beta_1^{g} - \beta_1^{g-1}) + \ldots + S_{ig-1}\,(\beta_{g-1}^{g} - \beta_{g-1}^{g\,-1}) + S_{ig}\,\beta_g^{g} \\ & + \mu_i\,(\tau^{g} - \tau^{g-1}) + \varepsilon_{i1}\,(\varphi_1^{g} - \varphi_1^{g-1}) + \ldots + \varepsilon_{ig-1}\,(\varphi_{g-1}^{g} - \varphi_{g-1}^{g-1}) + \varepsilon_{ig}\,\varphi_g^{g}. \end{aligned} \tag{4}$$

μ_i enters into (4) only to the extent that it has different effects on achievement in grades g-1 and g. A focus on gains can therefore be seen as removing a student fixed effect in achievement. Of course, if ability has different effects in different grades, it cannot be treated as a fixed effect and is not removed by differencing.

C. Decay

In general, the grade-g gain depends on the full history of school and time-varying non-school inputs. With restrictions on the β coefficients, however, past inputs may disappear from the

[9] Boardman and Murnane (1979), Rivkin et al. (2005) and Hanushek (1979) discuss the limitations of this sort of specification. Todd and Wolpin (2003) emphasize that family inputs will respond, most likely in a compensatory fashion, to school inputs. Note also that (2) rules out any complementarities between school and non-school inputs, or between inputs of the same type across grades. I maintain this assumption throughout.

gain score equation. For example, if effects of inputs persist forever without decay (i.e., if $\beta_h^g = \beta_g^g$ and $\varphi_h^g = \varphi_g^g$ for all h < g), then grade-g gains depend only on grade-g inputs:

(5) $\Delta A_{ig} = S_{ig}\, \beta_g^g + \mu_i\, (\tau^g - \tau^{g-1}) + \varepsilon_{ig}\, \varphi_g^g.$

Alternatively, if inputs decay at a constant rate (i.e., $\beta_h^g = \xi^{g-h}\, \beta_g^g$ and $\varphi_h^g = \xi^{g-h}\, \varphi_g^g$ for some $0 \leq \xi \leq 1$ and all h < g), (4) can be re-arranged to express grade-g gains as a function of current inputs and the lagged achievement level:

(6) $\Delta A_{ig} = A_{ig-1}\, (\xi - 1) + S_{ig}\, \beta_g^g + \mu_i\, (\tau^g - \xi\, \tau^{g-1}) + \varepsilon_{ig}\, \varphi_g^g.$[10]

Both zero decay and constant decay are strong restrictions on the educational production process, and it is easy to imagine ways in which they might be violated.[11] For example, vocabulary drills might raise students' scores on the end-of-grade test without having much effect on their long-run achievement, implying rapid decay. The rate of decay may vary with teaching styles: Compare a teacher who focuses on drills with one who is skilled at teaching her students to enjoy reading for pleasure. The achievement effects of the latter are likely to decay slowly if at all, and indeed it may have larger effects on students' long-run scores than in the short term. If so, the rate of decay may be faster for the "drills" treatment than for the "instill a taste for reading" treatment. There is little basis for either the zero-decay or the constant-decay restrictions.

The general model (4) imposes only one important restriction: There is no effect of *future* inputs on current gains. That is, neither S_{ih} nor ε_{ih} enter into the equation for ΔA_{ig}, g < h. This restriction arises naturally from the structure of the model, and forms the basis for my tests.

[10] See Harris and Sass (2006). If there is any measurement error in the annual test score, this will be a component of ε_{ig} that decays fully in the next year. Generalizing (6) to allow for this when other inputs do not decay so quickly introduces a correlation between A_{ig-1} and the error term, and some researchers (Koedel and Betts, 2007; Ladd and Walsh, 2002) instrument for A_{ig-1} with A_{ig-2}.

[11] The decay of past inputs is closely related to the degree to which test scores in different grades measure the same things and use the same scale. The no-decay and constant-decay properties are not invariant to re-scaling of test scores unless the same transformation is applied to each grade's scores.

D. A Simple VAM

It is difficult to measure and control for the full set of relevant schooling inputs, particularly as the sorts of inputs that are readily measured – class sizes, teacher experience, school spending – may be less important than the hard-to-capture efficiency with which these inputs are used (Hanushek, 1981, 1986). A widely-used alternative, and the focus of the current analysis, is to include a full set of indicators for students' grade-g teachers in S_{ig}. There are typically no other classroom-level controls; a teacher's "effect" is defined to equal the effect of being in a particular classroom and incorporates both the teacher's quality and the total effects of all other classroom-level determinants of performance (including, e.g., peers and class size).

The simplest value added specification is a regression of grade-g gain scores on indicators for grade-g teachers, perhaps with controls for a few student characteristics:

$$(7) \qquad \Delta A_{ig} = T_{ig}\, \beta_g^{\ g} + X_i\, \theta^g + e_{ig}.$$

Here, T_{ig} is an exhaustive set of teacher indicators and X_i is a vector of student characteristics. My notation treats X as non-time-varying: Most of the regressors that are typically available are constant across grades.[12] Equation (7), by excluding teachers from grades 1 through g-1, imposes the zero decay assumption discussed above, though this could be loosened by including past teacher assignments as additional explanatory variables.[13]

(7) attributes all school-level determinants of gain scores to teachers. As an alternative, one can normalize the β coefficients to have mean zero within each school. This avoids the need to model the assignment of students to schools, though because it prevents comparisons of teacher quality across schools it is unsuitable for most policy applications. I focus throughout on within-

[12] Typical regressors are race, gender, and free lunch status. Even the last varies little over time in practice.

[13] Some studies include the lagged achievement score as a control variable, either instrumented or not. As noted above, this allows for decay of past inputs, but at a constant, uniform rate.

school variation in β. Identification of this component is necessary but not sufficient for identification of the full β vector.

Estimates of specifications much like equation (7) form the basis for most value added analyses.[14] Many authors model the coefficients on the teacher indicators as random effects, while others use fixed effects techniques. The two types of models differ in their abilities to accommodate X variables: Random effects models typically require that teacher quality be orthogonal to any included controls (Ballou, et al., 2004), while fixed effects models can accommodate a correlation between teacher quality and observed controls, at some cost in efficiency. The two models rely on common assumptions about the relationship between teacher assignments and *unobserved* determinants of student achievement. As these assumptions are the focus of my analysis, my results (primarily from fixed effects estimators) are equally applicable to both sorts of models.

Value added analyses cannot identify teachers causal effects under *any* exogeneity assumptions if we maintain the conventional definition that sampling error must go to zero as the sample size grows toward infinity: Realistic assumptions would allow the number of teachers to grow with the number of students, keeping the number of students per teacher – the class size – approximately fixed and ensuring that sampling errors remain non-trivial. I focus on a weaker definition: I refer to teacher effects as "identified" if they would be accurately estimated as the number of students in the sample went to infinity with the number of teachers fixed. If teacher effects are identified under these unrealistic asymptotics, VAMs can be used in compensation and retention policy with appropriate correction for the sampling errors that arise with finite class sizes; if not, it is unreasonable to treat them as noisy but unbiased estimates of causal effects.

[14] See, e.g., Aaronson et al. (2007) and Rockoff (2004). The TVAAS model is expressed as a mixed model for level scores that depend on all lagged inputs, but the essential identification strategy is of this form. Other studies include the lagged achievement score as a control variable. See, e.g., Kane et al. (2006); Clotfelter et al. (2007); Goldhaber (2007); and Jacob and Lefgren (2005). As noted above, this allows for input effects to decay at a constant rate.

Even under this definition, identification of the β coefficients – or of summary statistics like the variance of β across teachers – requires strong assumptions about teacher assignments. Specifically, T_{ig} must be orthogonal to all other determinants of grade-g gains. Even with zero decay of teachers' effects, any correlation between T_{ig} and $(\mu_i - E[\mu_i \mid X_i])$ – if $\tau^g \neq \tau^{g-1}$ – or ε_{ig} will introduce bias. The first rules out static tracking, if permanent characteristics influence gain scores. The second puts strong restrictions on dynamic tracking. We require at least contemporaneous exogeneity, $E[T_{ig}\, \varepsilon_{ig}] = 0$, and most likely something stronger: Although in principle the grade-g teacher assignment could depend on the grade-h gain (h<g), this can be accommodated only if ε_{ig} is uncorrelated with ε_{ih}. Without strong restrictions on the serial correlation of ε, teacher assignments must be independent of test scores (conditional on X_i) in each previous grade.

E. The Student Fixed Effects VAM

Some studies adopt a richer specification that allows teacher assignments to depend on unobserved, permanent components of the error term, absorbing these components with student fixed effects in the gain equation (i.e. with student-specific trends in achievement).[15] The model is:

(8) $\Delta A_{ig} = T_{ig}\, \beta_g^{\,g} + D_i\, \mu + e_{ig},$

where D_i is a set of indicators for each student and μ a set of student-specific coefficients. This again relies on the assumption of zero decay.[16]

Estimation of (8) requires at least two observations on gain scores for each student. To illustrate the identifying assumptions, suppose that we observe gain scores in grades 1 and 2:

(9) $\Delta A_{i1} = T_{i1}\, \beta_1^{\,1} \qquad\qquad + D_i\, \mu + e_{i1}$

(10) $\Delta A_{i2} = \qquad\qquad T_{i2}\, \beta_2^{\,2} + D_i\, \mu + e_{i2}$

[15] See, e.g., Harris and Sass (2006), Koedel and Betts (2007); Jacob and Lefgren (2005); Rivkin et al. (2005); and Boyd et al. (2007).

[16] Even under the assumptions discussed in the text, strong assumptions about the decay process would be required to identify the effects of lagged teachers in an augmented version of (8) that allowed for decay.

The mean gain for student i is:

(11) $M_i \equiv ½ [\Delta A_{i1} + \Delta A_{i2}] = ½ [T_{i1}\, \beta_1^{\,1} + T_{i2}\, \beta_2^{\,2}] + D_i\mu + ½(e_{i1} + e_{i2})$.

The estimating equation subtracts the mean gain for student i from each year's gain score:

(12) $\Delta A_{ig} – M_i = ½T_{ig}\, \beta_g^{\,g} – ½\, T_{i3\text{-}g}\, \beta_{3\text{-}g}^{\,3\text{-}g} + ½\, e_{ig} – ½\, e_{i3\text{-}g}$,

Demeaning eliminates the $D_i\, \mu$ term. But both grades' teacher assignments and the transitory error terms from both grades' gain equations enter into the equations for each grade's de-meaned gain. Whenever the number of grades, G, is small enough that G^{-1} is non-negligible, as will be true in any imaginable value added analysis, a non-zero correlation between the error in one grade and the teacher assignment in that or any other grade will bias the estimated β coefficients.

The advantage of the student fixed effects model is that it is robust to correlations between teacher assignments and the permanent component of the test score error, μ_i, at least insofar as this component enters with the same loading into each grade's gain equation (that is, if $\tau^g – \tau^{g-1}$ is constant across g). But identification of the β coefficients requires strict exogeneity of teacher assignments conditional on μ: There can be no correlation between the grade-g teacher assignment and the time-varying residual ε in any past *or future* grade. This is stronger than the corresponding assumption for the simple VAM without decay, which even at its strongest required only that the grade-g teacher assignment be unrelated with *past* residuals. It is easy to imagine how it might be violated: It requires in effect that principals decide on classroom assignments for the remainder of a child's career on the day that the child begins kindergarten. If instead teacher assignments are updated each year and depend on both the student's permanent ability and her performance during the previous year, strict exogeneity would be violated.

A standard strategy in panel data analysis when strict exogeneity does not hold is to work with the first-differenced equation,

(13) $\Delta A_{i2} – \Delta A_{i1} = T_{i2}\, \beta_2^{\,2} – T_{i1}\, \beta_1^{\,1} + e_{i2} – e_{i1,}$

rather than the de-meaned equation, instrumenting for the explanatory variables with lagged values (Anderson and Hsiao, 1981; Arellano and Bond, 1991).[17] But note that here T_{ig} and T_{ig-1} are each vectors of hundreds or thousands of teacher indicators, and lagged teacher assignments are unlikely to have much predictive power for them (and are invalid in any case if the zero decay assumption fails). Thus, if teacher assignments may depend on student ability, there is no practical estimator that is robust to violations of strict exogeneity.

F. Testing the Identifying Assumptions

Panel data on teacher assignments can be used to test for violations of the exogeneity assumptions that underlie the basic and student fixed effects VAMs. An informal test of the basic VAM can be formed by examining the correlation between T_{ig} and ΔA_{ig-1}. A non-zero correlation – an apparent effect of grade-g teachers on gains in g-1 – suggests that T_{ig} is correlated with either μ_i or ε_{ig-1}.[18] Coupled with evidence that the residual from the gain equation is serially correlated, this would strongly imply that T is endogenous in (7).[19]

The student fixed effects VAM can rationalize this result as the consequence of a correlation between μ and teacher assignments. But an extension of the test can be used to evaluate the strict exogeneity assumption on which it relies. With this assumption, any apparent effect of (for example) 6th grade teachers on 4th grade gains arises only because 6th grade teacher assignments depend on and are correlated with μ. Grade-6 teachers who are assigned high-μ students will appear to have positive effects and those with low-μ students will appear to have negative effects. We should see the same pattern of apparent effects on grade-5 gains, positive for grade-6 teachers with

[17] Equation (13) can be estimated by OLS without instrumentation (Koedel and Betts, 2007, implement a strategy like this) if the grade-g error term is uncorrelated with teacher assignments in grades g-1, g, and g+1. This is slightly weaker (when G>2) than strict exogeneity. It is difficult to imagine an assignment process that would satisfy this but would not satisfy strict exogeneity, however.

[18] A formal test would have to rule out the potential explanation that T_{ig} is correlated with T_{ig-1} and that the omission of the latter drives the result. This might be done by controlling for T_{ig-1} directly.

[19] I describe below a detailed, restrictive set of assumptions under which T_{ig} might be correlated with $A_{i,g-1}$ without being correlated with the error term in (7). But these would not hold in general.

high-μ students and negative for teachers with low-μ students. An indication that a grade-6 teacher has different "effects" on grade-4 and grade-5 gain scores would indicate that not all of the omitted variables bias in (7) derives from μ, implying that the strict exogeneity assumption fails and that the student fixed effects VAM does not identify teachers' causal effects. I develop this test – a direct application of Chamberlain's (1982, 1984) correlated random effects model – formally in Section V.

III. Data and Sample Construction

My empirical analysis uses administrative data on public school students in North Carolina. North Carolina has been a leader in the development of linked longitudinal data on student achievement, and in 2006 was one of the first two states approved by the U.S. Department of Education to use "growth-based" accountability models. The data, assembled and distributed by the North Carolina Education Research Data Center (NCERDC), have undergone extensive cleaning to ensure accurate matches between the component administrative data systems. They have been used for several previous value added analyses (see, e.g., Clotfelter, et al., 2006; Goldhaber, 2007).

The dataset contains scores from end-of-grade tests in math and reading in grades 3 through 8. I focus on reading scores; analyses for math scores are available upon request. The tests purport to use a so-called "interval" scale, so that a one point increment corresponds to an equal amount of learning at each grade and at each point in the within-grade distribution.[20] I standardize the scale scores so that the distribution of 3rd grade scores has mean zero and standard deviation one. This preserves the interval scale.

The importance of interval scales to value added modeling is not widely appreciated.[21] As noted earlier, most VAMs rely on a zero-decay assumption: A 3rd grade teacher who raises her

[20] The scale scores are linear transformations of the estimated θ parameters from a 3-parameter Item Response Theory model (Sanford, 1996, p. 20).

[21] But see Ballou (2002) and Yen (1986) Many authors (see, e.g., Boyd, et al., 2007; Kramarz, Machin and Ouazad, 2007; Rivkin, et al., 2005) standardize scores separately in each grade. This destroys any interval scale unless the variance of achievement is indeed constant across grades.

students' end-of-grade scores by 1 point also adds one point to their scores in all future grades. This restriction is not scale-invariant, and is particularly implausible if scores do not have the interval property. The assumption that ability is equally important to each grade's gain is also sensitive to scaling. Interval scaling makes these assumptions plausible, but does not guarantee them. In the specifications below, I allow for both arbitrary, heterogeneous decay of teachers' effects and grade-specific effects of ability, as in (4). I also explore alternative scalings.

A. Empirical Properties of Gain Scores

Table 1 presents summary statistics for reading test scores and gains, computed over all available observations on students who were in 3rd grade in 1999, 4th grade in 2000, 5th grade in 2001, or 6th grade in 2002.

The table indicates that test scores are correlated about 0.82 in adjacent grades. A 1996 report estimates that the North Carolina reading score's test-retest reliability – based on administrations of alternative forms of the test one week apart – is 0.86 (Sanford, 1996, p. 45), indicating that 14% of the variance of test scores is transitory noise. This places an upward bound on the year-to-year correlations, one that is nearly met. Because gain scores retain all of the noise from the two component tests but eliminate much of the signal, they are necessarily less reliable, particularly when the signal is highly correlated over time. A simple calculation indicates that the correlation in true achievement between adjacent years is 0.93 and that individual gain scores have reliability around 0.3.[22]

The lower right portion of Table 1 shows the correlation between gains in different grades. The correlation between the grade-4 gain (the change in scores between grade 3 and grade 4) and the

[22] This and the following calculations are derived in an Appendix, and assume that the variance of achievement is constant across grades. Stake (1971) presents a nearly identical calculation; see also Kane and Staiger (2001, 2002). Rogosa (1995) argues that reliability is higher if gains are uncorrelated with initial levels, but Table 1 indicates that this is not the case here.

grade-5 gain is -0.43. Noise in the grade 4 score enters positively into the grade-4 gain and negatively into the grade-5 gain, biasing the observed correlation downward relative to the correlation in true gains. Based on gain score reliability of 0.3, the observed correlation is consistent with a correlation in true gains of -0.27. A similar calculation indicates that the correlation between true grade-5 and grade-6 gains is -0.23.

Finally, the correlation between observed grade-4 and grade-6 gains is 0.01. Though small, this is significantly different from zero. Using a reliability of 0.3, the observed correlation indicates a correlation between true gains of 0.03.

These calculations have two implications for value added modeling. First, gain scores are exceptionally noisy. It will be difficult to pick out effects on true gains with any reliability (Ballou, 2002), and mean reversion is likely to be an important factor. Second, the weak correlation between grade-4 and grade-6 gains suggests that there is little permanent heterogeneity across students in the rate of gain – either var(μ) is small or τ is nearly constant across grades.

B. Samples

The North Carolina data do not contain explicit teacher identifiers, but they do identify the teacher (or other school staff member) who administered the end-of-grade tests. In the elementary grades, this was usually the regular teacher. I follow Clotfelter et al. (2006) in using a linked personnel database to identify test administrators who had regular teaching assignments. I count a test administrator as a valid teacher if she taught a "self-contained" (i.e. all day, all subject) class for the relevant grade in the relevant year, if that class was not coded as Special Education or Honors, and if at least half of the tests that she administered were to students in the correct grade. 73% of 5th grade tests were administered by "valid" teachers.

In North Carolina, 6th grade students are typically in middle school and have different teachers for each subject. The end-of-grade exam need not be taken in the relevant subject-matter classroom, and students may not move together across classes. I do not attempt to identify "valid" 6th grade teachers, nor do I attempt to estimate any grade-6 teacher's causal effect. Instead, I use 6th grade class assignments as a source of information about student sorting, on the assumption that students who take the exam together share at least one class. I use the relationship between earlier achievement and grade-6 classroom assignments to identify the form that tracking takes. Classes are likely more heavily tracked in 6th grade than in earlier grades, so class groupings may proxy for unobserved student ability even better than do those in earlier grades.

I work with two samples. The first consists of 65,582 students from 860 schools who were in 5th grade in 2000-2001 and who could be matched with valid teachers in that year. I exclude students whose longitudinal records provide inconsistent measures of race or gender or multiple observations in any year, as these might indicate mismatches. I also exclude those whose teachers have fewer than 12 sample students or whose schools have fewer than two included teachers.

The second sample is more homogenous. To simplify the correlated random effects analysis, I use only students who were in 3rd grade in 1998-1999 and who progressed at the normal rate through 6th grade in 2001-2002, without skipped or repeated grades or missing test scores in any year. I exclude students who changed schools in 4th or 5th grade as well as those whose 4th or 5th grade teachers are invalid according to the definition outlined above. I do not require that the 3rd or 6th grade test administrator be a valid teacher, though I do track the identity of each, and I treat the group with which each student took the exam as a reasonably accurate proxy for the degree of across-classroom sorting.[23] I exclude schools with only a single teacher in the sample from any of grades 3, 4, 5, and 6.[24] The final sample consists of 21,101 students. There are 457 schools, 1,760

[23] The median school in my sample has 4 5th grade teachers but sees its students dispersed across 9 6th grade teachers. To avoid estimating effects from very small samples, I re-assign 6th grade teachers who had very small shares of the sample to a composite ID.

[24] Schools here are those that the students attend in grades 3-5. I also exclude a few students to eliminate perfect collinearity between teacher identifiers (as when an entire 4th grade class transitions to the same 5th grade teacher).

4th grade teachers, and 1,764 5th grade teachers represented. Students are grouped into 2,005 and 1,641 test administration groups in 3rd and 6th grade, respectively.

Table 2 presents summary statistics for the two samples and for the population, the latter including all available observations for each variable. The more homogenous samples, which will tend to exclude students who move frequently, have higher average achievement levels but similar patterns of gain scores. The Data Appendix describes each sample in more detail.

IV. Preliminary Estimates

Before implementing my formal tests, I present preliminary models that illustrate VAM estimation and suggest the importance of considering the dynamic path of educational production and teacher assignments. I focus on a very simple value added model with teacher effects but no other controls, fit to the first, broader sample:

$$(14) \qquad \Delta A_{i5} \;=\; T_{i5}\,\beta_5^5 \;+\; e_{i5}.$$

Model (14) has $R^2 = 0.098$ and adjusted $R^2 = 0.050$. By comparison, a model that includes only school effects has $R^2 = 0.048$ and $\overline{R}^2 = 0.034$.

The 3,013 β_5^5 coefficients can be summarized by their standard deviation across teachers, after normalizing them to have mean zero within each school.[25] This is 0.145, indicating that a student whose teacher is one standard deviation (of achievement levels) above average will see her achievement improve relative to the average by one seventh of a standard deviation over the course of the year. This corresponds to over one quarter of a standard deviation in the gain score distribution, and is similar to what has been found in other studies (e.g., Aaronson, et al., 2007; Kane, et al., 2006; Rivkin, et al., 2005).

[25] Both the mean and the standard deviation are weighted by the number of students taught by each teacher. The standard deviation is adjusted for the degrees of freedom absorbed by the normalization.

This overstates the standard deviation of true teacher effects, as it also incorporates a component due to sampling variance. Following Aaronson, Barrow, and Sander (2007), note that $E[(\hat{\beta}_{5j}^{5})^2] = (\beta_{5j}^{5})^2 + E[(\hat{\beta}_{5j}^{5} - \beta_{5j}^{5})^2]$, where β_{5j}^{5} is the effect of teacher j and $\hat{\beta}_{5j}^{5}$ is the corresponding estimate. Thus, the variance of the true β coefficients – the mean of $(\beta_{5j}^{5})^2$ across teachers – can be computed as the difference between the observed variance and the average sampling variance. I use a heteroskedasticity-robust estimator for this. The implied standard deviation of teachers' "true" effects is 0.106. This again resembles existing estimates.

A robust Wald test rejects (with a p value below 0.0001) the hypothesis that there is no variation in the β coefficients within schools. But this test obscures as much as it reveals, as there is a great deal of heterogeneity across schools in the apparent importance of teachers. Under the null hypothesis that teachers don't matter, the p value for a test that all of the teachers from school k have identical coefficients will have a uniform distribution on [0, 1]. By contrast, if there are true differences across teachers at school k, the p value will come from a distribution that peaks at zero and has lower density at high values. The rate at which the density declines depends on the test's power, and we may not be likely to reject the null hypothesis at small schools, but very high p values should be unlikely. Figure 1 shows the histogram of school-level p values. There is indeed a large spike at zero: At about 120 of the 860 schools in the sample, the p value is less than 0.05; at another 150 schools, the p value is between 0.05 and 0.2. But beyond this point, the density remains fairly stable. 327 schools have p values above 0.5, with no fall-off at higher values. The overall picture is of a mixture of a uniform distribution with one skewed toward zero. Taking the 327 (38%) schools with p values above 0.5 as coming from the uniform distribution, the implication is that there are no differences among teachers at perhaps three quarters of schools. The result that teachers matter appears to be driven by the remaining quarter of schools, where there are clear differences. This is not merely a reflection of low power: Figure 1 also shows the histogram of p values at schools with 5 or more teachers, where power should be greater. This has the same pattern as the overall sample.

A. Counterfactual value added estimates

As a first step toward evaluating the identifying assumptions discussed in Section II, I re-estimate model (14), substituting gain scores in other grades as the dependent variable. Table 3 summarizes the results of this analysis. Column 1 describes the model for grade-5 gains. Columns 2 and 3 describe models for grade-4 and grade-6 gains, respectively, and column 4 describes a model in which the dependent variable is the cumulative gain score from 3rd to 6th grade, $A_{i6} – A_{i3}$. In each case, the explanatory variables are grade-5 teacher indicators.

As grade-5 teachers likely have only small effects on grade-6 gains and can have no effects on grade-4 gains, we should not expect these specifications to have much explanatory power if teacher assignments are exogenous. In fact, the R^2 statistics are comparable to those in the original specification, and the hypothesis that the teacher effects are all zero is rejected at any reasonable confidence level in all four models. The bottom rows show the standard deviations of the estimated "effects." These are nearly as large when the dependent variable is the lagged or lead gain or the long-run cumulative gain as when it is the contemporaneous gain. This casts doubt on the causal interpretation of simple value added models like (14).

The upper portion of Table 4 presents correlations between the coefficients in the various models from Table 3. The correlation between a teacher's estimated effect on 5th and 4th grade gains is -0.41. The correlation between 5th and 6th grade effects is also strongly negative, -0.54. By contrast, the correlation between a teacher's effects on 5th grade gains and on cumulative gains across three years is positive but quite small, 0.12.[26]

[26] The cumulative effect is more strongly correlated with the effects on 4th and on 6th grade gains, even after correcting for sampling covariance deriving from the use of the same test scores in the cumulative and the 4th and 6th grade gains.

These correlations reflect the combination of true relationships and sampling error, which is almost certainly correlated across models. This is most obvious for comparisons between 5th grade gains and 4th or 6th grade gains, for which the sampling error is strongly negatively correlated: Any positive, non-persistent shock to, for example, students' 4th grade scores will inflate $\hat{\beta}_{5j}^{4}$ – the estimated effect on the 4th grade gain – and reduce $\hat{\beta}_{5j}^{5}$. The lower portion of Table 4 presents sampling-adjusted estimates of the correlations of the true coefficients across specifications.[27] The adjustment has remarkably little effect, and the correlations between coefficients in models for adjacent grades' gains are still quite negative. In other words, this simple VAM indicates that teachers who have positive effects on 5th grade gains tend to have negative effects on 4th and 6th grade gains. The correlation between the 5th grade model and the cumulative model remains small.

There are several candidate explanations for these results. Begin with the negative correlation between β_{5j}^{6} and β_{5j}^{5}, which we have established is not attributable to sampling error. This could be causal, if teachers who raise students' 5th grade gains reduce their future potential (perhaps by teaching "cramming" skills with positive short-term but negative long-term returns). It could also reflect bias from equation (14)'s failure to account for static tracking, persistent student heterogeneity in the rate of achievement growth that is correlated with teacher assignments. This explanation would suggest that the true variance of 5th grade teachers' effects on 5th grade gains is larger than is indicated by the basic VAM, as the better teachers are, on average, assigned students with flatter growth trajectories. Finally, the result could reflect dynamic tracking: If students are assigned to 5th grade teachers on the basis of transitory shocks in 4th grade, these could generate

[27] I discussed earlier how to compute $V(\beta_{5j}^{h})$ from $V(\hat{\beta}_{5j}^{h})$ and $V(\hat{\beta}_{5j}^{h} - \beta_{5j}^{h})$. A similar formula applies to covariances. Assuming that sampling errors are uncorrelated with true coefficients, $cov(\beta_{5j}^{g}, \beta_{5j}^{h}) = cov(\hat{\beta}_{5j}^{g}, \hat{\beta}_{5j}^{h}) - cov(\hat{\beta}_{5j}^{g} - \beta_{5j}^{g}, \hat{\beta}_{5j}^{h} - \beta_{5j}^{h})$. For computational simplicity, I restrict the sample for these calculations to students for whom both grade-g and grade-h gain scores are observed. I allow for arbitrary clustering of errors for the same student across grades.

differences across 5th grade teachers in growth paths in all future grades. In this case, it is unclear how estimates of β_5^5 from (14) relate to true causal effects.

The sizable effects of 5th grade teachers on 4th grade scores rule out the causal explanation, as teachers can have no causal effects on their students' *prior* gains. Students are evidently assigned to teachers in a way that correlates with past achievement gains.

V. The Full Lags and Leads Specification

A. Correlated Random Effects

Chamberlain's (1982, 1984) correlated random effects model can be used to distinguish between static and dynamic tracking as explanations for the non-causal estimates in Table 3. I develop the model in its general form, then describe its application in my data.

Begin with the general gain score equation (4), using teacher indicators as the only observed explanatory variables. This allows the grade-g gain to depend on individual ability and on the full history of teacher assignments and unobserved errors. Simplifying notation slightly,[28]

$$(15) \qquad \Delta A_{ig} = \sum_{h=1}^{g} T_{ih}\tilde{\beta}_h^g + \mu_i\tilde{\tau}^g + \sum_{h=1}^{g}\varepsilon_{ih}\tilde{\varphi}_h^g .$$

Now consider a linear projection of the permanent heterogeneity term, μ_i, onto the full sequence of teacher assignments in grades 1 through G.

$$(16) \qquad \mu_i = T_{i1}\,\lambda_1 + \ldots + T_{iG}\,\lambda_G + v_i,$$

with $E[v_iT_{ih}] = 0$ for h=1,…,G. If teacher assignments are independent of μ, all of the λ coefficients are identically zero; otherwise, some or all may be non-zero. Substituting (16) into (15), we obtain

$$(17) \qquad \Delta A_{ig} = \sum_{h=1}^{g} T_{ih}\left(\tilde{\beta}_h^g + \lambda_h\tilde{\tau}^g\right) + \sum_{h=g+1}^{G} T_{ih}\left(\lambda_h\tilde{\tau}^g\right) + v_i\tilde{\tau}^g + \sum_{h=1}^{g}\varepsilon_{ih}\tilde{\varphi}_h^g .$$

[28] The relation between the earlier and the new coefficients is: $\tilde{\beta}_h^g = \beta_h^g - \beta_h^{g-1}$ (for h<g); $\tilde{\beta}_g^g = \beta_g^g$; $\tilde{\tau}^g = \tau^g - \tau^{g-1}$; $\tilde{\varphi}_h^g = \varphi_h^g - \varphi_h^{g-1}$ (for h<g); and $\tilde{\varphi}_g^g = \varphi_g^g$.

Define $\pi_h^g = \widetilde{\beta}_h^g + \lambda_h \gamma^g$ when h≤g and $\pi_h^g = \lambda_h \gamma^g$ for h>g. Then (17) becomes

$$(18) \qquad \Delta A_{ig} = \sum_{h=1}^{G} T_{ih} \pi_h^g + v_i \gamma^g + \sum_{h=1}^{g} \varepsilon_{ih} \widetilde{\varphi}_h^g \; .$$

Recall that the identifying assumption of the student fixed effects VAM is that teacher assignments are strictly exogenous conditional on μ, $E[\varepsilon_{ih} \mid T_{i1}, \ldots, T_{iG}] = 0$ for each h. If so, an OLS regression of grade-g gains onto teacher indicators in grades 1 through G identifies the π_h^g coefficients.

Equation (17) places strong restrictions on these coefficients. If teacher assignments are uncorrelated with student ability (implying $\lambda_h = 0$) or if ability does not enter into the grade-g gain equation ($\widetilde{\tau}^g = \tau^g - \tau^{g-1} = 0$) then π_h^g should be identically zero for all h>g. Even if teacher assignments do depend on student ability, the restriction that π_h^g must be a scalar multiple of λ_h in each grade h > g permits a specification test. Specifically, if we observe gains in two grades, g and k, and teacher assignments in some later grade h>max(g,k), (17) implies that $\pi_h^g = \lambda_h \widetilde{\tau}^g$ and $\pi_h^k = \lambda_h \widetilde{\tau}^k$. Without loss of generality, we can normalize $\widetilde{\tau}^k = 1$. We then have $\pi_h^g = \pi_h^k * \widetilde{\tau}^g$. Chamberlain proposes using optimal minimum distance (OMD; see also Abowd and Card, 1989) to identify the restricted coefficients and to form an overidentification test statistic. A rejection of the restriction indicates that the strict exogeneity assumption for ε is incorrect.

B. Implementation

As noted earlier, I work with a relatively homogenous sample of students who attend the same school in grades 3 through 5, never skip grades or are held back, and have complete data in grades 3, 4, 5, and 6.[29] I model both 4th and 5th grade gains, including as predictor variables indicators for the school that the student attended in grades 3-5 and indicators for the teachers that the student had in 3rd, 4th, 5th, and 6th grades:

[29] It is not essential to the Chamberlain model that the full sequence of teacher assignments be observed. The same strategy applies when μ is projected only onto teacher indicators in grades 3-6, so long as there are no effects of teachers before 3rd grade on the gain scores considered.

(19) $\Delta A_{i4} = T_{i3}\,\pi_3^{\,4} + T_{i4}\,\pi_4^{\,4} + T_{i5}\,\pi_5^{\,4} + T_{i6}\,\pi_6^{\,4} + e_{i4}$.

(20) $\Delta A_{i5} = T_{i3}\,\pi_3^{\,5} + T_{i4}\,\pi_4^{\,5} + T_{i5}\,\pi_5^{\,5} + T_{i6}\,\pi_6^{\,5} + e_{i5}$.

Estimation of (19) and (20) presents computational difficulties, as each contains over six thousand dummy variables in five different sets. The approach that I take is described in detail in the Appendix. The homogeneity of my sample greatly reduces the computational burden: Because I exclude students who change schools and include only covariates that are measured within schools,[30] (19) and (20) can be estimated via a sequence of within-school regressions, each including only the indicators for teachers at the school in question. Each is estimated by system OLS, with standard errors that are robust to arbitrary heteroskedasticity and within-student, across-grade serial correlation. I then normalize the π_h^g coefficients to have mean zero across all grade-h teachers at the same school.

Recall from (17) that the π_h^g coefficients for current and past teachers (h<g) include causal effects ($\tilde{\beta}_h^g$) as well as the student sorting parameters $\lambda_h\,\tilde{\tau}^g$, while the coefficients for future teachers include only the sorting parameters. Intuitively, we might expect the former coefficients to be larger (in a variance sense) than are the latter. Table 5 presents the unadjusted and adjusted standard deviations of the π coefficients from (19) and (20). Diagonal elements – $\pi_4^{\,4}$ and $\pi_5^{\,5}$ – are shaded for emphasis. We see that each of the π_h^g coefficient vectors has substantial variation. Standard deviations are larger for current and past teachers than for future teachers, but not by much.

Table 6 presents goodness-of-fit statistics, for the full model and for restricted models that exclude one or more sets of teacher effects. The first column shows the number of model degrees of freedom, including 457 school effects and all teacher effects that can be separately identified. The second column shows the model R^2. The model with just school effects explains 0.06 (0.05) of the

[30] I allow a grade-6 teacher with students from several elementary schools to have separate "effects" on each group, in effect fully interacting T_{i6} with elementary school indicators. This allows tracking to depend on the base group – a teacher might get the best students from a low-achieving school and the worst students from a high-achieving school.

variance of grade-4 (5) gains. Adding indicators for the contemporaneous teacher raises this to 0.14 (0.13). Adding the remaining three sets of teacher indicators raises the R^2 to 0.33 in each grade.

Columns 3 and 4 report heteroskedasticity-robust Wald tests of restricted models that leave out one or more sets of teachers. These have χ^2 distributions under the null hypothesis. Every restricted model is rejected at the 0.001 level. Recall that if the strict exogeneity assumption holds, the coefficients for future teachers reflect only the degree to which students are sorted on the basis of their μ parameters. Thus, if there is no tracking these coefficients should equal zero. The clear rejection of this restriction, in the bottom row of each panel, indicates that there is indeed sorting. This rules out the basic VAM, as well as other specifications (including random effects models like that used in TVAAS) that rely on the same identifying assumptions.

Columns 5-7 report three fit statistics that attempt to indicate whether the additional covariates included in richer models have enough explanatory power to justify their inclusion. Column 5 shows $\overline{R}^2$. With one exception (the exclusion of 6th grade teachers from the model for 4th grade gains), $\overline{R}^2$ is maximized in the most saturated model that includes all four sets of teacher indicators. Columns 6 and 7 show the logs of the Akaike and Schwartz Information criteria, respectively. These criteria penalize saturated models much more than does $\overline{R}^2$, and in each case lower values correspond to better fit. Both indicate that the saturated model is uniformly worse than any of the restricted models. This is not particularly surprising, as the saturated model contains a large ratio of regressors (5,799) to observations (21,101). However, its poor performance on these criteria does not offer support for the more traditional VAM specification: Each criterion prefers the specification with only school effects to one including contemporaneous teacher effects as well.

Table 7 presents the correlations between grade-h teachers' estimated effects on grade-4 and grade-5 gains, corr(π_h^4, π_h^5), weighted by the number of students exposed to each teacher. To the extent that the λ_h coefficients – the factor loading of individual ability μ_i on the grade-h teacher assignment – are an important part of the variation in π_h^g, one would expect these correlations to be positive. In fact, the correlation is around -0.5 for grade-4 teachers (and -0.6 when adjusted for sampling error) and only slightly less negative for teachers in other grades.

To be sure, one can construct a story in which grade-4 teachers' causal effects on 4th and 5th grade gains are negatively correlated. Perhaps some teachers achieve high gain scores by cramming and coaching for the test, enabling their students to do well on the end-of-grade test with little effect on future performance. If indeed the grade-4 teacher can affect only A_{i4} and not A_{i5}, the lagged effect $\tilde{\beta}_4^5$ should solely reflect mean reversion and should be perfectly negatively correlated with $\tilde{\beta}_4^4$. Alternatively, if teachers' effects on grade-4 and grade-5 scores are uncorrelated and of equal magnitude, one should observe $\operatorname{corr}\left(\tilde{\beta}_4^4, \tilde{\beta}_4^5\right) = -0.5$. One can tell a similar story about why $\operatorname{corr}\left(\tilde{\beta}_3^4, \tilde{\beta}_3^5\right)$ might be negative.

It is more difficult to account for the negative correlation between grade-6 teachers' apparent "effects" on 4th and 5th grade gain scores. Indeed, this correlation indicates a rejection of strict exogeneity. Recall that $\pi_h^g = \left(\tilde{\tau}^g / \tilde{\tau}^k\right) \pi_h^k$ (up to sampling error) for each g, k < h. As $\tilde{\tau}^g / \tilde{\tau}^k$ is a scalar, $\operatorname{corr}(\pi_6^4, \pi_6^5)$ should equal one (again, up to sampling error) if $\tilde{\tau}^g / \tilde{\tau}^k$ is positive and minus one if it is negative.

I implement the formal test via optimal minimum distance (OMD), minimizing

(21) $$D \equiv ((\pi_6^4 \;\; \pi_6^5) - (\tilde{\tau}^4 \lambda_6 \;\; \tilde{\tau}^5 \lambda_6))\, W^{-1} \, ((\pi_6^4 \;\; \pi_6^5) - (\tilde{\tau}^4 \lambda_6 \;\; \tilde{\tau}^5 \lambda_6))'$$

over the scalars $\tilde{\tau}^4$ and $\tilde{\tau}^5$ and vector λ_6.[31] W is the cluster-robust variance matrix for (π_6^4, π_6^5).

In the full sample, π_6^g has 1,184 independent coefficients, yielding 1,183 degrees of freedom in D.[32] The first panel of Table 8 reports the (unweighted) standard deviations of the π coefficients

[31] These are identified only up to a scale parameter: Multiplying λ_6 by any non-zero constant and dividing $\tilde{\tau}^4$ and $\tilde{\tau}^5$ by the same constant has no effect on D. The composites $\tilde{\tau}^g \lambda_6$ and $\tilde{\tau}^5 / \tilde{\tau}^4$ are uniquely identified, however.

for the effects of grade-6 teachers on 4th and 5th grade gains, both unadjusted and adjusted for sampling error. The second panel reports the standard deviations of the fitted values from the restricted model, SD($\gamma_4\lambda_6$) and SD($\gamma_5\lambda_6$). The restricted parameterization is unable to fit the OLS coefficients well, and the best fit is obtained by aligning closely with the unrestricted estimates for π_6^5 (grade-6 teachers for grade-5 gains) and setting $\tilde{\tau}^4$ to near zero.

The next row shows the value of the objective, D, at the OMD estimates. Under the null hypothesis of strict exogeneity, this is distributed χ^2 with 1,183 degrees of freedom. This null is decisively rejected. We can thus conclude that teacher assignments are not strictly exogenous, even conditional on a fixed individual effect.

Given the importance of this result – it implies that VAMs which absorb student heterogeneity through fixed effects are misspecified – in Table 9 I explore its sensitivity to several alternative specifications. The first row repeats the estimates from Table 8. Row 2 uses scale scores that have been standardized separately in each grade. Row 3 uses gain scores that are standardized separately for each initial level, to allow for the possibility that the test may exhibit mean reversion or other sensitivities to the location in the score scale. Specifically, the dependent variable for the models summarized in this row is $\left(A_{ig} - E[A_{ig} \mid A_{ig-1}]\right) / \sqrt{V[A_{ig} \mid A_{ig-1}]}$. Hanushek, et al. (2005) and Jacob and Lefgren (2007) use standardizations of this form. Row 4 uses percentile scores in place of scale scores for all computations. In each case, I reject the restriction that the grade-6 teacher effects on grade-4 and grade-5 gains are proportional.

All of the above specifications exclude the score in year g-1 from the equation for the grade-g gain score. The last row of Table 9 presents a test based on a model that includes the once-lagged

[32] The complete list of grade-6 teachers at each school is perfectly collinear, implying a singular W. I drop the teacher at each school with the largest number of students, and use only the remaining π coefficients (relative to that teacher's π). This leaves 1,184 linearly independent coefficients for each grade.

score as well as all lagged teachers as explanatory variables. This takes a different form. Suppose that the causal model for grade-g achievement is:[33]

(22) $A_{ig} = T_{i1}\,\beta_1^{\,g} + \ldots + T_{ig}\,\beta_g^{\,g} + A_{ig-1}\,\chi + \mu_i + \eta_{ig}.$

Differencing from this the corresponding equation for the g-1 score, we get

(23) $\Delta A_{ig} = T_{i1}\,(\beta_1^{\,g} - \beta_1^{\,g-1}) + \ldots + T_{ig-1}\,(\beta_{g-1}^{\,g} - \beta_{g-1}^{\,g-1}) + T_{ig}\,\beta_g^{\,g} + \Delta A_{ig-1}\,\chi + \Delta\eta_{ig}.$

This eliminates the ability term μ_i, but even with strict exogeneity of teacher assignments the lagged gain score ΔA_{ig-1} is endogenous to $\Delta\eta_{ig}$. A_{ig-2} will serve as an instrument (Anderson and Hsiao, 1981, 1982). Strict exogeneity can be tested by including T_{ig+1} as an additional regressor in (23); if the assumption holds, the coefficients on future teachers should be zero. The last row of Table 9 presents the test of this restriction (using g=5). It is rejected decisively.

VI. Teacher Characteristics

Many value added studies contrast the large apparent effects of teachers indicated by VAMs like those discussed here with the small estimated effects of teachers' observed characteristics. The latter derive from simple regression specifications that replace teacher indicators with observed teacher characteristics. In light of the results thus far, it is reasonable to ask whether this sort of specification can identify the causal effects of teacher characteristics on student gains. I again investigate this by asking whether *future* teachers' characteristics predict current gains.

Table 10 presents the results. 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

[33] As discussed in Section IIC, including the lagged score adds no generality so long as the specification allows a flexible decay structure for the effects of lagged inputs. However, as many authors focus on models with lagged test scores, this is presented for completeness.

required to obtain certification in North Carolina.[34] My sample consists of students with complete records in grades 3-6 who attended the same school in grades 4 and 5 and had valid teacher matches in grades 3-5. I further discard students for whom I am unable to assemble complete characteristics for each of the teachers in grades 3-6, as well as those attending schools where fewer than 10 students meet the other criteria.

Columns 1 and 3 present basic estimates of the effects of 4th and 5th grade teachers on 4th and 5th grade gain scores, respectively. Each specification includes school fixed effects and in each case standard errors are clustered on the school.[35] Results echo those in the literature: A master's degree appears to make little difference, but teacher experience has an effect on student test scores (e.g., Clotfelter, et al., 2006, 2007; Goldhaber and Brewer, 1997; Hanushek and Rivkin, 2006).

Columns 2 and 4 generalize these specifications by adding controls for past and future teachers' characteristics. Several characteristics of past and future teachers have significant effects, and there is no indication that current teachers' characteristics are better predictors than are those of past and future teachers. The bottom rows of the table present hypothesis tests on several combinations of the coefficients. I cannot reject zero effects of 3rd grade teachers' characteristics on 4th grade gains, but the other tests indicate that the effects of past and future teachers' characteristics are significantly different from zero (at the 10 percent level).

Column 5 reports the sum of teachers' effects on grade-4 and grade-5 gains, as well as composite tests. None of the teacher characteristics have significant effects on cumulative gains. I reject the hypotheses that past teachers' effects are zero for both 4th and 5th grade gains (that is,

[34] I use the tests required for elementary certification for teachers in grades 3 through 5. For grade 6 teachers, I use middle-school-certification tests if they are available, or all available tests if not. Each test is standardized among North Carolina teachers who took them in the same year, then scores are averaged across tests (when multiple scores are used)

[35] This clustering is made possible by the shift from teacher fixed effects to a limited number of teacher characteristics (Kezdi, 2004). The clustered standard errors are robust to classroom-level error components. As it happens, clustering makes little difference to the results, suggesting that the earlier results were not much biased by the failure to allow for classroom-level errors in inference.

grade-3 teachers in column 2 and grade-3 and -4 teachers in column 4); that future teachers' effects are zero; and that all past and future teachers' effects are zero. I also test and (marginally) reject the hypothesis that the effects of grade-6 teachers' characteristics on grade-5 gains are a constant multiple of those on grade-4 gains, as in the correlated random effects model. The dynamic tracking found earlier evidently applies to teacher characteristics as well.

VII. Toward Identification

I have established thus far that teachers have apparent "effects" on students' *prior* achievement, and that these effects are both highly statistically significant and approximately as important as those on *current* gain scores. It is apparent that students are not even approximately randomly assigned to teachers. I have also investigated and rejected a leading explanation for this result, that students are sorted across teachers on the basis of a permanent component of achievement but that assignments are random conditional on this. Rather, it seems that teacher assignments respond dynamically to transitory shocks to student achievement.

Traditional VAMs do not identify teachers' causal effects in the presence of dynamic tracking. Non-experimental identification will require richer models that explicitly account for this tracking. In this Section, I describe two models that may permit identification, under differing assumptions about the information used in teacher assignments. Each model requires richer data than are available. I present approximations to each that are feasible given the available data. These cannot be treated as causal. They nevertheless suggest factors that should be the focus of future work, and offer an indication of the degree of bias in estimates based on the VAMs in common use.

Both of my models assume that the underlying rate of learning is homogenous across students (i.e. that τ^g is constant across g in equation (2)). The negligible student-level correlation between grade-4 and grade-6 gains suggests that this is plausible.

The models differ in their treatment of the source of the relationship between the transitory error, ε, and teacher assignments. The first model makes three assumptions: (i) there is no decay in the effect of non-teacher inputs, so the error term in equation (4) is simply $\varepsilon_{ig}\ \varphi_g^g$; (ii) this error is composed of a signal component and a noise component, $\varepsilon_{ig} = \varepsilon_{ig}^* + e_{ig}$, and the signal component is serially uncorrelated; (iii) teacher assignments depend in part on past realizations of the signal, ε_{i1}^*, $\ldots$, ε_{ig-1}^*, but not, conditional on this, on the noise component. Assumption (ii) attributes all serial correlation in observed gains to measurement error in the annual tests.[36] (iii) might hold if the test scores themselves were unavailable for use in tracking.

Under assumptions (i) through (iii), the non-random assignment of students to teachers can account for the correlation between gain scores and future teacher assignments that was seen earlier, but is ignorable for the effects of current and previous teachers. Causal effects can be identified from simple regressions of grade-g gains on teacher assignments in grades 1 through g.

Table 11 presents a comparison between estimates based on this model and the basic VAM, (14), for the effects of 4th grade teachers. The important distinction between the two – assuming that assumptions (i) through (iii) are satisfied – is the need to estimate lagged effects. If indeed inputs decay, the net effect of the 4th grade teacher is not simply the effect on 4th grade gains, β_4^4, but the cumulative effect on gains in all grades from 4 onward, $\beta_4^4 + \beta_4^5 + \ldots + \beta_4^G$. The available data do not permit the model to be extended beyond grade 5. I therefore truncate this series after two terms, focusing on $\beta_4^4 + \beta_4^5$. Similarly, one should ideally control for all previous teachers; I control only for those in grade 3 and afterward.

Column 1 of Table 11 presents the basic VAM, relating grade-4 gains to grade-4 teachers. Column 2 presents an augmented VAM with controls for 3rd grade teachers. Their effects are

[36] This is inconsistent with the calculations in Section IIIA, which indicated that corr(ε_{i4}^*, ε_{i5}^*) ≈ -0.25. But these were based on a known reliability of 0.86. If the test's reliability is in fact a bit lower, assumption (ii) could hold. The published reliability is based on a sample of only 70 students (Sanford, 1996), so should not be taken as precise.

approximately as large as those of 4th grade teachers. Column 3 presents the corresponding model for grade-5 gains, with controls for teachers in grades 3-5. Both 3rd and 4th grade teachers continue to have important effects in 5th grade. Column 4 presents teachers' cumulative effects, $\beta_h^4 + \beta_h^5$. For grade 4 teachers, these are slightly *smaller* than the contemporaneous effects. The explanation is in the lower portion of the table, showing the correlation between the T_{i4} effects from the various specifications. The estimates from the basic VAM (column 1) are almost perfectly correlated with those from the augmented model for grade-4 gains (column 2), indicating that omitted variables bias in the former is trivial. But effects on grade-5 gains are strongly negatively correlated with those on grade-4 gains. Thus, the correlation between contemporaneous and cumulative (over two grades) effects is only about 0.5.

My second proposed model is more realistic for the North Carolina data, where test scores appear to be available for use in teacher assignments. It replaces assumptions (i) - (iii) with a single assumption: $E[\varepsilon_{ig} \mid T_{i1}, \ldots, T_{ig}, \varepsilon_{i1}, \ldots, \varepsilon_{ig-1}] = E[\varepsilon_{ig} \mid \varepsilon_{i1}, \ldots, \varepsilon_{ig-1}]$. This would be appropriate if lagged test scores were the *only* information used in forming teacher assignments; if so, all of the information about future errors that is encoded in the teacher assignment sequence is also available from the error history.[37] Thus, the endogeneity of teacher assignments can be absorbed via controls for the full history of ε. Alternatively, one can include controls for all lags of A and T (though in this case the lagged T coefficients are not directly interpretable as estimates of β_h^g, which must be solved for recursively).

Table 12 presents a comparison between this model and the basic VAM for estimation of the contemporaneous effect of 5th grade teachers on 5th grade gains. Column 1 presents the basic VAM. Column 2 presents a specification that is augmented with controls for teachers in grades 3

[37] This implies that T_{i1} is randomly assigned, as there is no prior achievement history on which it can be based. This initial condition is required for the identification of lagged effects. Without it, only contemporaneous effects – the β_g^g parameters in (4), for g>1 – are identified.

and 4 (as in column 3 of Table 11).[38] Column 3 further adds controls for the full available history of test scores. Three scores are available: The end-of-grade tests in grades 3 and 4, plus a pre-test given at the beginning of grade 3. All three are highly significantly related to the grade-5 gain.[39] Their inclusion changes the estimated grade-5 teacher effects: The correlation of the β_5^5 coefficients from the augmented specification and those from the VAM with lagged teachers but no achievement history controls is only 0.82.

Neither of these models is perfectly realistic for North Carolina, where test scores might be used in teacher assignments but where – not least because there are no end-of-grade tests before 3rd grade – they are unlikely to be the only information used. However, the analyses in Tables 11 and 12 suggest that both heterogeneous rates of decay and the non-random assignment of students to teachers are important factors, and that simple VAMs which fail to account for them are not likely to be very informative about teachers' true value added.

VIII. Discussion

In the absence of random assignment, researchers often assume without evidence that the inclusion of large numbers of fixed effects will permit unbiased estimation of causal effects. My analysis indicates, at least in the case of teachers, that this assumption is unwarranted. Teachers are not as good as randomly assigned.

The results presented here invalidate many of the strategies that have been used to estimate teachers' effects. In particular, there is no basis for the continued inclusion of student fixed effects in value added models, as the assumptions that are required for this are clearly falsified by the data. It remains possible that the assumptions needed for identification without student fixed effects are

[38] This is an identical specification to that in column 3 of Table 11. The samples differ somewhat, as the sample used in Table 12 conditions on a complete test score history.

[39] The significance of the grade-3 scores indicates that assumption (ii) of the first model must be incorrect, as measurement error in grade-5 gains should be uncorrelated with grade-3 scores.

satisfied, though even here there must be changes in the methods used: There is clear evidence for heterogeneity in teachers' lagged effects on achievement gains in later grades, and a teacher's short-run effect appears to be a poor proxy for her cumulative effect.

It must also be emphasized that even when value added models incorporate sufficient flexibility to capture lagged effects, they rely on strong, unverified assumptions about the teacher assignment process. The VAMs in common use depend on incorrect assumptions, and richer models that are not falsified by the data yield notably different estimates of teachers' effects. Causal inference from observational data on student tests and teacher assignments calls for a great deal of caution and more attention to the plausibility of the identifying assumptions.

Value added estimates should be validated before being pressed into service in accountability and compensation policy. An obvious first step is to compare non-experimental estimates of individual teachers' effects in random assignment experiments with those based on pre- or post-experimental data (as in Cantrell, Fullerton, et al., 2007). But experiments are unlikely to resolve the issue. Experimental samples are typically small and unrepresentative, while most value added applications would apply to most or all teachers. As Figure 1 indicates, value added analyses are driven by a minority of schools; an experiment that excludes these schools is unlikely to resolve the important questions about universal value added systems. More attention should be paid as well to non-experimental evaluations, both validation studies[40] and falsification exercises like this one.

The questions investigated and methods used here have application beyond the estimation of teacher value added. Within education, similar observational estimates are used to measure the quality of schools (Kane and Staiger, 2002; Ladd and Walsh, 2002). Outside of education, models of firm and industry wage effects that include worker fixed effects (Abowd, et al., 1999) are structurally

[40] Jacob and Lefgren (2005) and Harris and Sass (2007) show that VAM estimates are correlated with principals' ratings of teacher performance. More work along these lines is needed.

similar to the student fixed effects VAM, and rely on similar (equally implausible) assumptions. Evidence about the "effects" of future schools and employers on current outcomes would be informative about the validity of both strategies.

References

Aaronson, Daniel; Lisa Barrow and William Sander (2007). "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 24(1), 95-135.

Abowd, John M. and David Card (1989). "On the Covariance Structure of Earnings and Hours Changes." *Econometrica* 57(2), March, 411-45.

Abowd, John M. and Francis Kramarz (1999). "The Analysis of Labor Markets Using Matched Employer-Employee Data." In O. C. Ashenfelter and D. Card, ed., *Handbook of Labor Economics, Volume 3b.* Amsterdam: North-Holland, 2629-710.

Abowd, John M.; Francis Kramarz and David N. Margolis (1999). "High Wage Workers and High Wage Firms." *Econometrica* 67(2), March, 251-333.

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

____ (1982). "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics* 18(1), January, 47-82.

Arellano, Manuel (2001). "Panel Data Models: Some Recent Developments." In J. J. Heckman and E. Leamer, ed., *Handbook of Econometrics.* Elsevier Science B.V, 3229-96.

Arellano, Manuel and Stephen Bond (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58(2), April, 277-97.

Ashenfelter, Orley C. and David Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *The Review of Economics and Statistics* 67(4), November, 648-60.

Ballou, Dale (2002). "Sizing up Test Scores." *Education Next* 2(2), Summer, 10-15.

Ballou, Dale; William Sanders and Paul Wright (2004). "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Educational and Behavioral Statistics* 29(1), Spring, 37-65.

Boardman, Anthony E. and Richard J. Murnane (1979). "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement." *Sociology of Education* 52(2), April, 113-21.

Bock, R. Darrell; Richard Wolfe and Thomas H. Fisher (1996). "A Review and Analysis of the Tennessee Value-Added Assessment System." Nashville, Tennessee: Comptroller of the Treasury, Office of Education Accountability.

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

Braun, Henry I. (2005a). "Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models," *Policy Information Perspective* ETS Policy Information Center.

____ (2005b). "Value-Added Modeling: What Does Due Diligence Require?" In R. W. Lissitz, ed., *Value Added Models in Education: Theory and Applications.* Maple Grove, Minn.: JAM Press, 19-39.

Cantrell, Steven; Jon Fullerton; Thomas J. Kane and Douglas O. Staiger (2007). "National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment." Unpublished manuscript, May 25.

Chamberlain, Gary (1982). "Multivariate Regression Models for Panel Data." *Journal of Econometrics* 18, 5-46.

____ (1984). "Panel Data." In Z. Griliches and M. D. Intriligator, ed., *Handbook of Econometrics.* Amsterdam: Elsevier North-Holland, 1248-318.

Clotfelter, Charles T.; Helen F. Ladd and Jacob L. Vigdor (2006). "Teacher-Student Matching and the Assessment of Teacher Effectiveness."*NBER Working Paper* 11936, January

____ (2007). "How and Why Do Teacher Credentials Matter for Student Achievement?"*NBER Working Paper* 12828, January.

Dee, Thomas S. and Benjamin J. Keys (2004). "Does Merit Pay Reward Good Teachers? Evidence from a Randomized Experiment." *Journal of Policy Analysis and Management* 23(3), 471-88.

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

Goldhaber, Dan and Dominic J. Brewer (1997). "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity." *Journal of Human Resources* 32(3), Summer, 505-23.

Gordon, Robert; Thomas J. Kane and Douglas O. Staiger (2006). "Identifying Effective Teachers Using Performance on the Job," *White Paper* 2006-01. Washington, D.C.: The Hamilton Project, 1-35.

Hanushek, Eric A. (1979). "Conceptual and Empirical Issues in the Estimation of Educational Production Functions." *Journal of Human Resources* 14(3), Summer, 351-88.

____ (1981). "Throwing Money at Schools." *Journal of Policy Analysis and Management* 1(1), Fall, 19-41.

____ (1986). "The Economics of Schooling: Production and Efficiency in Public Schools." *Journal of Economic Literature* 49(3), September, 1141-77.

____ (2002). "Teacher Quality." In L. T. Izumi and E. M. Williamson, ed., *Teacher Quality.* Hoover Press,

Hanushek, Eric A.; John F. Kain; Daniel M. O'Brien and Steven G. Rivkin (2005). "The Market for Teacher Quality."*NBER Working Paper* 11154, February.

Hanushek, Eric A. and Steven G. Rivkin (2006). "Teacher Quality." In E. A. Hanushek and F. Welch, ed., *Handbook of the Economics of Education.* Amsterdam: Elsevier North-Holland, 2-28.

Harris, Douglas N. and Tim R. Sass (2006). "Value-Added Models and the Measurement of Teacher Quality." Preliminary Draft,*Unpublished manuscript, Florida State University*, April

____ (2007). "What Makes for a Good Teacher and Who Can Tell?"*Unpublished manuscript, Florida State University*, July.

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

Jacob, Brian A. and Lars Lefgren (2005). "Principals as Agents: Subjective Performance Measurement in Education."*NBER Working Paper 11463*, June

____ (2007). "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education."*NBER Working Paper* June

Kane, Thomas J.; Jonah E. Rockoff and Douglas O. Staiger (2006). "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City."*NBER Working Paper 12155*, March

Kane, Thomas J. and Douglas O. Staiger (2001). "Improving School Accountability Measures." Working Paper,*NBER Working Paper* 8156.

____ (2002). "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives* 16(4), Fall, 91-114.

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

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

Koppich, Julia E. (2005). "All Teachers Are Not the Same: A Multiple Approach to Teacher Compensation." *Education Next* 2005(1), Winter, 13-15.

Kramarz, Francis; Stephen Machin and Amine Ouazad (2007). "What Makes a Grade? The Respective Contributions of Pupils, Schools and Peers in Achievement in English Primary Education."*Unpublished manuscript*, April

Ladd, Helen F. and Randall P. Walsh (2002). "Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right." *Economics of Education Review* 21, 1-17.

McCaffrey, Daniel F.; J. R. Lockwood; Daniel M. Koretz and Laura S. Hamilton (2003). "Evaluating Value-Added Models for Teacher Accountability." Santa Monica, CA: RAND.

Nye, Barbara; Spyros Konstantopoulos and Larry V. Hedges (2004). "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis* 26(3), Fall, 237-57.

Peterson, Paul E. (2006). "Of Teacher Shortages and Quality." *Education Next* 6(2), Spring, 5.

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

Rockoff, Jonah E. (2004). "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 92(2), May, 247-52.

Rogosa, David R. (1995). "Myths and Methods: 'Myths About Longitudinal Research,' Plus Supplemental Questions." In J. M. Gottman, ed., *The Analysis of Change.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, 3-65.

Sanders, William L. and Sandra P. Horn (1994). "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education* 8, 299-311.

____ (1998). "Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research." *Journal of Personnel Evaluation in Education* 12(3), 247-56.

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

Sanders, William L.; Arnold M. Saxton and Sandra P. Horn (1997). "The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment." In J. Millman, ed., *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure.* Thousand Oaks, CA: Corwin, 137-62.

Sanford, Eleanor E. (1996). "North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics," *Technical Report* #1. Division of Accountability/Testing, Office of Instructional and Accountability Services, North Carolina Department of Public Instruction.

Stake, Robert E. (1971). "Testing Hazards in Performance Contracting." *Phi Delta Kappan* 52(10), June, 583-89.

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

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

Yen, Wendy (1986). "The Choice of Scale for Educational Measurement: An Irt Perspective." *Journal of Educational Measurement* 23(4), Winter, 299-325.

Figure 1. Distribution of p values from school-level tests of hypothesis that all teachers at school have identical effects.

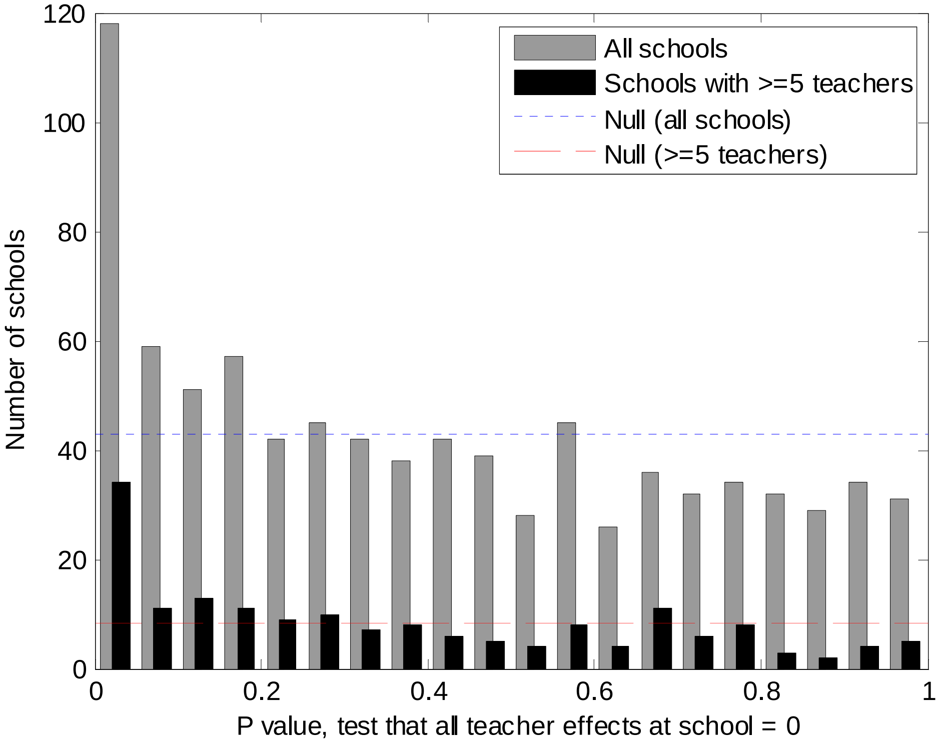

Table 1. Across-grade correlations of reading test scores and test score gains

		Achievement levels				Gain scores			
		Grade 3	Grade 4	Grade 5	Grade 6	Grade 4	Grade 5	Grade 6	Cumulative
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
N		96,678	96,243	97,199	98,080	88,982	89,352	89,945	80,724
Mean		0.00	0.34	0.98	1.15	0.34	0.63	0.17	1.13
SD		1.00	0.98	0.84	0.94	0.60	0.56	0.53	0.65
Correlations									
Achievement levels	Grade 3	1							
	Grade 4	0.81	1						
	Grade 5	0.77	0.82	1					
	Grade 6	0.76	0.80	0.82	1				
Gain scores	Grade 4	-0.29	0.29	0.06	0.05	1			
	Grade 5	-0.24	-0.51	0.07	-0.19	-0.43	1		
	Grade 6	0.15	0.16	-0.11	0.47	0.01	-0.42	1	
	Cumulative	-0.38	-0.04	0.02	0.27	0.57	0.12	0.46	1

Table 2. Summary Statistics

	Population		Sample 1		Sample 2	
	(1)	(2)	(3)	(4)	(5)	(6)
	Students	**Schools**				
Total	129,665	1,316	65,582	860	21,101	457
Count by # of 5th grade teachers represented in the sample						
1 5th grade teacher	1,947	117	0	0	0	0
2 5th grade teachers	4,240	128	8,556	202	1,527	59
3-5 5th grade teachers	44,026	608	48,425	600	16,118	353
>5 5th grade teachers	79,452	463	8,601	58	3,456	45
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Female	48%		50%		51%	
Black	30%		28%		18%	
Hispanic	5%		4%		2%	
Other non-white	5%		4%		3%	
Consistent student record	98%		100%		100%	
Data available for						
Grade 3	78%		87%		100%	
Grade 4	78%		92%		100%	
Grade 5	79%		99%		100%	
Grade 6	80%		93%		100%	
Changed schools in						
Grade 4	26%		21%		0%	
Grade 5	23%		19%		0%	
Grade 6	80%		95%		93%	
"Valid" teacher in						
Grade 3	78%		84%		86%	
Grade 4	77%		87%		100%	
Grade 5	73%		100%		100%	
Grade 6	0%		0%		0%	
Reading test score						
Grade 3	0.00	1.00	0.04	0.98	0.25	0.90
Grade 4	0.34	0.98	0.39	0.97	0.59	0.89
Grade 5	0.98	0.84	1.01	0.83	1.18	0.76
Grade 6	1.15	0.94	1.19	0.92	1.38	0.86
Gain score						
Grade 4	0.34	0.60	0.33	0.58	0.34	0.56
Grade 5	0.63	0.56	0.62	0.55	0.59	0.53
Grade 6	0.17	0.53	0.16	0.53	0.19	0.52
Cumulative	1.13	0.65	1.11	0.63	1.12	0.62

Table 3. Simple models for 5th grade teachers' value added

	Dependent variable			
	5th grade gain score	4th grade gain score	6th grade gain score	Cumulative gain, 4th - 6th grades
	(1)	(2)	(3)	(4)
Fit statistics				
N	59,104	54,377	59,535	51,275
# of teachers	3,013	3,013	3,013	3,013
# of schools	860	860	860	860
R2	0.098	0.085	0.095	0.096
Adjusted R2	0.050	0.031	0.047	0.039
Just school effects: R2	0.048	0.044	0.052	0.053
Just school effects: Adj. R2	0.034	0.028	0.038	0.037
Test, teacher effects = 0				
Test statistic	2,953	2,318	2,650	2,274
DF	2,153	2,153	2,153	2,153
5% critical value	2,262	2,262	2,262	2,262
p value	<0.001	0.007	<0.001	0.035
Standard deviation of teacher effects				
Unadjusted	0.145	0.140	0.130	0.154
Adjusted	0.103	0.081	0.085	0.087

Notes: Standard deviations of teacher effects are weighted by the number of students assigned to each teacher, with degrees of freedom adjustments to account for the adjustment of teacher effects to have mean zero at each school. Adjusted variances are computed by subtracting the weighted mean heteroskedasticity-robust sampling variance.

Table 4. Across-grade correlations of 5th grade teacher effects, simple models

	Dependent variable			
	4th grade gain score	5th grade gain score	6th grade gain score	Cumulative gain, 4th - 6th grades
	(1)	(2)	(3)	(4)
Unadjusted				
4th grade gain score	1	-0.41	0.04	0.56
5th grade gain score	-0.41	1	-0.54	0.12
6th grade gain score	0.04	-0.54	1	0.40
Cumulative gain, 4th - 6th grades	0.56	0.12	0.40	1
Adjusted for sampling covariances				
4th grade gain score	1	-0.38	0.06	0.56
5th grade gain score	-0.38	1	-0.69	0.15
6th grade gain score	0.06	-0.69	1	0.28
Cumulative gain, 4th - 6th grades	0.56	0.15	0.28	1

Notes: Each correlation is computed from specifications like those in Table 3, but limited to students with data on both dependent variables. Correlations are weighted by the number of such students in this subsample. Sampling covariances are estimated allowing for heteroskedasticity and arbitrary clustering within students across grades.

Table 5. Standard deviations of teacher effects in models with controls for past and future teachers

	Model for 4th grade gain score		Model for 5th grade gain score	
	Unadjusted	Adjusted	Unadjusted	Adjusted
	(1)	(2)	(3)	(4)
3rd grade teacher	0.21	0.14	0.18	0.10
4th grade teacher	0.22	0.15	0.20	0.13
5th grade teacher	0.20	0.12	0.20	0.13
6th grade teacher	0.17	0.10	0.17	0.11

Notes: Statistics are computed from separate specifications for 4th and 5th grade gain scores, each including school fixed effects and fixed effects for 3rd, 4th, 5th, and 6th grade teachers. Across-teacher variances are weighted by the number of students taught. Adjusted variances are computed by subtracting the weighted average of the heteroskedasticity-robust sampling variance of the teachers' effects. Estimates corresponding to the contemporaneous teacher are shaded.

Table 6. Models with controls for all past and future teachers

			Robust Wald test (relative to full model)		More conservative criteria		
	Model DF	R2	DF	Test stat	Adj. R2	Akaike Info.	Schwartz Info.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable is 4th grade gain score							
Full model	5,799	0.33			0.082	-1.00	1.18
Restricted model							
Omitting one set of teacher effects at a time							
Excluding 3rd grade teachers	4,251	0.25	1,548	2,127	0.064	-1.04	0.57
Excluding 4th grade teachers	4,496	0.26	1,303	2,014	0.056	-1.02	0.68
Excluding 5th grade teachers	4,492	0.28	1,307	1,621	0.079	-1.04	0.65
Excluding 6th grade teachers	4,615	0.28	1,184	1,441	0.084	-1.04	0.70
Omitting several teacher effects together							
All teachers excluded	457	0.06	5,342	5,773	0.036	-1.16	-0.99
All but current teachers excluded	1,760	0.14	4,039	4,357	0.063	-1.13	-0.47
All past teachers excluded	4,251	0.25	1,548	2,127	0.064	-1.04	0.57
All future teachers excluded	3,308	0.22	2,491	2,822	0.081	-1.09	0.16
Dependent variable is 5th grade gain score							
Full model	5,799	0.33			0.082	-1.15	1.04
Restricted model							
Omitting one set of teacher effects at a time							
Excluding 3rd grade teachers	4,251	0.27	1,548	1,838	0.082	-1.20	0.41
Excluding 4th grade teachers	4,496	0.26	1,303	1,899	0.061	-1.16	0.53
Excluding 5th grade teachers	4,492	0.26	1,307	1,810	0.066	-1.17	0.52
Excluding 6th grade teachers	4,615	0.28	1,184	1,619	0.075	-1.18	0.57
Omitting several teacher effects together							
All teachers excluded	457	0.05	5,342	5,886	0.034	-1.30	-1.13
All but current teachers excluded	1,764	0.13	4,035	4,567	0.052	-1.26	-0.60
All past teachers excluded	2,948	0.19	2,851	3,440	0.060	-1.22	-0.11
All future teachers excluded	4,615	0.28	1,184	1,619	0.075	-1.18	0.57

Notes: N=21,101. Each model includes 457 effects for the school attended in grades 3-5; 6th grade teachers are interacted with indicators for these schools. The tests in columns 3-4 reject each restriction at the 0.001 level. Akaike (AIC) and Schwartz (BIC) statistics are presented in logs.

Table 7. Correlation between teacher effects on 4th and 5th grade gains, by teacher grade, full model

	Grade 3 teacher	Grade 4 teacher	Grade 5 teacher	Grade 6 teacher
	(1)	(2)	(3)	(4)
Unadjusted	-0.37	-0.52	-0.41	-0.42
Adjusted	-0.25	-0.61	-0.38	-0.40

Table 8. Optimal minimum distance estimates

	6th grade teacher effect on	
	4th grade gain	5th grade gain
	(1)	(2)
Unconstrained estimates		
SD, unadjusted	0.21	0.21
SD, adjusted	0.12	0.14
Constrained (OMD) estimates		
SD, unadjusted	0.02	0.18
SD, adjusted	0.01	0.10
Ratio, $\tilde{\tau}^{5}/\tilde{\tau}^{4}$	8.2	
Overidentification test		
Statistic (χ^2 under null)	1,607	
DF	1,183	
95% critical value	1,264	
P value	<0.0001	

Note: Standard deviations are not weighted by student enrollment.

Table 9. Tests for overidentifying restrictions, alternative specifications

		Objective function	DF	5% critical value	p
		(1)	(2)	(3)	(4)
1)	Base model	1,607	1,183	1,264	<0.001
2)	Gain in standardized scores	1,584	1,183	1,264	<0.001
3)	Standardized (by base level) gain	1,659	1,183	1,264	<0.001
4)	Gain in percentile scores	1,616	1,183	1,264	<0.001
5)	Control for lagged score (2SLS; Wald for $\beta_6^5=0$)	1,349	1,184	1,265	<0.001

Notes: Tests in rows 1-4 are of the hypothesis that the vector of teacher effects on 4th grade gains is a scalar multiple of the vector of effects on 5th grade gains. Each is computed as the objective function from OMD estimates, as reported (for the base model) in Table 8, and each is distributed chi2 under the null hypothesis. The test in row 5 is of the hypothesis that the vector of 6th grade teacher effects on 5th grade gains is identically zero; it, too, is chi2 under the null.

Table 10. Estimates of the effects of teacher characteristics on student gains

		Grade 4 gain*100		Grade 5 gain*100		Cumulative
		(1)	**(2)**	**(3)**	**(4)**	**(5)**
MA degree	Grade 3 teacher		1.64		-1.81	-0.17
			(1.31)		(1.14)	(1.82)
	Grade 4 teacher	0.88	0.82		-0.14	0.68
		(1.42)	(1.41)		(1.33)	(1.68)
	Grade 5 teacher		1.69	-1.35	-1.38	0.31
			(1.27)	(1.23)	(1.23)	(1.67)
	Grade 6 teacher		**2.51**		-0.20	2.31
			(1.16)		(1.07)	(1.72)
Experience	Grade 3 teacher		-0.12		0.00	-0.12
			(0.06)		(0.06)	(0.09)
	Grade 4 teacher	0.09	0.09		-0.02	0.07
		(0.07)	(0.07)		(0.07)	(0.09)
	Grade 5 teacher		-0.04	**0.13**	**0.13**	0.10
			(0.07)	(0.06)	(0.06)	(0.10)
	Grade 6 teacher		0.01		**-0.11**	-0.10
			(0.06)		(0.05)	(0.09)
Experience < 2	Grade 3 teacher		0.56		1.95	2.51
			(2.14)		(1.81)	(3.21)
	Grade 4 teacher	**-5.79**	**-5.86**		**6.66**	0.80
		(1.90)	(1.90)		(1.86)	(2.62)
	Grade 5 teacher		-1.30	-0.13	-0.32	-1.62
			(1.92)	(1.91)	(1.89)	(2.88)
	Grade 6 teacher		-0.32		1.19	0.86
			(1.73)		(1.68)	(3.21)
Praxis score	Grade 3 teacher		-0.60		0.15	-0.45
			(0.74)		(0.66)	(1.13)
	Grade 4 teacher	-0.19	-0.17		-0.22	-0.39
		(0.80)	(0.80)		(0.78)	(0.96)
	Grade 5 teacher		-1.01	0.83	0.92	-0.09
			(0.71)	(0.82)	(0.82)	(1.10)
	Grade 6 teacher		-1.28		0.22	-1.06
			(0.68)		(0.62)	(0.96)
R2		0.076	0.077	0.068	0.070	
p values for restrictions						**Joint tests**
All current teacher characteristics = 0		0.001	0.001	0.216	0.176	0.002
All prior teacher characteristics = 0			0.249		0.007	0.005
All future teacher characteristics = 0			0.087		0.062	0.033
All past and future teacher characteristics = 0			0.115		0.002	0.002
Grade-6 coefficients in (2) and (4) are proportional						0.056

Notes: Sample restricted to students for whom complete teacher data is available in grades 3-6 and who attended the same school in grades 4 and 5. N=16,386 from 632 schools. All specifications include fixed effects for the (grade 4 and 5) school, and standard errors are clustered on the school. Column 5 reports composite tests for the combined restrictions in columns 2 and 4, allowing for clustering across grades as well as within. The "scalar" hypothesis is that the ratio of the grade 6 teacher coefficients in column 4 to those in column 2 is constant; this test is based on the objective function in an optimal minimum distance estimator.

Table 11. Sensitivity of estimates of 4th grade teacher effect to allowing for lagged effects

	Sparse model for 4th grade gains	Model with lagged effects		
		4th grade gains	5th grade gains	Cumulative effect
	(1)	(2)	(3)	(4)
Standard deviation of teacher effects				
Unadjusted				
Grade 3 teacher		0.20	0.18	0.22
Grade 4 teacher	0.19	0.20	0.19	0.19
Grade 5 teacher			0.19	0.19
Adjusted				
Grade 3 teacher		0.13	0.09	0.14
Grade 4 teacher	0.13	0.14	0.13	0.11
Grade 5 teacher			0.12	0.12
Correlation of 4th grade teacher effect across specifications				
Unadjusted				
(1) Sparse model	1	0.94	-0.50	0.48
(2) 4th grade effect	0.94	1	-0.52	0.52
(3) 5th grade effect	-0.50	-0.52	1	0.46
(4) Cumulative effect	0.48	0.52	0.46	1
Adjusted				
(1) Sparse model	1	0.9996	-0.68	0.50
(2) 4th grade effect	1.00	1	-0.67	0.51
(3) 5th grade effect	-0.68	-0.67	1	0.39
(4) Cumulative effect	0.50	0.51	0.39	1

Notes: Sparse model includes only school effects and effects for the current teacher. Models in columns 2 and 3 add controls for the 3rd grade teacher and (in column 3) the 4th grade teacher. Cumulative effect is the sum of the effects in columns 2 and 3.

Table 12. Alternative specifications for 5th grade gain scores

	Basic model	Controls for lagged teachers	Controls for lagged teachers & scores
	(1)	(2)	(3)
Standard deviation of teacher effects			
Unadjusted			
Grade 3 teacher		0.21	0.14
Grade 4 teacher		0.21	0.14
Grade 5 teacher	0.17	0.19	0.19
Adjusted			
Grade 3 teacher		0.13	0.07
Grade 4 teacher		0.15	0.08
Grade 5 teacher	0.11	0.12	0.11
Coefficients on lagged scores			
Grade 3 score, beginning of year			0.077
			(0.006)
Grade 3 score			0.252
			(0.007)
Grade 4 score			-0.581
			(0.007)
Correlations of grade-5 teacher effects			
Unadjusted			
(1) Basic model	1	0.87	0.68
(2) Lagged teachers	0.87	1	0.80
(3) Lagged teachers & scores	0.68	0.80	1
Adjusted			
(1) Basic model	1	0.99	0.76
(2) Lagged teachers	0.99	1	0.82
(3) Lagged teachers & scores	0.76	0.82	1

Notes: All teacher effects are normalized to mean zero within each school. Basic model (column 1) includes only effects for the current teacher. Model in column 2 adds controls for grade-3 and grade-4 teachers. Column 3 also adds controls for achievement scores at the beginning of grade 3 and at the end of grades 3 and 4.