

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

JESSE ROTHSTEIN

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

## I. INTRODUCTION

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

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

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

© 2010 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology.

*The Quarterly Journal of Economics*, February 2010

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

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

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

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

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

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

My estimates also point to an important substantive result. To the extent that any of the VAMs that I consider identify causal effects, they indicate that teachers' long-run effects are at best weakly proxied by their immediate impacts. A teacher's effect in the year of exposure—the universal focus of value added analyses—is correlated only .3 to .5 with her cumulative effect over two years, and even less with her effect over three years. Accountability policies that rely on measures of short-term value added would do an extremely poor job of rewarding the teachers who are best for students' longer-run outcomes.

An important caveat to the empirical results is that they may be specific to North Carolina. Students in other states or in individual school districts might be assigned to classrooms in ways that satisfy the assumptions required for common VAMs. But at the least, VAM-style analyses should attempt to evaluate the model assumptions, perhaps with methods like those used here. Models that rely on incorrect assumptions are likely to yield misleading estimates, and policies that use these estimates in hiring, firing, and compensation decisions may reward and punish teachers for the students they are assigned as much as for their actual effectiveness in the classroom.

Section II reviews the use of preassignment variables to test exogeneity assumptions. Section III introduces the three VAMs,

discusses their implicit assumptions, and describes my proposed tests. Section IV describes the data. Section V presents results. Section VI attempts to quantify the biases that nonrandom classroom assignments produce in VAM-based analyses. Section VII presents evidence on teachers' long-run effects. I conclude, in Section VIII, by discussing some implications for the design of incentive pay systems in education.

## II. USING PANEL DATA TO TEST EXCLUSION RESTRICTIONS

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

This insight has been most fully explored in the literature on the effect of job training on wages and employment. Today's wage or employment status is quite informative about tomorrow's, even controlling for all observables. Evidence that assignment to job training is correlated with lagged wage dynamics indicates that simple specifications for the effect of training on outcomes are likely to yield biased estimates (Ashenfelter 1978). Richer models of the training assignment process may absorb this correlation while permitting identification (Heckman, Hotz, and Dabos 1987). But even these models may impose testable restrictions on the relationship between treatment and the outcome history

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

(Ashenfelter and Card 1985; Card and Sullivan 1988; Jacobson, LaLonde, and Sullivan 1993).<sup>3</sup>

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

Only a few studies have attempted to validate VAMs. Harris and Sass (2007) and Jacob and Lefgren (2008) show that value added coefficients are weakly but significantly correlated with principals’ ratings of teacher performance. Of course, if principal decisions about classroom assignments created bias in the VAMs, causality could run from principal opinions to estimated value added rather than the reverse. More relevant to the current analysis, Kane and Staiger (2008) demonstrate that VAM estimates from observational data are approximately unbiased predictors of teachers’ effects when students are randomly assigned. Although I examine a question closely related to that considered by Kane and Staiger, my larger and more representative sample permits me to extend their analysis in two ways. First, I have much more statistical power. This enables me to identify biases that are substantively important but that lie well within Kane and Staiger’s confidence intervals. Second, my sample resembles the sort that would be used for any VAM intended as a teacher compensation or retention tool. In particular, it includes teachers specializing in students (e.g., late readers) who cannot be readily identified and excluded from large-scale analyses. The likely exclusion of such teachers from Kane and Staiger’s sample quite plausibly avoids the most severe biases in observational VAM estimates.<sup>4</sup>

3. Of course, these sorts of tests cannot diagnose all model violations. If treatment assignments depend on unobserved determinants of future outcomes that are uncorrelated with the outcome history, the treatment effect estimator may be biased even though treatment is uncorrelated with past outcomes.

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

### III. STATISTICAL MODEL AND METHODS

This section develops the statistical framework for VAM analysis and introduces my tests. I begin by defining the parameters of interest in Section III.A. In Section III.B, I introduce the three VAMs that I consider. Section III.C describes the exclusion restrictions that the VAM requires to permit identification of the causal effects of interest and develops the implications of these restrictions for the relationship between the current teacher and lagged outcome. Section III.D discusses the implementation of the tests.

#### III.A. Defining the Problem

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

I am interested in whether common VAMs identify classroom effects with arbitrarily large samples. I therefore sidestep small-sample issues by considering the properties of VAM estimates as the number of students grows with the number of teachers (and classrooms) fixed.<sup>5</sup> If classroom effects are identified under these unrealistic asymptotics, VAMs may be usable in compensation and retention policy with appropriate allowances for the sampling errors that arise with finite class sizes;<sup>6</sup> if not, these corrections are likely to go awry.

A final important distinction is between identification of the variance of teacher quality and identification of individual teachers' effects. I focus exclusively on the latter. It is impractical

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

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

to report each of several thousand teachers' estimated effects, however. I therefore report only the implied standard deviations (across teachers) of teachers' actual and counterfactual effects, along with tests of the hypothesis that the teacher effects are all zero.<sup>7</sup>

### III.B. Data Generating Process and the Three VAMs

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

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

Here,  $\beta_{hgc}$  is the effect of being in classroom  $c$  in grade  $h$  on the grade- $g$  test score, and  $c(i, h) \in \{1, \dots, J_h\}$  indexes the classroom to which student  $i$  is assigned in grade  $h$ .  $\mu_i$  is individual ability. We might expect the achievement gap between high-ability and low-ability students to grow over time; this would correspond to  $\tau_k > \tau_g > 0$  for each  $k > g$ .  $\varepsilon_{ih}$  captures all other inputs in grade  $h$ , including those received from the family, nonclassroom peers, and the community. It might also include developmental factors: A precocious child might have positive  $\varepsilon$ s in early grades and negative  $\varepsilon$ s in later grades as her classmates caught up. As this example shows,  $\varepsilon$  is quite likely to be serially correlated within students across grades. Finally,  $v_{ig}$  represents measurement error in the grade- $g$  test relative to the student's "true" grade- $g$  achievement. This is independent across grades within students.<sup>8</sup>

A convenient restriction on the time pattern of classroom effects is uniform geometric decay,  $\beta_{hgc} = \beta_{hg'c} \lambda^{g'-g}$  for some  $0 \leq \lambda \leq 1$  and all  $h \leq g \leq g'$ . A special case is  $\lambda = 1$ , corresponding to perfect persistence. Although my results do not depend on these restrictions, I impose them as needed for notational simplicity.

7. Rivkin, Hanushek, and Kain (2005) develop a strategy for identifying the variance of teachers' effects, but not the effect of individual teachers, under weaker assumptions than are required by the VAMs described below.

8. I define the  $\beta$  parameters to include any classroom-level component of  $v_{ig}$  and assume that  $v_{ig}$  is independent across students in the same classroom.

I consider nonuniform decay in Section VII. Note that there is no theoretical basis for restrictions on the decay of nonclassroom effects (i.e., on  $\phi_{hg}$ ).

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

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

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

a regression of score levels (or, equivalently, of gains) on classroom indicators and the lagged score,

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

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

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

All three are widely used.<sup>9</sup> VAM2 and VAM3 can both be seen as generalizations of VAM1: Constraining  $\lambda = 1$  converts VAM2 to VAM1, whereas constraining  $\mu_i = 0$  converts VAM3.

### III.C. Exclusion Restrictions and Falsification Tests

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

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

*The Gain Score Model (VAM1).* First-differencing the production function (1), we can write the grade- $g$  gain score as

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

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

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

$$(3) \quad E[e_{1ig} | c(i, g) = c] = 0.$$

The most natural model that is consistent with (3) is for assignments to depend only on student ability,  $\mu_i$ , and for ability to have the same effect on achievement in grades  $g$  and  $g - 1$  (i.e.,  $\Delta \tau_g = 0$ ). With these restrictions, VAM1 can be seen as the first-difference estimator for a fixed effects model, with strict exogeneity of classroom assignments conditional on  $\mu_i$ . By contrast, (3) is not likely to hold if  $c(i, g)$  depends, even in part, on  $\omega_{ig-1}$ ,  $v_{ig-1}$ , or  $A_{ig-1}$ .

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

$$(4) \quad E[\Delta A_{ig-1} | c(i, g) = c] = \Delta \alpha_{g-1} + E[\beta_{g-1,g-1,c(i,g-1)} | c(i, g) = c] + E[e_{1ig-1} | c(i, g) = c].$$

The first term is constant across  $c$  and can be neglected. The second term might vary with  $c$  if (for example) a principal compensated for a bad teacher in grade  $g - 1$  by assignment to a better-than-average teacher in grade  $g$ . This can be absorbed by examining the across- $c(i, g)$  variation in  $\Delta A_{ig-1}$  *controlling for*  $c(i, g - 1)$ . I estimate specifications of this form below.<sup>10</sup> Any

10. This is a test of the hypothesis that students are randomly assigned to grade- $g$  classrooms *conditional on the  $g - 1$  classroom*. This test is uninformative unless there is independent variation in  $c(i, g - 1)$  and  $c(i, g)$ . To take one example, Nye, Konstantopoulos, and Hedges (2004) use data from the Tennessee STAR class size experiment to study teacher effects. In STAR, "streaming" was quite common, and in many schools there is zero independent variation in third grade classroom assignments controlling for second grade assignments. In this case, identification of teacher effects rests entirely on the assumption that past teachers' effects do not decay.

remaining variation across grade- $g$  classrooms in  $g - 1$  gains, after controlling for  $g - 1$  classroom assignments, must indicate that students are sorted into grade- $g$  classrooms on the basis of  $e_{1ig-1}$ .

Sorting on  $e_{1ig-1}$  would not necessarily violate (3) if  $e_{1ig}$  were not serially correlated. But the definition of  $e_{1ig}$  above indicates four sources of potential serial correlation. First, ability  $\mu_i$  appears in both  $e_{1ig}$  and  $e_{1ig-1}$  (unless  $\Delta\tau_g = 0$ ). Second, the  $\varepsilon_{ig}$  process may be serially correlated. Third, even if  $\varepsilon$  is white noise,  $\Delta\omega_{ig}$  is a moving average of order  $g - 1$  (absent strong restrictions on the  $\phi$  coefficients). Finally,  $\Delta v_{ig}$  is an MA(1), degenerate only if  $\text{var}(v) = 0$ .<sup>11</sup> Thus, (3) is not likely to hold if  $E[e_{1ig-1} | c(i, g)]$  is nonzero.

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

$$(5) \quad A_{ig} = \check{\alpha}_g + A_{ig-1}\lambda + \beta_{ggc(i, g)} + e_{2ig},$$

where  $\check{\alpha}_g = \alpha_g - \alpha_{g-1}\lambda$ . This can equivalently be expressed as a model for the grade- $g$  gain, by subtracting  $A_{ig-1}$  from each side of (5). In either case, the error is

$$(6) \quad e_{2ig} = \mu_i (\tau_g - \tau_{g-1}\lambda) + \sum_{h=1}^{g-1} \varepsilon_{ih} (\phi_{hg} - \phi_{hg-1}\lambda) + \varepsilon_{ig} + (v_{ig} - v_{ig-1}\lambda).$$

As before, each of the terms in (6) is likely to be serially correlated.

The exclusion restriction for VAM2 is that  $e_{2ig}$  is uncorrelated with  $c(i, g)$  conditional on  $A_{ig-1}$ . This would hold if  $c(i, g)$  were randomly assigned conditional on  $A_{ig-1}$ . It is unlikely to hold if assignments depend on  $e_{2ig-1}$  or on any of its components (including  $\mu_i$ ).<sup>12</sup> As with the VAM1, I test the VAM2 exclusion restriction by

11. In Rothstein (2008), I conclude that  $\Delta v_{ig}$  accounts for as much as 80% of the variance of  $\Delta A_{ig}$ .

12. Alternatively, if  $\tau_g - \tau_{g-1}\lambda$  is constant across  $g$ , (5) can be seen as a fixed effects model with a lagged dependent variable.  $\lambda$  and  $\beta_{gg}$  can be identified via IV or GMM (instrumenting for  $\Delta A_{ig-1}$  in a model for  $\Delta A_{ig}$ ) if  $c(i, g)$  depends on  $\mu_i$  but is strictly exogenous conditional on this (Anderson and Hsiao 1981; Arellano and Bond 1991). See, for example, Koedel and Betts (2007). Value added researchers typically apply OLS to (5). This is inconsistent for  $\lambda$  and identifies  $\beta_{ggc}$  only if  $c(i, g)$  is random conditional on  $A_{ig-1}$ .

reestimating the model with the  $g - 1$  gain as the dependent variable. By rearranging the lag of (5), we can write the  $g - 1$  gain as

$$(7) \quad \Delta A_{ig-1} = \lambda^{-1} (\check{\alpha}_g + A_{ig-1}(\lambda - 1) + \beta_{g-1,g-1,c(i,g-1)} + e_{2ig-1}).$$

Thus, the grade- $g$  classroom assignment will have predictive power for the gain in grade  $g - 1$ , controlling for the  $g - 1$  achievement level, if grade- $g$  classrooms are correlated either with the  $g - 1$  teacher's effect (i.e., with  $\beta_{g-1,g-1,c(i,g-1)}$ ) or with  $e_{2ig-1}$ .<sup>13</sup> As in VAM1, the former can be ruled out by controlling for  $g - 1$  classroom assignments; the latter would indicate a violation of the VAM2 exclusion restriction if  $e_2$  is serially correlated.

*The Fixed Effects in Gains Model (VAM3).* For the final VAM, we return to equation (2) and to the earlier assumption of zero decay of teachers' effects.<sup>14</sup> The student fixed effects used in VAM3 absorb any variation in  $\mu_i$  (assuming that  $\Delta\tau_g = 1$  for each  $g$ ). Thus, the VAM3 error term is  $e_{3ig} = \Delta\omega_{ig} + \Delta v_{ig}$ .

The reliance on fixed effects, combined with the small time dimension of student data sets, means that VAM3 requires stronger assumptions than the earlier models. To avoid bias in the teacher effects  $\beta_{ggc}$ , even in large samples, teacher assignments must be strictly exogenous conditional on  $\mu_i$ :  $E[e_{3ih} | c(i, g)] = 0$  for all  $g$  and all  $h$  (Wooldridge 2002, p. 253).<sup>15</sup> Conditional strict exogeneity means that the same information,  $\mu_i$  or some function of it, is used to make teacher assignments in each grade. This requires, in effect, that principals decide on classroom assignments for the remainder of a child's career before she starts kindergarten. If teacher assignments are updated each year in response to the student's performance during the previous year, strict exogeneity is violated.

13. The test can alternatively be expressed in terms of a model for the score level in  $g - 2$ . (Simply rearrange terms in (7).) The VAM2 exclusion restriction of random assignment conditional on  $A_{ig-1}$  will be rejected if the grade- $g$  classroom predicts  $A_{ig-2}$  conditional on  $A_{ig-1}$ .

14. Although VAM1 and VAM2 can easily be generalized to allow for nonuniform decay, VAM3 cannot.

15. For practical value added implementations, it is rare to have more than three or four student grades, so asymptotics based on the  $g$  dimension are infeasible. One approach if strict exogeneity does not hold is to focus on the first difference of (2). OLS estimation of the first-differenced equation requires that  $c(i, g)$  be uncorrelated with  $e_{3ig-1}$ ,  $e_{3ig}$ , and  $e_{3ig+1}$ . Though this is weaker than strict exogeneity, it is difficult to imagine an assignment process that would satisfy one but not the other. If the OLS requirements are not satisfied, the only option is IV/GMM (see note 12), instrumenting for both the  $g$  and  $g - 1$  classroom assignments. Satisfactory instruments are not apparent.

As before, my test is based on analyses of the apparent effects of grade  $g$  teachers on gains in prior grades. Consider estimation of VAM1, without the student fixed effects that are added in VAM3. If teacher assignments depend on ability, this will bias the VAM coefficients and will lead me to reject the VAM1 exclusion restriction. But the conditional strict exogeneity assumption imposes restrictions on the coefficients from the VAM1 falsification test. Under this assumption, the only source of bias in VAM1 is the omission of controls for  $\mu_i$ . As  $\mu_i$  enters into *every* grade's gain equation, grade- $g$  teachers should have the same apparent effects on  $g - 2$  gains as they do on  $g - 1$  gains. An indication that these differ would indicate that omitted time-varying determinants of gains are correlated with teacher assignments, and therefore that assignments are not strictly exogenous.

Following Chamberlain (1984), consider a projection of  $\mu$  onto the full sequence of classroom assignments in grades 1 through  $G$ :

$$(8) \quad \mu_i = \xi_{1c(i,1)} + \dots + \xi_{Gc(i,G)} + \eta_i.$$

$\xi_{hc}$  is the incremental information about  $\mu_i$  provided by the knowledge that the student was in classroom  $c$  in grade  $h$ , conditional on classroom assignments in all other grades. Substituting (8) into (2), we obtain

$$(9) \quad \Delta A_{ig} = \Delta \alpha_g + \sum_{h=1}^G \pi_{hgc(i,h)} + \eta_i + e_{3ig},$$

where  $\pi_{ggc} = \xi_{gc} \Delta \tau_g + \beta_{ggc}$  and  $\pi_{hgc} = \xi_{hc} \Delta \tau_g$  for  $h \neq g$ . Under conditional strict exogeneity,  $E[e_{3ih} | c(i, 1), \dots, c(i, G)] = 0$  for each  $h$ , and the fact that (8) is a linear projection ensures that  $\eta_i$  is uncorrelated with the regressors as well. An OLS regression of grade- $g$  gains onto classroom indicators in grades 1 through  $G$  thus estimates the  $\pi_{hgc}$  coefficients without bias. When  $G \geq 3$ , the underlying parameters are overidentified. To see this, note that

$$(10) \quad \pi_{hgc} = \xi_{hc} \Delta \tau_g = \xi_{hc} \Delta \tau_{g-1} \frac{\Delta \tau_g}{\Delta \tau_{g-1}} = \pi_{h,g-1,c} \frac{\Delta \tau_g}{\Delta \tau_{g-1}}$$

for all  $h > g$ : The coefficient for grade- $h$  classroom  $c$  in a model of gains in grade  $g$  is proportional to the same coefficient in a model of gains in  $g - 1$ . If there are  $J_h$  grade- $h$  classrooms in the sample, this represents  $J_h - 1$  overidentifying restrictions on

the  $2J_h$  elements of the vectors  $\Pi_{hg} = \{\pi_{hg1} \dots \pi_{hgJ_h}\}'$  and  $\Pi_{hg-1} = \{\pi_{h,g-1,1} \dots \pi_{h,g-1,J_h}\}'$ .<sup>16</sup>

To test these restrictions, I estimate the the  $J_h$ -vector  $\Xi_h$  and the scalars  $\Delta\tau_1$  and  $\Delta\tau_2$  that minimize

$$(11) \quad D = \left( \begin{pmatrix} \hat{\Pi}_{hg-1} \\ \hat{\Pi}_{hg} \end{pmatrix} - \begin{pmatrix} \Xi_h \Delta\tau_{g-1} \\ \Xi_h \Delta\tau_g \end{pmatrix} \right)' W^{-1} \left( \begin{pmatrix} \hat{\Pi}_{hg-1} \\ \hat{\Pi}_{hg} \end{pmatrix} - \begin{pmatrix} \Xi_h \Delta\tau_{g-1} \\ \Xi_h \Delta\tau_g \end{pmatrix} \right),$$

using the sampling variance of  $(\hat{\Pi}'_{hg-1} \hat{\Pi}'_{hg})'$  as  $W$ . Under the null hypothesis of strict exogeneity, the minimized value  $D$  is distributed  $\chi^2$  with  $J_h - 1$  degrees of freedom.<sup>17</sup> If  $D$  is above the 95% critical value from this distribution, the null is rejected. Intuitively, the correlation between corresponding elements of the coefficient vectors  $\Pi_{hg-1}$  and  $\Pi_{hg}$ , representing apparent “effects” of grade- $h$  teachers on gains in grades  $g - 1$  and  $g$  ( $g < h$ ), should be 1 or  $-1$  under the null; a correlation far from this would suggest that the exclusion restriction is violated.

### III.D. Implementation

To put the three VAMs in the best possible light, I focus on estimation of within-school differences in classroom effects. For many purposes, one might want to make across-school comparisons. But students are not randomly assigned to schools, and those at one school may gain systematically faster than those at another for reasons unrelated to teacher quality. Random assignment to classrooms within schools is at least somewhat plausible. To isolate within-school variation, I augment each of the estimating equations discussed above with a set of indicators for the school attended.<sup>18</sup> The tests for VAM1 and VAM2 then amount to tests of whether students are (conditionally) randomly assigned to

16. When  $G > 3$ , there are many such pairs of vectors that must be proportional. Even when  $G = 3$ , there are additional overidentifying restrictions created by similar proportionality relationships for teachers’ effects on *future* gains. These restrictions might fail either because strict exogeneity is violated or because teachers’ effects decay (that is,  $\beta_{hh} \neq \beta_{hg}$  for some  $g > h$ ). I therefore focus on restrictions on the coefficients for teachers’ effects on *past* gains, as these provide sharper tests of strict exogeneity.

17. Although there are  $J_h + 2$  unknown parameters, they are underidentified: Multiplying  $\Xi_h$  by a constant and dividing  $\Delta\tau_{g-1}$  and  $\Delta\tau_g$  by the same constant does not change the fit.

18. This makes  $W$  singular in (11). For the OMD analysis of VAM3, I drop the elements of  $\pi_{gh}$  that correspond to the largest class at each school.

classrooms *within schools*. They resemble tests of successful randomization in stratified experiments, treating schools as strata.

Intuitively, I will reject random assignment if replacing a set of school indicators with grade- $g$  grade classroom indicators adds more explanatory power for  $g - 1$  gains than would be expected by chance alone. Let  $S_g$  and  $T_g$  be matrices of indicators for grade- $g$  schools and classrooms. These are collinear, so to eliminate this I define  $\tilde{T}_g$  as the submatrix of  $T_g$  that results from excluding the columns corresponding to one classroom per school. The VAM1 test is based on a simple regression:

$$(12) \quad \Delta A_{g-1} = \alpha + S_g \delta + \tilde{T}_g \beta + e.$$

The identifying assumption of VAM1 is rejected if  $\beta \neq 0$ . I use a heteroscedasticity-robust score test (Wooldridge 2002, p. 60) to evaluate this. I also estimate versions of (12) that include controls for grade- $(g - 1)$  classroom assignments. To test VAM2, I simply add a control for  $A_{g-1}$  on the right-hand side of (12).

It is clear from the definition of  $\tilde{T}_g$  that only schools with multiple classrooms per grade can contribute to the analysis. One might be concerned that schools with only two or three classrooms will be misleading, as even with random assignment of students to classrooms there will be substantial overlap in the composition of a student's grade- $g$  and grade- $(g - 1)$  classrooms. The Online Appendix presents a Monte Carlo analysis of the VAM1 and VAM2 tests in schools of varying sizes. The VAM1 test has appropriate size even with just two classrooms per school, so long as the number of students per classroom is large. (Recall that I focus on large-class asymptotics.) With small classes, the asymptotic distribution of the test statistic is an imperfect approximation, and as a result the test over-rejects slightly. When there are twenty students per class, the test of VAM1 has size around 10%. With empirically reasonable parameter values, the VAM2 test performs similarly.<sup>19,20</sup>

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

20. Kinsler (2008) claims that the VAM3 test also overrejects in simulations. In personal communication, he reports that the problem disappears with large classes.

I also report the standard deviation of the teacher coefficients (the  $\beta$ s in (12)) themselves. The standard deviation of the estimated coefficients necessarily exceeds that of the true coefficients (those that would be identified with large samples of students per teacher, even if these are biased estimates of teachers' true causal effects). Aaronson, Barrow, and Sander (2007) propose a simple estimator for the variance of the true coefficients across teachers. Let  $\beta$  be a mean-zero vector of true projection coefficients and let  $\hat{\beta}$  be an unbiased finite-sample estimate of  $\beta$ , with  $E[\beta'(\hat{\beta} - \beta)] = 0$ . The variance (across elements) of  $\beta$  can be written as

$$(13) \quad E[\beta'\beta] = E[\hat{\beta}'\hat{\beta}] - E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)].$$

$E[\hat{\beta}'\hat{\beta}]$  is simply the variance across teachers of the coefficient estimates.<sup>21</sup>  $E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)]$  is the average heteroscedasticity-robust sampling variance. I weight each by the number of students taught.

Specifications that include indicators for classroom assignments in several grades simultaneously—such as that used for the test of VAM3—introduce two complications. First, the coefficients for teachers in different grades can only be separately identified when there is sufficient shuffling of students between classrooms. If students are perfectly streamed—if a student's classmates in third grade are also his or her classmates in fourth grade—the third and fourth grade classroom indicators are collinear. I exclude from my samples a few schools where inadequate shuffling leads to perfect collinearity. Second, these regressions are difficult to compute, due to the presence of several overlapping sets of fixed effects. As discussed in the Online Appendix, this difficulty is avoided by restricting the samples to students who do not switch schools during the grades for which classroom assignments are controlled.

#### IV. DATA AND SAMPLE CONSTRUCTION

The specifications described in Section III require longitudinal data that track students' outcomes across several grades, linked to classroom assignments in each grade. I use administrative data on elementary students in North Carolina public schools, assembled and distributed by the North Carolina

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

Education Research Data Center. These data have been used for several previous value added analyses (see, e.g., Clotfelter, Ladd, and Vigdor [2006]; Goldhaber [2007]).

I examine end-of-grade math and reading tests from grades 3 through 5, plus “pretests” from the beginning of third grade (which I treat as second grade tests). I standardize the scale scores separately for each subject–grade–year combination.<sup>22</sup>

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

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

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

Table I presents summary statistics. I show statistics for the population, for the base sample, and for my most restricted sample

22. The original score scale is meant to ensure that one point corresponds to an equal amount of learning at each grade and at each point in the within-grade distribution. Rothstein (2008) and Ballou (2009) emphasize the importance of this property for value added modeling. All of the results here are robust to using the original scale.

TABLE I  
SUMMARY STATISTICS

	Population		Base sample		Most restricted sample	
	Mean (1)	SD (2)	Mean (3)	SD (4)	Mean (5)	SD (6)
# of students	99,071		60,740		23,415	
# of schools	1,269		868		598	
1 fifth grade teacher	122		0		0	
2 fifth grade teachers	168		207		122	
3-5 fifth grade teachers	776		602		440	
>5 fifth grade teachers	203		59		36	
# of fifth grade classrooms	4,876		3,040		2,116	
# of fifth grade classrooms w/valid teacher match	3,315		3,040		2,116	
Female (%)	49		50		51	
Black (%)	29		28		23	
Other nonwhite (%)	8		7		6	
Consistent student record (%)	99		100		100	
Complete test score record, G4-5 (%)	88		99		100	
G3-5 (%)	81		91		100	
G2-5 (%)	72		80		100	
Changed schools between G3 and G5 (%)	30		27		0	
Valid teacher assignment in grade 3 (%)	68		78		100	
grade 4 (%)	70		86		100	
grade 5 (%)	72		100		100	
Fr. of students in G5 class in same G4 class	0.22	[0.19]	0.22	[0.17]	0.30	[0.19]
Fr. of students in G5 class in same G3 class	0.15	[0.15]	0.15	[0.13]	0.28	[0.18]

TABLE I  
(CONTINUED)

	Population		Base sample		Most restricted sample	
	Mean	SD	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)
Math scores						
Third grade (beginning of year)	0.11	[0.97]	0.14	[0.96]	0.20	[0.96]
Third grade (end of year)	0.09	[0.94]	0.11	[0.94]	0.19	[0.91]
Fourth grade (end of year)	0.04	[0.97]	0.07	[0.97]	0.20	[0.93]
Fifth grade (end of year)	0.00	[1.00]	0.09	[0.98]	0.20	[0.94]
Third grade gain	-0.02	[0.70]	-0.02	[0.69]	0.00	[0.69]
Fourth grade gain	-0.02	[0.58]	-0.01	[0.58]	0.01	[0.56]
Fifth grade gain	-0.01	[0.55]	0.01	[0.55]	-0.01	[0.53]
Reading scores						
Third grade (beginning of year)	0.08	[0.98]	0.12	[0.98]	0.17	[0.98]
Third grade (end of year)	0.08	[0.95]	0.11	[0.94]	0.19	[0.91]
Fourth grade (end of year)	0.04	[0.98]	0.07	[0.97]	0.18	[0.93]
Fifth grade (end of year)	0.00	[1.00]	0.07	[0.97]	0.17	[0.94]
Third grade gain	0.01	[0.76]	0.00	[0.75]	0.01	[0.75]
Fourth grade gain	-0.02	[0.59]	-0.02	[0.59]	0.00	[0.57]
Fifth grade gain	-0.01	[0.59]	0.00	[0.58]	-0.02	[0.57]

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

(used for estimation of equation (9)). The last is much smaller than the others, largely because I require students to have attended the same school in grades 3 through 5 and to have valid teacher matches in each grade. Table I indicates that the restricted sample has higher mean fifth grade scores than the full population. This primarily reflects the lower scores of students who switch schools frequently.<sup>23</sup> Average fifth grade gains are similar across samples. The Online Appendix describes each sample in more detail.

As discussed above, my tests can be applied only if there is sufficient reshuffling of classrooms between grades. Table A2 in the Online Appendix shows the fraction of students' fifth grade classmates who were also in the same fourth grade classes, by the number of fourth grade classes at the school. Complete reshuffling (combined with equal-sized classes) would produce 0.5 with two classes, 0.33 with three, and so on. The actual fractions are larger than this, but only slightly. In schools with exactly three fifth grade teachers, for example, 35% of students' fifth grade classmates were also their classmates in fourth grade. In only 7% of multiple-classroom schools do the fourth and fifth grade classroom indicators have deficient rank.

Table II presents the correlation of test scores and gains across grades and subjects. The table indicates that fifth grade scores are correlated above .8 with fourth grade scores in the same subject, whereas correlations with scores in earlier grades or other subjects are somewhat lower. Fifth grade gains are strongly negatively correlated with fourth grade levels and gains in the same subject and weakly negatively correlated with those in the other subject. The correlations between fifth and third grade gains are small but significant both within and across subjects.

VAM3 is predicated on the notion that student ability is an important component of annual gains. Assuming that high-ability students gain faster, this would imply positive correlations between gains in different years. There is no indication of this in Table II. One potential explanation is that noise in the annual tests introduces negative autocorrelation in gains, but I conclude elsewhere (Rothstein 2008) that even true gains are negatively

23. Table I shows that average third and fourth grade scores in the "population" are well above zero. The norming sample that I use to standardize scores in each grade consists of all students in that grade in the relevant year (i.e., of all third graders in 1999), whereas only those who make normal progress to fifth grade in 2001 are included in the sample for columns (1) and (2). The low scores of students who repeat grades account for the discrepancy.

TABLE II  
CORRELATIONS OF TEST SCORES AND SCORE GAINS ACROSS GRADES

	Summary statistics		Correlations				N
			Fifth grade score		Fifth grade gain		
	Mean	SD	Math	Reading	Math	Reading	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<b>Math scores</b>							
G5	0.02	1.00	1	.78	.29	.08	70,740
G4	0.07	0.97	.84	.73	-.27	-.07	61,535
G3	0.09	0.95	.80	.70	-.02	-.03	57,382
G3 pretest	0.08	0.97	.71	.64	.00	-.03	50,661
<b>Reading scores</b>							
G5	0.01	1.00	.78	1	.10	.31	70,078
G4	0.06	0.97	.73	.82	-.05	-.29	61,535
G3	0.09	0.95	.70	.78	-.01	-.05	57,344
G3 pretest	0.08	0.99	.59	.65	.00	-.05	50,629
<b>Math gains</b>							
G4-G5	0.01	0.55	.29	.10	1	.25	61,349
G3-G4	-0.01	0.58	.11	.07	-.41	-.07	56,171
G2-G3	0.02	0.70	.08	.05	-.02	.01	50,615
<b>Reading gains</b>							
G4-G5	0.00	0.58	.08	.31	.25	1	60,987
G3-G4	-0.02	0.59	.08	.10	-.08	-.41	56,159
G2-G3	0.02	0.75	.09	.10	-.01	.02	50,558

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

autocorrelated. This strongly suggests that VAM3 is poorly suited to the test score data generating process.

## V. RESULTS

Tables III, IV, and V present results for the three VAMs in turn. I begin with VAM1, in Table III. I regress fifth grade math and reading gains (in columns (1) and (2), respectively) on indicators for fifth grade schools and classrooms, excluding one classroom per school. In each case, the hypothesis that all of the classroom coefficients are zero (i.e., that classroom indicators have no explanatory power beyond that provided by school indicators) is decisively rejected. The VAM indicates that the within-school standard deviations of fifth grade teachers' effects on math and reading are 0.15 and 0.11, respectively. This is similar to what

TABLE III  
EVALUATION OF VAM1: REGRESSION OF GAIN SCORES ON TEACHER INDICATORS

	(1) Fifth grade gain		(2) Fifth grade gain		(3) Fourth grade gain		(4) Fourth grade gain		(5) Fifth grade gain		(6) Fifth grade gain		(7) Fourth grade gain		(8) Fourth grade gain	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading	Math	Reading
Teacher coefficients																
Fifth grade teachers																
Unadjusted SD	0.179	0.160	0.134	0.142	0.134	0.142	0.134	0.142	0.197	0.181	0.197	0.181	0.151	0.168	0.151	0.168
Adjusted SD	0.149	0.113	0.077	0.084	0.077	0.084	0.077	0.084	0.163	0.126	0.163	0.126	0.090	0.105	0.090	0.105
<i>p</i> -value	<.001	<.001	.016	.002	.016	.002	.016	.002	<.001	<.001	<.001	<.001	.035	<.001	.035	<.001
Fourth grade teachers																
Unadjusted SD									0.188	0.181	0.188	0.181	0.220	0.193	0.220	0.193
Adjusted SD									0.150	0.125	0.150	0.125	0.182	0.140	0.182	0.140
<i>p</i> -value									<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Exclude invalid fourth grade teacher assignments & fifth grade movers?	n	n	n	n	n	n	n	n	y	y	y	y	y	y	y	y
# of students	55,142	55,142	55,142	55,142	55,142	55,142	55,142	55,142	40,661	40,661	40,661	40,661	40,661	40,661	40,661	40,661
# of fifth grade teachers	3,038	3,038	3,038	3,038	3,038	3,038	3,038	3,038	2,761	2,761	2,761	2,761	2,761	2,761	2,761	2,761
# of schools	868	868	868	868	868	868	868	868	783	783	783	783	783	783	783	783
<i>R</i> <sup>2</sup>	.195	.100	.132	.086	.132	.086	.132	.086	.297	.176	.297	.176	.254	.174	.254	.174
Adjusted <i>R</i> <sup>2</sup>	.148	.047	.081	.033	.081	.033	.081	.033	.203	.066	.203	.066	.154	.064	.154	.064

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

TABLE IV  
EVALUATION OF VAM2: REGRESSIONS WITH CONTROLS FOR LAGGED SCORE LEVELS

	Fifth grade gain		Fourth grade gain		Fifth grade gain		Fourth grade gain	
	Math	Reading	Math	Reading	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher coefficients								
Fifth grade teachers								
Unadjusted SD	0.176	0.150	0.120	0.129	0.191	0.169	0.138	0.150
Adjusted SD	0.150	0.109	0.067	0.076	0.161	0.121	0.079	0.091
<i>p</i> -value	<.001	<.001	.040	.007	<.001	<.001	.162	.001
Fourth grade teachers								
Unadjusted SD					0.160	0.162	0.182	0.175
Adjusted SD					0.121	0.109	0.142	0.126
<i>p</i> -value					<.001	<.001	<.001	<.001
Continuous controls								
Fourth grade math score	-0.317 (0.004)	0.239 (0.004)	0.368 (0.004)	-0.213 (0.004)	-0.292 (0.004)	0.255 (0.005)	0.332 (0.005)	-0.229 (0.005)
Fourth grade reading score	0.195 (0.004)	-0.383 (0.004)	-0.218 (0.004)	0.380 (0.004)	0.189 (0.004)	-0.387 (0.005)	-0.206 (0.005)	0.379 (0.005)
Exclude invalid fourth grade teacher assignments & fifth grade movers?	n	n	n	n	y	y	y	y
# of students	55,142	55,142	55,142	55,142	40,661	40,661	40,661	40,661
# of fifth grade teachers	3,038	3,038	3,038	3,038	2,761	2,761	2,761	2,761
# of schools	868	868	868	868	783	783	783	783
<i>R</i> <sup>2</sup>	.313	.249	.274	.237	.385	.315	.354	.307
Adjusted <i>R</i> <sup>2</sup>	.273	.206	.231	.193	.302	.224	.268	.215

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

TABLE V  
CORRELATED RANDOM EFFECTS EVALUATION OF VAM3: GAIN SCORE SPECIFICATION WITH STUDENT FIXED EFFECTS

	Math			Reading		
	Third grade (1)	Fourth grade (2)	Corr(1),(2)) (3)	Third grade (4)	Fourth grade (5)	Corr(4),(5)) (6)
Standard deviation of teacher effects, adjusted	Unrestricted model					
Fifth grade teacher	0.135	0.099	-.04	0.144	0.123	-.06
Fourth grade teacher	0.136	0.193	-.07	0.160	0.163	-.08
Third grade teacher	0.228	0.166	-.36	0.183	0.145	-.24
Fit statistics						
$R^2$	.314	.376		.245	.284	
Adjusted $R^2$	.129	.209		.042	.092	
	Restricted model (optimal minimum distance)					
Ratio, effect on G4/effect on G3		0.14			1.17	
SD of G5 teacher effects	0.126	0.018		0.088	0.103	
Objective function		2,136			2,174	
95% critical value		1,684			1,684	
$p$ -value		<.001			<.001	

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

has been found in other studies (e.g., Aaronson, Barrow, and Sander [2007]; Rivkin, Hanushek, and Kain [2005]).

Columns (3) and (4) present falsification tests in which fourth grade gains are substituted for the fifth grade gains as dependent variables, with the specification otherwise unchanged. The standard deviation of fifth grade teachers' "effects" on fourth grade gains is 0.08 in each subject, and the hypothesis of zero association is rejected in each specification.<sup>24</sup> In both the standard deviation and statistical significance senses, fifth grade classroom assignments are slightly more strongly associated with fourth grade reading gains than with math gains.

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

These results strongly suggest that the exclusion restrictions for VAM1 are violated. To demonstrate this conclusively, however, we need to show that the residual in VAM1,  $e_{1ig}$ , is serially correlated. To examine this, I reestimated VAM1 for fourth grade teachers' effects on fourth grade gains. The correlation between  $\hat{e}_{1i4}$  and  $\hat{e}_{1i5}$  is  $-.38$  in math and  $-.37$  in reading.

The negative serial correlation of  $e_1$  implies that students with high gains in fourth grade will tend to have low gains in fifth grade, and vice versa. Because VAM1 evidently does not

24. The table shows analytic  $p$ -values based on the  $F$  distribution. As noted earlier, simulations suggest that my tests over-reject slightly. When I use the empirical distribution of test statistics from an appropriately calibrated Monte Carlo simulation (discussed in the Online Appendix) to construct  $p$ -values, these are .031 and .004, respectively.

adequately control the determinants of classroom assignments, it gives unearned credit to teachers who are assigned students who did poorly in fourth grade, as these students will predictably post unusually high fifth grade gains when they revert toward their long-run means. Similarly, teachers whose students did unusually well in fourth grade will be penalized by the students' fall back toward their long-run means in fifth grade. Indeed, an examination of the VAM1 coefficients indicates that fifth grade teachers whose students have above average fourth grade gains have systematically lower estimated value added than teachers whose students underperformed in the prior year. Importantly, this pattern is stronger than can be explained by sampling error in the estimated teacher effects; it reflects true mean reversion and not merely measurement error.

Table IV repeats the falsification exercise for VAM2. The structure is identical to that of Table III. Columns (1) and (2) present estimates of the basic VAM for fifth grade teachers' effects on fifth grade gains, controlling for fourth grade math and reading scores. The standard deviations of fifth grade teachers' effects are nearly identical to those in Table III. Columns (3) and (4) substitute fourth grade gains as the dependent variable. Once again, we see that fifth grade teachers are strongly predictive, more so in reading than in math.<sup>25</sup> Columns (5)–(8) augment the specification with controls for fourth grade teachers. The fifth grade teacher coefficients are no longer jointly significant in the fourth grade math gain specification, though they remain quite large in magnitude. They are still highly significant in the specification for fourth grade reading gains.

The VAM2 residuals, like those from VAM1, are nontrivially correlated between fourth and fifth grades,  $-.21$  for math gains and  $-.19$  for reading. They are also correlated across subjects:  $-.14$  between fourth grade reading and fifth grade math. Thus, the evidence that fifth grade teacher assignments are correlated with the fourth grade residuals indicates that the VAM2 exclusion restriction is violated, regardless of whether the dependent variable is the math or the reading score. As before, fifth grade teachers' effects on fifth grade scores are negatively correlated with their counterfactual "effects" on fourth grade gains, suggesting that mean reversion in student achievement—combined

25.  $p$ -values based on Monte Carlo simulations (see note 24) are .086 and .018 in columns (3) and (4), respectively.

with nonrandom classroom assignments—is an important source of bias in VAM2.

To implement the VAM3 falsification test, I begin by selecting the subsample with nonmissing third and fourth grade gains; valid teacher assignments in grades 3, 4, and 5; and continuous enrollment at the same school in all three grades. I exclude 26 schools where the three sets of indicators for teachers in grades 3, 4, and 5 (dropping one teacher in each grade from each school) are collinear. I then regress both the third and fourth grade gains on school indicators and on each of the three sets of teacher indicators.<sup>26</sup>

Table V reports estimates for math gains, in columns (1) and (2), and for reading gains, in columns (4) and (5). The first panel shows the standard deviations (adjusted for sampling error) of the coefficients for each grade's teachers. Gains in each subject and in each grade are substantially correlated with classroom assignments in all three grades. Although  $p$ -values are not shown, in all twelve cases the hypothesis of zero effects is rejected. Columns (3) and (6) report the across-teacher correlations between the coefficients in the models for third and fourth grade gains (i.e., between  $\Pi_{g3}$  and  $\Pi_{g4}$ ). The most important correlation is that for fifth grade teachers,  $-.04$  for math and  $-.06$  for reading. Recall that strict exogeneity implies that the fifth grade teacher coefficients in the model for fourth grade gains should be proportional to the corresponding coefficients in the model for third grade gains,  $\Pi_{54} = (\Delta\tau_4/\Delta\tau_3)\Pi_{53}$ , implying a correlation of  $\pm 1$ . The near-zero correlations strongly suggest that a single ability factor is unable to account for the apparent "effects" of fifth grade teachers on gains in earlier grades.

Indeed, these correlations are direct evidence against the VAM3 identifying assumption of conditional strict exogeneity. The lower panel of Table V presents OMD estimates of the restricted model.<sup>27</sup> For math scores, the estimated ratio  $\Delta\tau_4/\Delta\tau_3$  is  $0.14$ , implying that student ability is much more important to third grade than to fourth grade gains. Thus, the constrained estimates

26. It is not essential to the correlated random effects test that the full sequence of teacher assignments back to grade 1 be observed, but the test may over-reject if classroom assignments in grades 3–5 are correlated with those in first and second grade and if the latter have continuing effects on third and fourth grade gains. Recall, however, that VAM3 assumes such lagged effects away.

27. The OMD analysis uses a variance–covariance matrix  $W$  that is robust to arbitrary heteroscedasticity and within-student, between-grade clustering. See the Online Appendix.

imply negligible coefficients for fifth grade teachers in the equation for fourth grade gains and do a very poor job of fitting the unconstrained estimate of the standard deviation of these coefficients, 0.099. The test statistic  $D$  is 2,136, and the overidentifying restrictions are overwhelmingly rejected. In the reading specification, the  $\Delta\tau_4/\Delta\tau_3$  ratio is close to one, and the restricted model allows meaningful coefficients on fifth grade teachers in both the third and fourth grade gain equations, albeit much less variability than is seen in the unconstrained model. But the test statistic is even larger here, and the restricted model is again rejected. We can thus conclude that fifth grade teacher assignments are not strictly exogenous with respect to either math or reading gains, even conditional on single-dimensional (subject-specific) student heterogeneity. The identifying assumption for VAM3 is thus violated.

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

## VI. HOW MUCH DOES THIS MATTER?

The results in Section V indicate that the identifying assumptions for all three VAMs are violated in the North Carolina data. However, if classroom assignments nearly satisfied the assumptions underlying the VAMs, the models might yield almost unbiased estimates of teachers’ causal effects. In this section, I

use the degree of sorting on prior outcomes to quantify the magnitude of the biases resulting from nonrandom assignments. I focus on VAM1 and VAM2, as the lack of correlation between third and fifth grade gains (Table II) strongly suggests that the additional complexity and strong maintained assumptions of VAM3 are unnecessary.

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

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

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

TABLE VI  
MAGNITUDE OF BIAS IN VAM1 AND VAM2 RELATIVE TO A RICHER SPECIFICATION  
THAT CONTROLS FOR ALL PAST OBSERVABLES

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

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

Of course, the exercise carried out here can only diagnose bias in VAM1 and VAM2 from selection on *observables*—variables that can easily be included in the VAM specification. In a companion paper (Rothstein 2009), I attempt to quantify the bias that is likely to result from selection on unobservables. Following the intuition of Altonji, Elder, and Taber (2005) that the weight of observable (to the econometrician) and unobservable variables in classroom assignments is likely to mirror their relative weights in predicting achievement, one can use the degree of sorting on observables to estimate the importance of unobservables and therefore the magnitude of the bias in estimated teacher effects. Under varying assumptions about the amount of information that parents and principals have, I find that the bias from nonrandom assignments is quite plausibly 75% as large (in standard deviation terms) as the estimates of teachers' effects in VAM1, and perhaps half this large in VAM2.<sup>28</sup>

To provide a better sense of the import of nonrandom classroom assignments for the value of VAMs in teacher compensation

28. Kane and Staiger's (2008) comparison of experimental and nonexperimental value added estimates would be unlikely to detect biases of this magnitude.

and retention decisions, I simulate true and estimated teacher effects with joint distributions resembling those reported in Table VI and in Rothstein (2009). For each of several scenarios characterizing the assignment of students to classrooms, I generate 10,000 teachers' true effects and coefficients from VAMs 1, 2, and 4.<sup>29</sup> I assume that true effects and biases are both normally distributed, and that the VAM coefficients are free of sampling error. I then compute three statistics to summarize the relationship of the VAM estimates to teachers' true effects: the correlation between teachers' true effects and the VAM coefficients, the rank correlation, and the fraction of teachers with true effects in the top quintile who are indicated to be in the top quintile by the VAMs.

Results are presented in Table VII. Each panel corresponds to a distinct assumption about the classroom assignment process. In the first panel, I assume that selection is solely on the basis of the observed test score history. Using the model for reading scores from Table VI, the standard deviation of teachers' true effects is 0.148, and the standard deviations of the biases in VAM1 and VAM2 are 0.054 and 0.028, respectively. Columns (4)–(6) show the reliability of teacher quality under different metrics. True effects and ranks are very highly correlated with the effects and ranks indicated by VAMs 1 and 2. From 79% to 90% of teachers who are in the top quintile of the actual quality distribution are judged to be so by the simple VAMs.

But this analysis assumes, implausibly, that selection is solely on observables. Panels B–E present alternative estimates that allow variables that are not controlled even in VAM4 to play a role in classroom assignments, as in Rothstein (2009). In Panel B, I assume that classroom assignments depend both on the test score history that is reported in my data and on a second, unobserved history (e.g., student grades) that provides an independent, equally noisy measure of the student's trajectory through grades 2–4. Allowing for this moderate degree of selection-on-unobservables notably degrades the performance of VAM1, but VAM2 and VAM4 continue to perform reasonably well. In Panel C, I assume that there are two separate unobserved achievement measures. Performance degrades still further; although the correlations between true effects and the VAM2 and VAM4 estimates

29. It is not possible to use the estimates from Table VI directly because I wish to abstract from the role of sampling error. The simulation is described in greater detail in the Online Appendix.

TABLE VII  
SIMULATIONS OF THE EFFECTS OF STUDENT SELECTION AND HETEROGENEOUS DECAY  
ON TEACHER QUALITY ESTIMATES

	Data generating process			Simulation: comparisons between true effects and those indicated by VAM		
	SD of truth	SD of bias	(2) as % of (1)	Correlation	Rank correlation	Reliability of top quintile ranking
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Selection is on observables						
VAM1	0.148	0.054	36%	.93	.93	0.79
VAM2	0.148	0.028	19%	.98	.98	0.90
VAM4	0.148	0	0%	1.00	1.00	1.00
Panel B: Selection is on history of two tests, one observed						
VAM1	0.148	0.124	84%	.77	.75	0.62
VAM2	0.148	0.049	33%	.95	.94	0.82
VAM4	0.148	0.028	19%	.98	.98	0.89
Panel C: Selection is on history of three tests, one observed						
VAM1	0.148	0.137	92%	.74	.73	0.60
VAM2	0.148	0.060	40%	.93	.92	0.78
VAM4	0.148	0.041	28%	.96	.96	0.85
Panel D: Selection is on true and observed achievement history						
VAM1	0.148	0.166	112%	.64	.63	0.52
VAM2	0.148	0.089	60%	.86	.85	0.70
VAM4	0.148	0.078	53%	.89	.88	0.73
Panel E: Selection on unobservables is like selection on observables						
VAM1	0.148	0.212	143%	.57	.56	0.49
VAM2	0.148	0.140	95%	.73	.71	0.59
VAM4	0.148	0.147	99%	.71	.70	0.58
Panel F: Selection conforms to VAM assumptions, but effects of interest are those on the following year's score						
VAM1	0.118	0.148	125%	.42	.40	0.38
VAM2	0.110	0.147	133%	.33	.32	0.34

*Notes.* Estimates in column (1) are taken from the rich specification for reading in Table VI (Panels A–E) and from columns (2) and (4) of Table VIII (Panel F). Column (2) is from Table VI, columns (2) and (4) in Panel A, and is computed from the models reported in Table VIII in Panel F. In Panels B–E, estimates from Table 10 of Rothstein (2009) are used, with an adjustment for the different test scale used here. See the Online Appendix for details. Columns (4)–(6) are computed by drawing 10,000 teachers from normal distributions with the standard deviations described in columns (1) and (2). Estimates of the correlation between teachers' true effects and the bias in their estimated effects (–.33 for VAM 1 and –.43 for VAM2) are used in Panel A. In Panels B–E, this correlation is constrained to zero. In Panel F, the estimated correlation is used again; this is –.38 for VAM1 and –.43 for VAM2. “Reliability of top quintile” in column (6) is the fraction of teachers whose true effects are in the top quintile who are estimated to be in the top quintile by the indicated VAM.

remain large, only about four-fifths of top-quintile teachers are judged to be so by the two VAMs.

Panel D allows even more unobserved information to be used in classroom assignments: I assume that the principal knows the

student's true achievement in grades 2–4. Now, even VAM4 is correlated less than .9 with teachers' true effects, and less than three-fourths of true top-quintile teachers get top-quintile ratings from any the VAMs. Finally, Panel E presents an extreme scenario corresponding to Altonji, Elder, and Taber's (2005) assumption that selection on unobservables is like selection on observables. This is not realistic, as principals cannot perfectly predict student achievement, but it provides a useful bound for the degree of bias that nonrandom classroom assignments might produce in VAM-based estimates. This bound is tight enough to be informative: Even in this worst case, the VAMs retain some signal, and VAM2 and VAM4 continue to classify correctly over half of top-quintile teachers.

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

Moreover, Table VII omits three major sources of error in VAM-based quality measures that would magnify the misclassification rates seen there. First, I have suppressed the role of sampling error that would inevitably arise in VAM-based estimates. It is well documented (Lockwood, Louis, and McCaffrey 2002; McCaffrey et al. 2009) that this alone produces high misclassification rates. Second, all of the analyses in this paper are based on comparisons of teachers within schools. As in most other value added studies, I make no effort to measure across-school differences in teacher quality. But most policy applications of value added would require comparisons across as well as within schools. Because students are not even approximately randomly assigned to schools, these comparisons are likely to be less informative about causal effects than are the within-school comparisons considered here.

Finally, I have assumed that teachers' effects on their students' end-of-grade scores are the sole outcome of interest. This may be incorrect. In particular, if teachers can allocate effort between teaching to the test and raising students' long-run learning

trajectories (e.g., by working to instill a love of reading), one would like to reward the second rather than the first. This suggests that the effects that matter may be those on students' long-run outcomes rather than on their end-of-grade scores. I consider this issue in the next section.

## VII. SHORT-RUN VS. LONG-RUN EFFECTS

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

When only a single grade's teacher assignment is included, VAM2 implicitly assumes that teachers' effects decay at a uniform, geometric rate ( $\beta_{g,g+s} = \beta_{gg}\lambda^s$  for  $\lambda \in [0, 1]$ ), whereas VAM1 assumes zero decay ( $\lambda = 0$ ). It is not clear that either restriction is reasonable.<sup>30</sup> Although several studies have estimated  $\lambda$ ,<sup>31</sup> all have done so under the restriction that decay is uniform. As a final investigation, I analyze the validity of this restriction by comparing a grade- $g$  teacher's initial effect in grade  $g$  with her longer-run effect on scores in grade  $g + 1$  or  $g + 2$ , without restricting the relationships among them.<sup>32</sup> If in fact teachers' effects decay uniformly, the initial and longer-run effects should be perfectly correlated (except for sampling error).

I begin by estimating VAM1 and VAM2 for third, fourth, and fifth grade scores or gains, augmenting each specification with controls for past teachers back to third grade. I then compute third

30. Although a full discussion is beyond the scope of this paper, assumptions about "decay" are closely related to issues of test scaling and content coverage (Martineau 2006; Rothstein 2008; Ballou 2009). To illustrate, consider a third grade teacher who focuses on addition and subtraction. This will raise her students' third grade scores but may do little for their performance on a fifth grade multiplication test.

31. See, for example, Sanders and Rivers (1996), Konstantopoulos (2007), and Andrabi et al. (2009).

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

TABLE VIII  
PERSISTENCE OF TEACHER EFFECTS IN VAMS WITH LAGGED TEACHERS

	VAM1		VAM2	
	Math	Reading	Math	Reading
	(1)	(2)	(3)	(4)
Cumulative effect of fourth grade teachers over two years				
Standard deviation of fourth grade teacher effects, adjusted				
On fourth grade scores	0.184	0.150	0.188	0.140
On fifth grade scores	0.108	0.118	0.118	0.110
Average persistence of fourth grade teacher's immediate effect one year later	0.269	0.325	0.320	0.262
Correlation (effect on fourth grade, effect on fifth grade), adjusted	.455	.413	.511	.334
Cumulative effect of third grade teachers over three years				
Standard deviation of third grade teacher effects, adjusted				
On third grade scores	0.218	0.172	0.209	0.167
On fourth grade scores	0.136	0.126	0.120	0.130
On fifth grade scores	0.185	0.199	0.129	0.147
Average persistence of third grade teacher's immediate effect two years later	0.335	0.394	0.277	0.394
Correlation (effect on third grade, effect on fifth grade), adjusted	.395	.341	.450	.447

*Notes.*  $N = 23,415$ . Sample is identical to that used in Table VI. Effects of fourth grade teachers on fifth grade scores and of third grade teachers on fourth and fifth grade scores are cumulative effects. For VAM1, the specification for gains in grade  $g$  includes controls for teachers in grades 3 through  $g$ , and the cumulative effect of the grade  $h$  teacher on the grade  $g$  gain is the sum of the effects in  $h, h + 1, \dots, g$ . For VAM2, the specification is augmented with controls for math and reading scores in grade  $g - 1$ . The calculation of cumulative effects is described in footnote 31. "Average persistence" is the coefficient from a regression of effects on fifth grade scores on effects on fourth (Panel A) or third (Panel B) scores, and indicates the expected effect on fifth grade scores for a teacher whose initial effect was +1. All standard deviations, correlations, and persistence parameters are adjusted for the influence of sampling error, as described in Online Appendix B.

and fourth grade teachers' cumulative effects over one, two, and (for third grade teachers) three years. Table VIII presents summary statistics for these cumulative effects. I show their standard deviation; the implied average persistence of teachers' first-year effects (computed as  $\lambda = \text{cov}(\beta_{44}, \beta_{45}) / \text{var}(\beta_{44})$ ); and the correlation between the initial and cumulative effects. All statistics are adjusted for sampling error in the  $\beta$  coefficients. Three aspects of the results are of note. First, there is much more variation in fourth grade teachers' effects on fourth grade scores than in those same teachers' effects on fifth grade scores. With uniform decay at rate  $(1 - \lambda)$ ,  $\text{var}(\beta_{g,g+s}) = \lambda^{2s} \text{var}(\beta_{gg})$ , so this is consistent with the mounting recent evidence that teachers' effects decay importantly in the year after contact (Kane and Staiger 2008; Andrabi et al. 2009; Jacob, Lefgren, and Sims forthcoming). Second, the average

persistence of fourth grade teachers' effects one year later is only around 0.3, again consistent with recent evidence.<sup>33</sup> Third, the data are not even approximately consistent with the notion that this persistence rate is uniform across teachers: The correlation between teachers' first-year effects and their two year cumulative effects is much less than one, ranging between .33 and .51 depending on the model and subject. Three-year cumulative effects show a similar pattern, correlated around .4 with the immediate effect. Even if we assume that the VAM-based estimates can be treated as causal, a teacher's first-year effect is a poor proxy for his or her longer-run impact.

The final panel of Table VII explores the implications of this analysis for teacher quality measurement. I use the estimates in Table VIII as parameters for my simulation to compare traditional end-of-year VAM coefficients to teachers' longer-run (two-year) effects, treating the latter as the "truth." The results are not encouraging. Correlations are well below .5, and only about a third of teachers in the top quintile of the distribution of two-year cumulative effects are also in the top quintile of the one-year effect distribution. It is apparent that misspecification of the outcome variable produces extreme amounts of misclassification. Note, moreover, that this analysis assumes that the VAM1 and VAM2 exclusion restrictions are valid. A full account of the utility of VAMs for identifying good teachers would need to combine the analyses of lagged effects and endogenous classroom assignments. This would imply even higher rates of misclassification than are produced by either on its own.

## VIII. DISCUSSION

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

33. In other contexts, experiments have shown short-term effects on test scores that do not persist, as well as long-term effects on other outcomes (see, e.g., Schweinhart et al. [2005]). If teachers' effects had this form, we might wish to focus on short-run rather than long-run test score effects. But there is no direct evidence that teacher effects follow this pattern.

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

First, it is clear that an important priority in educational research should be to build richer VAMs that can accommodate dynamic sorting of students to classrooms. In contrast, there is little apparent need to allow for permanent heterogeneity in students' rates of growth. One approach might be to assume that classroom assignments depend on the principal's best prediction of students' unobserved ability, with predictions updated each year based on student grades and test scores. None of the VAMs considered here can accommodate assignments of this form, which on its face seems quite plausible, but approaches like those taken by Altonji, Elder, and Taber (2005) and Rothstein (2009) may be useful.

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

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

The present analysis also has implications beyond the specific application to measuring teacher productivity. Estimates of the

quality of schools and of the effects of firms on workers' wages use identical econometric models, and rely on similar exclusion restrictions. Evidence about the "effects" of future schools and employers on current outcomes would be informative about the validity of both sets of estimates.

Finally, the results here have important implications for the use of existing VAMs in education policy. My results indicate that policies based on these VAMs will reward or punish teachers who do not deserve it and fail to reward or punish teachers who do. The literature on pay-for-performance suggests some consequences of this result. First, and most clearly, the stakes attached to VAM-based measures should be relatively small. Baker (1992, 2002) considers a performance measure that is less than perfectly correlated with the worker's contribution to firm output. He notes that high-stakes compensation will create incentives for workers to direct excess effort to the unproductive component of the performance measure. In education, this might take the form of teachers lobbying their principals to be assigned the "right" students who will yield predictably high value added scores. In Baker's model, misallocation of effort can be kept to a tolerable level by keeping the variable component of compensation small.<sup>34</sup> Another argument for low stakes in VAM-based compensation is provided by Hölmstrom and Milgrom (1991), who discuss implications of the results presented in Section VII above: If short-term test scores are poor proxies for the dimensions of achievement that really matter, it may be better to forgo or limit incentive pay rather than encourage excessive teaching to the test.

A second and more speculative suggestion is that VAM-based estimates should be used as only one among several inputs into an accountability system that also incorporates principals' subjective ratings (see, e.g., Baker, Gibbons, and Murphy [1994]). There are two reasons for this. First, principals may have information about the direction of the bias in a particular teacher's VAM-based estimate that is not otherwise available to the econometrician, so incorporation of their opinions might lead to better-targeted incentives (Hölmstrom 1979). Second, use of the VAM as the sole basis for teacher compensation and/or retention would permit principals to reward or punish teachers only through the assignment of desirable or undesirable students. Anecdotally, this

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

is an important management tool for principals, who may induce disfavored teachers to resign by assigning them difficult students. But there is evidence that teacher–student matching is an important determinant of student learning (Dee 2005; Clotfelter, Ladd, and Vigdor 2006), so manipulation of matches can have real efficiency consequences. If the principal’s subjective judgment is incorporated directly into the incentive scheme, he or she will be able to allocate students to teachers to maximize output without sacrificing his or her ability to influence rewards and sanctions. Of course, this suggestion presumes high-quality principals who have enough time to observe teachers’ classrooms and enough training to distinguish good from bad teachers. Without this, neither subjective evaluations nor VAM-based estimates that depend importantly on classroom assignments are likely to provide much useful information.

GOLDMAN SCHOOL OF PUBLIC POLICY, UNIVERSITY OF CALIFORNIA, BERKELEY, AND  
NATIONAL BUREAU OF ECONOMIC RESEARCH

#### REFERENCES

- Aaronson, Daniel, Lisa Barrow, and William Sander, “Teachers and Student Achievement in the Chicago Public High Schools,” *Journal of Labor Economics*, 25 (2007), 95–135.
- Abowd, John M., and Francis Kramarz, “The Analysis of Labor Markets Using Matched Employer-Employee Data,” in *Handbook of Labor Economics*, Vol. 3B, Orley C. Ashenfelter and David Card, eds. (Amsterdam: North-Holland, 1999).
- Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools,” *Journal of Political Economy*, 113 (2005), 151–184.
- Anderson, T. W., and Cheng Hsiao, “Estimation of Dynamic Models with Error Components,” *Journal of the American Statistical Association*, 76 (1981), 598–609.
- Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc, Do Value-Added Estimates Add Value? Accounting for Learning Dynamics, unpublished manuscript, Harvard, 2009.
- Arellano, Manuel, and Stephen Bond, “Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations,” *Review of Economic Studies*, 58 (1991), 277–297.
- Ashenfelter, Orley, “Estimating the Effect of Training Programs on Earnings,” *Review of Economics and Statistics*, 60 (1978), 47–57.
- Ashenfelter, Orley, and David Card, “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs,” *Review of Economics and Statistics*, 67 (1985), 648–660.
- Baker, George P., “Incentive Contracts and Performance Measurement,” *Journal of Political Economy*, 100 (1992), 598–614.
- , “Distortion and Risk in Optimal Incentive Contracts,” *Journal of Human Resources*, 37 (2002), 728–751.
- Baker, George P., Robert Gibbons, and Kevin J. Murphy, “Subjective Performance Measures in Optimal Incentive Contracts,” *Quarterly Journal of Economics*, 109 (1994), 1125–1156.

- Ballou, Dale, "Test Scaling and Value-Added Measurement," *Education Finance and Policy*, 4 (2009), 351–383.
- Boardman, Anthony E., and Richard J. Murnane, "Using Panel Data to Improve Estimates of the Determinants of Educational Achievement," *Sociology of Education*, 52 (1979), 113–121.
- Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff, "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools," Center for Analysis of Longitudinal Data in Education Research, Working Paper 10, 2007.
- Braun, Henry I., "Using Student Progress To Evaluate Teachers: A Primer on Value-Added Models," ETS Policy Information Center, Manuscript, 2005a.
- , "Value-Added Modeling: What Does Due Diligence Require?" in *Value Added Models in Education: Theory and Applications*, Robert W. Lissitz, ed. (Maple Grove, MN: JAM Press, 2005b).
- Card, David, and Daniel Sullivan, "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56 (1988), 497–530.
- Chamberlain, Gary, "Panel Data," in *Handbook of Econometrics*, Vol. II, Z. Griliches and M. D. Intriligator, eds. (Amsterdam: Elsevier North-Holland, 1984).
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor, "Teacher–Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources*, 41 (2006), 778–820.
- Dee, Thomas S., "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review*, 95 (2005), 158–165.
- Goldhaber, Dan "Everyone's Doing It, but What Does Teacher Testing Tell Us about Teacher Effectiveness?" *Journal of Human Resources*, 42 (2007), 765–794.
- Harris, Douglas N., and Tim R. Sass, Value-Added Models and the Measurement of Teacher Quality, unpublished manuscript, 2006.
- , What Makes for a Good Teacher and Who Can Tell? unpublished manuscript, 2007.
- Heckman, James J., V. Joseph Hotz, and Marcelo Dabos, "Do We Need Experimental Data to Evaluate the Impact of Manpower Training on Earnings?" *Evaluation Review*, 11 (1987), 395–427.
- Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81 (1986), 945–960.
- Hölmstrom, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, 10 (1979), 74–91.
- Hölmstrom, Bengt, and Paul Milgrom "Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7 (1991), 24–52.
- Imbens, Guido W., and Thomas Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142 (2008), 615–635.
- Jacob, Brian A., and Lars Lefgren, "What Do Parents Value in Education? An Empirical Examination of Parents' Revealed Preferences for Teachers," *Quarterly Journal of Economics*, 122 (2007), 1603–1637.
- , "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education," *Journal of Labor Economics*, 25 (2008), 101–136.
- Jacob, Brian A., Lars Lefgren, and David Sims, "The Persistence of Teacher-Induced Learning Gains," *Journal of Human Resources*, forthcoming.
- Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan, "Earnings Losses of Displaced Workers," *American Economic Review*, 83 (1993), 685–709.
- Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger, "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City," *Economics of Education Review*, 27 (2008), 615–631.
- Kane, Thomas J., and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," National Bureau of Economic Research Working Paper No. 14607, 2008.
- Kezdi, Gabor, "Robust Standard Error Estimation in Fixed Effects Panel Models," *Hungarian Statistical Review*, 9 (2004), 95–116.
- Kinsler, Josh, Estimating Teacher Value-Added in a Cumulative Production Function, unpublished manuscript, University of Rochester, 2008.

- Koedel, Cory, and Julian R. Betts, "Re-Examining the Role of Teacher Quality in the Educational Production Function," University of Missouri Department of Economics, Working Paper 07-08, 2007.
- Konstantopoulos, Spyros, "How Long Do Teacher Effects Persist?" IZA Discussion Paper No. 2893, 2007.
- Lazear, Edward P., "Pay Equality and Industrial Politics," *Journal of Political Economy*, 97 (1989), 561–580.
- Lockwood, J. R., Thomas A. Louis, and Daniel F. McCaffrey, "Uncertainty in Rank Estimation: Implications for Value-Added Modeling Accountability Systems," *Journal of Educational and Behavioral Statistics*, 27 (2002), 255.
- Martineau, Joseph A., "Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability," *Journal of Educational and Behavioral Statistics*, 31 (2006), 35–62.
- McCaffrey, Daniel F., J. R. Lockwood, Daniel M. Koretz, and Laura S. Hamilton, "Evaluating Value-Added Models for Teacher Accountability," RAND, Report, 2003.
- McCaffrey, Daniel F., Tim R. Sass, J. R. Lockwood, and Kata Mihaly, "The Intertemporal Stability of Teacher Effect Estimates," *Education Finance and Policy*, 4 (2009), 572–606.
- Milgrom, Paul R., "Employment Contracts, Influence Activities, and Efficient Organization Design," *Journal of Political Economy*, 96 (1988), 42–60.
- Monk, David H., "Assigning Elementary Pupils to Their Teachers," *Elementary School Journal*, 88 (1987), 167–187.
- Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges, "How Large Are Teacher Effects?" *Educational Evaluation and Policy Analysis*, 26 (2004), 237–257.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica*, 73 (2005), 417–458.
- Rosenbaum, Paul R., and Donald B. Rubin, "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79 (1984), 516–524.
- Rothstein, Jesse, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Princeton University Education Research Section, Working Paper 25, 2008.
- , "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy*, 4 (2009), 537–571.
- Sanders, William L., and June C. Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," University of Tennessee Value-Added Research and Assessment Center, Research Progress Report, 1996.
- Sanders, William L., Arnold M. Saxton, and Sandra P. Horn, "The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment," in *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* Jason Millman, ed. (Thousand Oaks, CA: Corwin, 1997).
- Schweinhart, L. J., J. Montie, Z. Xiang, W. S. Barnett, C. R. Belfield, and M. Nores, *Lifetime Effects: The High/Scope Perry Preschool Study Through Age 40* (Ypsilanti, MI: High/Scope Press, 2005).
- Todd, Petra E., and Kenneth I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113 (2003), F3–F33.
- Wainer, Howard, "Introduction to a Special Issue of the Journal of Educational and Behavioral Statistics on Value-Added Assessment," *Journal of Educational and Behavioral Statistics*, 29 (2004), 1–3.
- Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002).