

# Appendix:

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

Jesse Rothstein

Princeton University and NBER

May 2009

This document contains four appendices: A data appendix (A), a technical appendix (B), a discussion of Monte Carlo analyses of my proposed tests (C), and an appendix describing additional specifications that were not included in the main text (D).

### **A Data**

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

Each of my analyses uses subsets of this sample that have complete data on test scores and teacher assignments for enough grades to permit the analysis. A student might be excluded

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

Appendix Table A2 reports statistics on mixing of classrooms between 4th and 5th grades. This uses a somewhat different sample than other tables, consisting of all students with valid records and valid teacher matches in both grades 4 and 5 who did not switch schools or make abnormal progress between grades. Using this sample, I count the number of 4th grade classes at the school, and I compute for each student the fraction of her 5th grade classmates who were also in her 4th grade class. I average this over the full sample and over subsamples defined by the number of 4th grade teachers at the school. I also identify schools where dummies for the  $J_4$  4th grade teachers and  $J_5$  5th grade teachers have rank less than  $J_4 + J_5 - 1$ , indicating perfect collinearity of at least one teacher assignment with the others, and re-compute the statistic excluding observations from those schools.

## **B Technical Details**

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

### **B.1 School-level normalizations**

As discussed in the text, each of my regressions includes fixed effects for the school attended, and coefficients on teacher indicators are normalized to have mean zero at the school level. This normalization is easiest to describe if the sample consists of only a single school. Let  $T$  be an  $N$ -by- $J$  matrix of indicators for having been taught by each of the  $J$  teachers in a particular

grade at that school. Many of my regressions take the form

$$y = \alpha + T\beta + \varepsilon. \quad (\text{B.1})$$

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

$$V \equiv \begin{pmatrix} V_b & 0_J \\ 0_J' & 0 \end{pmatrix}, \quad (\text{B.2})$$

where  $0_J$  is a column vector of  $J$  zeros.

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

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

$$D = \begin{pmatrix} D_1 & 0 & \cdots & 0 \\ 0 & D_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & D_S \end{pmatrix}. \quad (\text{B.3})$$

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

## B.2 Sampling-adjusted standard deviations

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

The (weighted) variance of teachers' estimated effects is

$$\hat{\text{var}}(\hat{\theta}) = \frac{1}{J-S} \hat{\theta}' \text{diag}\{\bar{n}\} \hat{\theta}, \quad (\text{B.4})$$

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

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

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

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

$$\Delta A_{i3} = T_{i3}\pi_{33} + T_{i4}\pi_{43} + T_{i5}\pi_{53} + e_{3i3} \quad (\text{B.5})$$

$$\Delta A_{i4} = T_{i3}\pi_{34} + T_{i4}\pi_{44} + T_{i5}\pi_{54} + e_{3i4}. \quad (\text{B.6})$$

Two computational challenges arise. First, not all of the  $\pi$  coefficients can be separately computed. The particular problem arises because (as discussed below) I restrict the sample to students who do not change schools. The fitted values of the regressions would be unchanged were we to add a constant  $c$  to each element of the  $\pi_{gh}$  corresponding to a teacher at a particular

school  $j$  and subtract the same constant from the similarly-defined elements of  $\pi_{kh}$  for some  $k \neq g$ . As a result, the mean of  $\pi_{gh}$  across all teachers in grade  $g$  at school  $j$  cannot be separately identified. I augment (B.5) and (B.6) with school indicators, then select one teacher in each grade at each school to exclude from the regressions.<sup>1</sup> I treat the excluded  $\pi$  coefficient as zero, with sampling variance zero. After estimating the regression, I normalize the coefficients of (B.5) and (B.6) to have mean zero across teachers in each grade at each school, using the procedure described above.

The second issue derives from the sheer size of the regression. Even after excluding the over-identified coefficients, each of the  $T_{ig}$  vectors has over 2,200 elements, and the full regression (after dropping redundant indicators) has 5,501 regressors. Numerical inversion of a matrix of this dimension may introduce inaccuracies. By focusing on samples of students who do not switch schools I can simplify the computation. Re-order the independent variables in equations (B.5) and (B.6) as  $X = [X_{(1)}, X_{(2)}, \dots, X_{(j)}]$ , where  $X_{(j)}$  contains the indicator for school  $j$  and the indicators for all teachers (in all three grades) at school  $j$ . Any sample student who ever appears in school  $j$  never appears in any other school, so  $X'_{(j)}X_{(k)} = 0$  for all  $j \neq k$ . This ensures that  $X'X$  is block-diagonal:

$$X'X = \begin{pmatrix} X'_{(1)}X_{(1)} & 0 & \cdots & 0 \\ 0 & X'_{(2)}X_{(2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X'_{(j)}X_{(j)} \end{pmatrix}. \quad (\text{B.7})$$

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

$$(X'X)^{-1} = \begin{pmatrix} (X'_{(1)}X_{(1)})^{-1} & 0 & \cdots & 0 \\ 0 & (X'_{(2)}X_{(2)})^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (X'_{(j)}X_{(j)})^{-1} \end{pmatrix}. \quad (\text{B.8})$$

Each block has dimension of only a few dozen, so inversion is straightforward. The  $\pi$  coefficients (before the within-school normalization) and robust sampling variances are readily computed from  $(X'X)^{-1}$ . The covariances between the coefficients of equations (B.5) and (B.6)

---

<sup>1</sup>The sample used for these regressions excludes schools where, due to insufficient mixing, the  $[T_{i3} \ T_{i4} \ T_{i5}]$  submatrix corresponding to teachers at the school has rank less than  $J_{s3} + J_{s4} + J_{s5} - 2$ .

can be computed with

$$\text{cov}(\Pi^4, \Pi^5) = (X'X)^{-1} X' \text{diag}(\hat{e}_{i4}\hat{e}_{i5}) X (X'X)^{-1}. \quad (\text{B.9})$$

This implicitly clusters on the individual student, and is equivalent to applying system OLS (aka the seemingly unrelated regression estimator) to the simultaneous equations (B.5) and (B.6).

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

In a few cases (e.g. columns 5-8 of Table IV), I include continuous regressors  $Z$  along with the school and teacher indicators from several grades. These regressions have the form

$$y = X\Pi + Z\psi + \varepsilon. \quad (\text{B.10})$$

Letting  $W = [X \ Z]$  and  $\Lambda = [\Pi' \ \psi']'$ , we have  $y = W\Lambda + \varepsilon$ . Because the  $\psi$  coefficients are common across schools,  $W'W$  is no longer block-diagonal, and the school-by-school strategy described above cannot be used directly here. In these models, I use a brute-force OLS regression estimator (implemented in Matlab) to compute the regression of the school de-meaned  $y$  on the de-meaned  $W$ . This may introduce numerical inaccuracy in the estimated coefficients,  $\hat{\Lambda}$ . To avoid this, I use an iterative algorithm to obtain improved coefficient estimates. At each iteration  $t$  (beginning with  $t=1$ ), there are two steps:

1. Treat the  $\psi$  parameters as known, using values from the previous iteration,  $\hat{\psi}_{t-1}$ . Regress  $y - Z\hat{\psi}_{t-1}$  on  $X$ . The methods used in the previous section can be applied here, as  $X'X$  is block diagonal with blocks corresponding to schools. Label the resulting coefficients  $\hat{\Pi}_t$ .
2. Treating the  $\hat{\Pi}_t$  coefficients as known, regress  $y - X\hat{\Pi}_t$  on  $Z$ .  $Z$  typically contains only a handful of variables, so this is simple to calculate. Label the resulting coefficients  $\hat{\psi}_t$ , and use these as inputs to step 1 on the next iteration.

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

This is essentially the Gauss-Seidel method, though the structure of the problem makes it possible to use only two sub-vectors of the full parameter vector  $\Lambda$  rather than stepping through

each element of  $\Lambda$  separately as in typical implementations. It is a contraction mapping on the sum of squared errors, so the coefficients necessarily converge to the OLS coefficients. Abowd et al. (2002) use a similar (in spirit, though not in detail) computational strategy.

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

## **B.5 Simulation of VAM estimates with selection on unobservables**

Table VII of the main paper presents simulated comparisons between teachers' true effects and those obtained from potentially biased VAMs. The simulations are necessary because I wish to focus on the role of (asymptotic) bias in the VAM estimates, while estimates from finite samples are dominated by sampling error. The use of simulated data comes at a slight cost: I must assume a specific joint distribution for teachers' causal effects and the VAM-based estimates of these causal effects. I assume that these are jointly normally distributed. The standard deviations of each and the correlation between the two are parameters that are varied across panels.

The simulations in Panels A and F of Table VII are based on the estimates from Tables VI and VIII, respectively. In each of these tables, I treat the plim of the estimates from a particular specification as the "true" effects of interest, and I ask how other specifications compare to this. I compute the difference between the "true" effects and those from the "incorrect" specifications, then use the methods described above to estimate the standard deviation that would be observed for this difference were it free of sampling error. (This requires stacking the two specifications. Standard errors are clustered on the student.) These methods can be readily extended to estimate sampling-adjusted correlations. In Table VI, the bias in the VAM1 reading specification is correlated -0.33 with the VAM4 estimates; for VAM2, the correlation is -0.43. In Table VIII, the bias in 4th grade teachers' one-year effects relative to their effects on 5th grade scores is correlated -0.38 with the latter effects in Column 2. The correlation in Column 4 is -0.43.

The parameters in Panels B-E are computed differently. Rothstein (forthcoming) uses methods developed by Altonji et al. (2005) to estimate the bias in VAM-based estimates when selection into classrooms is based on unobservables as well as on observed variables. However, those estimates are based on a somewhat different sample than is used here, and on differently-scaled test scores. To make the estimates of the standard deviation of the bias in VAMs 1, 2, and 4 (called VAM2, VAM3, and VAM4 in Rothstein (forthcoming)) comparable to those used here, I rescale by the ratio of the standard deviation of the VAM-based estimates in Table VI (0.127,

0.121, and 0.148 for VAM1, VAM2, and VAM4, respectively) to the corresponding standard deviations (0.114, 0.106, and 0.100) in Table 10 of Rothstein (forthcoming). This yields the bias standard deviations listed in Column 2 of Table VII. I assume that the distribution of teachers’ true effects is the same as in Panel A, with a standard deviation of 0.148, and that bias is uncorrelated with the truth. There is little basis for the latter assumption, but it does not make much difference in the results: When I use a correlation of zero in Panel A, the results are identical.

Given estimates of the standard deviation of teachers’ true effects, the standard deviation of the “incorrect” VAM coefficients, and the correlation between the two, it is straightforward to simulate data from their joint distribution. I draw 10,000 teachers from this joint distribution. I use these simulated data to compute the correlation between true and VAM-based effects, the Spearman rank correlation, and the fraction of teachers who are actually in the top quintile who appear to be in the top quintile according to the VAM in question. Note that the symmetry of the normal distribution ensures that the misclassification rate for the bottom quintile is identical.

## C Monte Carlo analysis of VAM1 and VAM2 tests

To evaluate the performance of the VAM1 and VAM2 tests in samples resembling those used in Tables III and IV, I conducted extensive Monte Carlo analyses. I began with a simplified version of the general data generating process introduced in Section III of the main paper. Letting  $A_{ijsg}^*$  represent the true achievement at the end of grade  $g$  of student  $i$  in grade- $g$  classroom  $j$  in school  $s$ , I assume that:

$$A_{ijs3}^* = \xi_{s3} + \beta_{js3} + u_{ijs3} \quad (\text{C.1})$$

$$A_{ijs4}^* = A_{ijs3}^* + \xi_{s4} + \beta_{js4} + u_{ijs4}. \quad (\text{C.2})$$

Here,  $\xi_{sg}$  is a school-by-grade effect,  $\beta_{jsg}$  is the contemporaneous effect of the grade- $g$  teacher, and  $u_{ijsg}$  is an error term, independent across students in the same classroom but not necessarily across grades for the same student. Observed achievement equals true achievement plus random measurement error:  $A_{ijsg} = A_{ijsg}^* + e_{ijsg}$ , with  $e_{ijsg}$  independent across grades and across students. All terms are assumed to be normally distributed, potentially correlated across grades but with no correlation between terms (between  $\beta$  and  $\xi$ , for example).

I simulate three sample configurations, two rules for the assignment of students to 5th grade classrooms, and seven sets of parameters governing the data generating process. All sample configurations involve 400 schools. In the first, there are 2 classrooms per grade per school

and 20 students per classroom. In the second, there are still 2 classrooms but 100 students per classroom. In the third, there are 4 classrooms, each with 20 students.

My first rule assigns students randomly to 5th grade classrooms. The second makes assignments random conditional on  $A_{ijs4}$ . To implement this, I compute the student's percentile rank (scaled to range from 0 to 1) in the 4th grade score distribution. I then add to this a random number, uniform on  $[0, 1]$ . Within each school, students are sorted on the resulting sum, with the top 20 or 100 assigned to one classroom, the next to the next classroom, and so on. In both rules, students are assigned to 4th grade classrooms so as to create moderate collinearity between the 3rd and 4th grade classroom assignments: Half of the students in each 3rd grade class are streamed into the same 4th grade class, while the remaining students are randomly assigned to 4th grade classes within the school.

The seven sets of parameter values are as indicated in Table C1. Each entry reports the standard deviation of the grade 3 term, the standard deviation of the grade 4 term, and the correlation between them. For example, the upper left entry indicates that in the baseline specification,  $SD(\xi_{s3}) = 0.5$ ,  $SD(\xi_{s4}) = 0.25$ , and  $corr(\xi_{s3}, \xi_{s4}) = 0.8$ .

The baseline scenario is chosen to resemble plausible parameter values. Scenario U turns off all between-classroom variation, while Scenario T rules out any within-classroom heterogeneity. Scenarios M1-M4 are intermediate cases in which the magnitude of teachers' effects is gradually reduced from 0.8 to an empirically plausible 0.1. For each combination of sample configuration, assignment rule, and data generating process, I generated 500 or 1000 simulated samples. With each sample, I applied the VAM1 and VAM2 tests, as in columns 3 and 4 of Tables III and IV. Appendix Table C2 reports the fraction of the resulting p-values that were below 0.05.

The first two rows shows that the VAM1 and VAM2 tests have size around 10% for the first sample configuration when students are randomly assigned to 5th grade classrooms within schools, regardless of the DGP. This is not ideal, but neither is it extremely severe. A decision rule that rejects the null hypothesis only if the F test's p-value is below 0.025 would have size around 5% across all seven DGPs. Note that most of the p-values in Tables III and IV are far below this threshold. Using the estimates in the "baseline" simulation, I compute empirical p-values for the tests in Columns 3 and 4 of Table III. These are 0.031 and 0.004, respectively.<sup>2</sup>

Rows 4-6 show that the over-rejection is eliminated when the sample includes sufficient numbers of students per class and ameliorated (particularly for VAM2) to a lesser degree when the sample includes more than two classrooms per school. Evidently, the over-rejection in the

---

<sup>2</sup>That is, 3.1% of the simulation draws for the baseline DGP, using the first sample configuration and random assignment of students to classrooms, yielded "p-values" (based on the F distribution) for VAM1 smaller than the 0.016 reported in Column 3 of Table III.

first rows is largely a function of the small number of students per classroom.

The second panel shows rejection rates when the assignment to 5th grade classrooms is random only conditional on the 4th grade score. The VAM1 test, appropriately, rejects 100% of the time in this case, regardless of the DGP or the sample size. In the Baseline, U, and M4 DGPs, VAM2 performs about as well with conditional random assignment as with true random assignment. Assuming the “baseline” DGP and random assignment conditional on the lagged score, I compute simulated p-values for the tests in Columns 3 and 4 of Table IV of 0.086 and 0.018, respectively. In DGPs where the teacher effects are more prominent, however, VAM2 rejects at very high rates even with conditional random assignment, 100% of the time in T and M1. The problem arises because the 4th grade teachers’ effects introduce extreme clustering of 4th grade gains that is not accounted for by the variance-covariance matrix used for the test. Fortunately, with more realistic data generating processes clustering does not appear to be a problem and rejection rates are in the 5-10% range. Note also that the specifications in the final columns of Table IV, which include controls for 4th grade teachers, would absorb the clustering that is the source of the problem here. Additional Monte Carlo estimates, not reported in Table C1, indicate that these specifications bring the size of the M1, M2, and M3 tests down to levels similar to those seen in the M4 specification.

## **D Additional Specifications**

### **D.1 Teachers’ observable characteristics**

VAMs are used not only to estimate individual teachers’ effects, but also to assess the relationship of teacher quality with teachers’ observed characteristics (see, e.g., Clotfelter et al., 2006, 2007; Goldhaber and Brewer, 1997; Hanushek and Rivkin, 2006). These analyses replace the teacher indicators in VAM1, VAM2, or VAM3 with vectors of teacher observables – education, experience, etc. The tests developed in the main text can be applied to these models as well. Appendix Table D1 presents results for mathematics. (Results for reading are similar and are available from the author.) I focus on a short vector of teacher characteristics: An indicator for whether the teacher has a master’s degree, a linear experience measure, an indicator for whether the teacher has less than two years of experience, and the teacher’s score on the Praxis tests required to obtain elementary certification in North Carolina.<sup>3</sup> As in the other analyses, I restrict attention to students who can be assigned to valid teachers in each grade for which teacher char-

---

<sup>3</sup>Each test is standardized among North Carolina teachers who took it in the same year, then (when multiple scores are available) scores are averaged across tests.

acteristics will be controlled and who do not switch schools between grades. I further exclude students for whom I am unable to assemble complete characteristics for each of the relevant teachers.

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

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

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

Columns 5 and 6 present the correlated random effects analysis that I use to evaluate VAM3, modeling 3rd and 4th grade gains, respectively, as functions of the characteristics of teachers in grades 3 through 5. I consider two restricted models, one that constrains student ability to enter identically into each grade's gain score equation and another that allows different ability coefficients in different grades. The former model – corresponding to the version of VAM3 that is uniformly used in the literature – implies that the 5th grade teacher coefficients in columns 5 and 6 of Table VI should be equal. In fact, we see a significant negative coefficient for the no experience indicator in the model for 4th grade gains and a marginally significant ( $t=1.67$ ) positive coefficient in the model for 3rd grade gains. The hypothesis of equal effects is decisively rejected ( $p=0.02$ ). The less restrictive model requires only that the coefficients in columns 5 and 6 be proportional to one another. This restriction is consistent with the data ( $p=0.81$ ). However, the OMD estimates indicate a factor of proportionality of  $-0.92$ . If we normalize  $\tilde{\tau}_3 = 1$ , defining “ability” to have a positive effect on 3rd grade gains, the model indicates that high ability

students gain much *less* during 4th grade than their low ability peers.

An alternative interpretation of this extremely counterintuitive result is that the test is unable to detect violations of strict exogeneity in this context. The correlated random effects test has power against violations of strict exogeneity only if classroom assignments depend on factors that are correlated with the included variables. As all of the coefficients except those for the inexperienced teacher indicator are small and far from statistically significant, and as even the inexperienced teacher coefficients are consistent with the model only with implausible coefficient estimates, the simplest interpretation is that VAM3 is poorly suited to identifying the effects of teacher characteristics on student achievement. Indeed, when I extend the analysis to use the characteristics of 6th grade teachers – students are typically in middle school in 6th grade, and ability tracking is more pronounced – to strengthen the overidentification test (see Rothstein, 2008), I reject proportionality of the 6th grade teacher coefficients.

Despite the violations of the VAM identifying assumptions seen in Appendix Table D1, qualitative conclusions about the importance of teacher characteristics do not appear to be very sensitive to the inclusion of controls for observables. Appendix Table D2 presents estimates of the effects of 5th grade teacher characteristics on 5th grade gains, with varying control variables. Estimates change notably between Column 1 (VAM1, without controls) and Column 2 (VAM2, controlling for the 4th grade score). Changes are smaller but still non-trivial as we move across the remaining columns, which gradually add controls for more lagged scores, for teacher characteristics in earlier grades, for earlier classroom assignments, and eventually for the amount of time that the student reports having spent watching TV in 4th grade and the number of absences from school in that grade. In the math specifications, the Praxis effect is about 50% larger in Column 7 than in Column 2. In the reading specification, the effect of an inexperienced teacher declines by about 20%.

## **D.2 Distinguishing between teacher and classroom effects using cross-cohort comparisons**

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

Let  $\beta_{ty}$  be the effect of being taught by teacher  $t$  in year  $y$ . (I suppress grade subscripts for notational simplicity.) We can decompose this into a permanent component associated with

the teacher,  $\theta_t$ , and a time-varying component  $v_{ty}$  associated with transitory aspects of the classroom in year  $y$ :  $\beta_{ty} = \theta_t + v_{ty}$ . If we assume that the non-random assignments of students to classrooms are completely transitory – that the pre-assignment characteristics of students in the teacher’s classroom in year  $t$  uncorrelated both with the characteristics of students in year  $t + 1$  and with the teacher’s true effect  $\theta_t$  – then the bias in  $\hat{\beta}_{ty}$  will be uncorrelated from one year to the next. A decomposition of  $\hat{\beta}$  into permanent and transitory components – a regression of  $\hat{\beta}_{ty}$  onto teacher indicators – would yield unbiased estimates of the permanent teacher components  $\theta_t$ . Alternatively, the variance of  $\theta_t$  across teachers can be estimated from the between-year covariance of  $\beta$ :

$$E [\beta_{ty}\beta_{t,y+1}] = E [\theta_t^2] + E [v_{ty}v_{t,y+1}] + E [\theta_tv_{ty}] + E [\theta_tv_{t,y+1}]. \quad (\text{D.1})$$

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

This strategy relies crucially on the assumption that the assignments are uncorrelated across years. If some teachers are repeatedly assigned students with high expected gains that are not controlled in the VAM, this will create a covariance between  $v_{ty}$  and  $v_{t,y+1}$ , and therefore will bias the estimates of  $\theta_t$  and  $E [\theta_t^2]$ . To evaluate whether assignments are in fact uncorrelated across years, I used students who were in 5th grade in 2000 to estimate a regression of 5th grade gains on all prior scores, absorbing 5th grade classroom indicators. This resembles the rich VAM4 from Section VI, but it excludes classroom indicators from prior grades. Using the coefficients from this regression, I form predicted 5th grade gains for each 5th grade student in both 2000 and 2001, excluding the classroom effects, then average these to the classroom level. These mean predicted gains represent bias in single-cohort estimates of VAM1. I also residualize the predicted gains against 4th grade scores to obtain the bias in VAM2. I then correlate the average predicted gains (or residualized predicted gains) of a teacher’s students in 2000 with those for the same teacher’s students in 2001.

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

## References

- Abowd, John M., Robert H. Creecy, and Francis Kramarz**, “Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data,” March 2002. Unpublished manuscript.
- Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber**, “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools,” *Journal of Political Economy*, February 2005, 113 (1), 151–184.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor**, “Teacher-Student Matching and the Assessment of Teacher Effectiveness,” *Journal of Human Resources*, Fall 2006, 41 (4), 778–820.
- , —, and —, “How and Why Do Teacher Credentials Matter for Student Achievement?,” Working paper 12828, National Bureau of Economic Research January 2007.
- Goldhaber, Dan and Dominic J. Brewer**, “Why Don’t Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity,” *Journal of Human Resources*, Summer 1997, 32 (3), 505–523.
- Hanushek, Eric A. and Steven G. Rivkin**, “Teacher Quality,” in Eric A. Hanushek and Finis Welch, eds., *Handbook of the Economics of Education*, Vol. 2, Amsterdam: Elsevier North-Holland, 2006, pp. 2–28.
- , **John F. Kain, Daniel M. O’Brien, and Steven G. Rivkin**, “The Market for Teacher Quality,” February 2005.
- Kane, Thomas J. and Douglas O. Staiger**, “Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation,” July 21, 2008. Unpublished manuscript.
- Rothstein, Jesse**, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” Working Paper 25, Princeton University Education Research Section, 2008.
- , “Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables,” *Education Finance and Policy*, forthcoming, 0.

**Appendix Table A1. Construction of analytical samples**

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

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

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

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

**Appendix Table C1. Parameters for Monte Carlo simulations**

<b>Scenario</b>	<b>School effects (<math>\xi</math>)</b>	<b>Teacher effects (<math>\beta</math>)</b>	<b>True residual (<math>u</math>)</b>	<b>Measurement error (<math>e</math>)</b>
Baseline	(0.5, 0.25, 0.8)	(0.1, 0.1, -0.5)	(0.8, 0.4, -0.25)	(0.2, 0.2, 0)
U	(0, 0, 0)	(0, 0, 0)	(0.8, 0.8, 0)	(0, 0, 0)
T	(0, 0, 0)	(0.8, 0.8, 0)	(0, 0, 0)	(0, 0, 0)
M1	(0, 0, 0)	(0.8, 0.8, 0)	(0.8, 0.8, 0)	(0.2, 0.2, 0)
M2	(0, 0, 0)	(0.4, 0.4, 0)	(0.8, 0.8, 0)	(0.2, 0.2, 0)
M3	(0, 0, 0)	(0.2, 0.2, 0)	(0.8, 0.8, 0)	(0.2, 0.2, 0)
M4	(0, 0, 0)	(0.1, 0.1, 0)	(0.8, 0.8, 0)	(0.2, 0.2, 0)

Notes: Each cell contains a triplet, describing the standard deviation of the random variable in the 3rd grade score equation, the standard deviation in the 4th grade score equation, and the correlation between the two. The data generating process is as in equations C1 and C2:  $A_3^* = \xi_3 + \beta_3 + u_3$ ,  $A_4^* = A_3^* + \xi_4 + \beta_4 + u_4$ ,  $A_3 = A_3^* + e_3$ , and  $A_4 = A_4^* + e_4$ .

**Appendix Table C2. Monte Carlo analysis of rejection rates with VAM1 and VAM2 tests**

	Production function						
	Baseline	U	T	M1	M2	M3	M4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>Random assignment</i>							
400 schools, 2 classes per school, 20 students per class (2500 replications)							
VAM1	9.1%	9.8%	10.1%	9.7%	9.6%	9.3%	9.2%
VAM2	9.6%	9.3%	10.2%	9.7%	9.3%	9.4%	9.8%
400 schools, 2 classes per school, 100 students per class (2000 replications)							
VAM1	5.1%	4.9%	5.4%	5.3%	4.9%	4.7%	4.8%
VAM2	5.7%	5.2%	5.1%	5.7%	5.7%	5.9%	5.7%
400 schools, 4 classes per school, 20 students per class (1000 replications)							
VAM1	7.5%	8.1%	11.5%	8.2%	8.9%	8.5%	7.9%
VAM2	7.1%	7.4%	9.6%	7.7%	7.2%	7.1%	7.0%
<i>Random assignment conditional on G4 score</i>							
400 schools, 2 classes per school, 20 students per class (2500 replications)							
VAM1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
VAM2	10.3%	9.2%	100.0%	100.0%	52.4%	12.2%	9.9%
400 schools, 2 classes per school, 100 students per class (2000 replications)							
VAM1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
VAM2	5.9%	4.8%	100.0%	100.0%	100.0%	16.4%	5.6%
400 schools, 4 classes per school, 20 students per class (1000 replications)							
VAM1	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
VAM2	8.2%	6.8%	100.0%	100.0%	54.1%	9.2%	6.7%

Notes: See Appendix C and Table C1 for descriptions of the simulation and of the data generating process. Under the null hypothesis that the test's size is 5%, a 95% confidence interval for the rejection rate in 2,500 replications is [4.2, 5.9]. With 2,000 replications, the confidence interval is [4.0, 6.0]; with 1,000 replications, it is [3.7, 6.4]. Rejection rates that lie outside of these intervals are shown in bold.

**Appendix Table D1. Tests of models for the effects of teacher observable characteristics on math gains**

	VAM1		VAM2		VAM3 (correlated random effects)	
	5th grade	4th grade	5th grade	3rd grade	3rd grade	4th grade
	(1)	(2)	(3)	(4)	(5)	(6)
<b>5th grade teacher</b>						
MA degree	-0.05 (1.30)	-1.49 (0.99)	-0.75 (1.30)	-0.74 (0.90)	2.20 (1.43)	-1.12 (1.04)
Experience	0.09 (0.07)	0.05 (0.05)	0.07 (0.07)	0.06 (0.05)	-0.04 (0.07)	-0.02 (0.05)
1(Experience < 2)	<b>-5.35</b> (1.88)	-2.87 (1.55)	<b>-5.95</b> (1.84)	-2.02 (1.41)	3.65 (2.19)	<b>-4.13</b> (1.61)
Praxis score	1.50 (0.80)	<b>1.32</b> (0.61)	<b>2.26</b> (0.77)	0.41 (0.54)	-1.03 (0.82)	1.03 (0.62)
<b>4th grade teacher</b>						
MA degree	-1.93 (1.30)	2.83 (1.53)	-1.19 (1.12)	1.92 (1.23)	0.67 (1.33)	<b>3.25</b> (1.62)
Experience	-0.10 (0.07)	0.07 (0.08)	-0.09 (0.06)	0.05 (0.06)	-0.07 (0.07)	0.13 (0.08)
1(Experience < 2)	<b>5.21</b> (1.75)	<b>-5.77</b> (2.00)	<b>3.76</b> (1.57)	<b>-3.96</b> (1.66)	1.06 (2.09)	<b>-5.89</b> (2.16)
Praxis score	-1.48 (0.76)	<b>2.18</b> (0.89)	-0.72 (0.65)	1.29 (0.72)	0.17 (0.81)	<b>2.53</b> (0.94)
<b>3rd grade teacher</b>						
MA degree					0.25 (1.91)	0.72 (1.44)
Experience					0.18 (0.11)	<b>-0.16</b> (0.08)
1(Experience < 2)					-0.58 (3.05)	-1.04 (2.24)
Praxis score					0.34 (1.07)	-0.05 (0.80)
<b>4th grade scores (*100)</b>						
Math			<b>0.69</b> (0.01)	<b>0.64</b> (0.01)		
Reading			<b>0.21</b> (0.01)	<b>0.22</b> (0.01)		
N	20,251	20,251	20,251	20,251	18,239	18,239
R2	0.147	0.142	0.264	0.278	0.105	0.145
p-value, G5 teacher coeffs. = 0	<0.01	0.02	<0.01	0.11	0.13	0.04
Restricted specification, G5 teacher effects are equal in G3, G4 models						
p-value					0.02	
Restricted specification, G5 teacher effects are proportional in G3, G4 models						
Ratio, effect on G4 to effect on G3					-0.92	
p-value for overid. test					0.81	

Note: Dependent variables are math gain scores (Columns 1, 2, 5, and 6) or level scores (Columns 3-4) in the relevant grade, multiplied by 100. Standard errors are clustered on the school. Bold coefficients are significant at the 5% level.

**Appendix Table D2. Sensitivity of effects of teacher observable characteristics to VAM controls**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>Panel A: Dependent variable is 5th grade math gain</i>							
G5 teacher: MA degree	-0.10 (0.78)	-0.84 (0.75)	-0.80 (0.71)	-1.13 (0.77)	-1.03 (0.90)	-0.88 (1.22)	-0.84 (1.23)
G5 teacher: Experience	0.02 (0.04)	0.04 (0.04)	0.03 (0.04)	0.04 (0.04)	0.07 (0.05)	0.00 (0.07)	0.00 (0.07)
G5 teacher: Experience < 2	<b>-3.75</b> (1.14)	<b>-4.94</b> (1.08)	<b>-5.11</b> (1.05)	<b>-5.38</b> (1.16)	<b>-4.33</b> (1.28)	<b>-6.00</b> (1.86)	<b>-5.93</b> (1.89)
G5 teacher: Praxis score	0.62 (0.49)	<b>1.33</b> (0.45)	<b>1.34</b> (0.43)	<b>1.24</b> (0.47)	<b>1.49</b> (0.55)	<b>1.87</b> (0.74)	<b>1.93</b> (0.74)
Controls for:							
G4 score, same subj.		y					
G3-G4 scores, both subj.			y				
G2-G4 scores, both subj.				y	y	y	y
G3-G4 teacher characteristics					y		
G3-G4 teacher dummies						y	y
G4 TV watching & absences							y
N	48,753	48,753	48,753	43,170	27,741	43,042	42,715
p, all 5th grade teacher coeffs = 0	0.002	<0.001	<0.001	<0.001	<0.001	0.001	0.001
<i>Panel B: Dependent variable is 5th grade reading gain</i>							
G5 teacher: MA degree	-1.15 (0.81)	<b>-1.56</b> (0.73)	-0.79 (0.64)	-1.17 (0.67)	-1.31 (0.79)	-1.78 (1.11)	-1.84 (1.12)
G5 teacher: Experience	0.08 (0.04)	<b>0.11</b> (0.04)	<b>0.09</b> (0.03)	<b>0.10</b> (0.04)	<b>0.11</b> (0.04)	<b>0.14</b> (0.06)	<b>0.14</b> (0.06)
G5 teacher: Experience < 2	<b>-3.63</b> (1.12)	<b>-4.62</b> (1.01)	<b>-4.30</b> (0.91)	<b>-4.16</b> (0.98)	<b>-3.25</b> (1.25)	<b>-3.65</b> (1.79)	<b>-3.71</b> (1.82)
G5 teacher: Praxis score	0.31 (0.48)	<b>0.92</b> (0.42)	0.47 (0.37)	0.53 (0.40)	0.45 (0.47)	0.68 (0.71)	0.66 (0.72)
N	48,753	48,753	48,753	43,170	27,741	43,042	42,715
p, all 5th grade teacher coeffs = 0	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	0.001

Notes: All specifications include school indicators, and all are clustered on the school. "G3-G4 teacher dummies" are indicators for each possible sequence of 3rd and 4th grade classrooms. Sample sizes vary due to availability of independent variables. Bold coefficients are significant at the 5% level.