

Student sorting and bias in value added estimation: Selection on observables and unobservables

Jesse Rothstein*

June 5, 2008

Abstract

Non-random assignment of students to teachers can bias value added estimates of teachers' causal effects. Rothstein (2008) shows that typical value added models indicate large counter-factual effects of 5th grade teachers on students' 4th grade learning, implying that assignments do not satisfy the imposed assumptions. This paper quantifies the resulting biases in estimates of 5th grade teachers' causal effects from several value added models, under varying assumptions about the assignment process. Under selection on observables, models for gain scores without controls or with only a single lagged score control are subject to important bias, but models with controls for the full test score history are nearly free of bias. I consider several scenarios for selection on unobservables, using the across-classroom variance of observed variables to calibrate each. Results indicate that even well-controlled models may be substantially biased, with the magnitude of the bias depending on the amount of information available for use in classroom assignments.

*Princeton University and NBER. Mailing address: Industrial Relations Section, Firestone Library, Princeton, NJ 08544. E-mail: jrothst@princeton.edu. I thank Nathan Wozny and Enkeleda Gjerci for research assistance and the Industrial Relations Section and the Center for Economic Policy Studies at Princeton for financial support. I am grateful to the North Carolina Education Data Research Center and the North Carolina Department of Public Instruction for assembling and making available the data used in this study. This work has benefited from helpful conversations with Gordon Dahl, Ed Glaeser, Brian Jacob, David Lee, and Diane Schanzenbach, and from comments from Jane Cooley.

1 Introduction

Proposals to consider teacher quality in hiring, compensation, and retention require adequate measures of quality. This is increasingly defined in terms of educational outputs, as reflected in student performance, rather than by teacher inputs like graduate degrees and experience. In order for output-based quality measures to be of use, they must reflect teachers' *causal* effects on the student outcomes of interest, not pre-existing differences among students for which the teacher cannot be given credit or blame.

If students were known to be randomly assigned to teachers, there would be no systematic differences in students' potential outcomes across teachers, so straightforward comparisons of mean end-of-year achievement would provide unbiased estimates of teachers' effects.¹ But there is good reason to think that teachers are not in fact randomly assigned. Principals may attempt to group students of similar ability together, so as to permit more focused teaching to students' skill levels, or they may try to spread high- and low-ability students across classrooms. Teachers who are thought to be particularly skilled at teaching, e.g., reading skills may be assigned students who are in need of extra reading help. Students who are known to create trouble together may be intentionally assigned to different classrooms. Teachers who the principal would like to reward may be given the easiest-to-teach students, and teachers who the principal would like to induce to find another job may be given the troublemakers.² Finally, parents, perceiving teacher assignments as important determinants of their children's success, may intervene to ensure that their students are given

¹There would still be the problem of accounting for sampling variation in the estimates: Because each teacher is in contact with only a few dozen students per year, annual estimates of teacher effects are quite noisy, and compensation schemes based on these estimates would have to be robust to the mis-identification of teacher quality that results from this noise. But existing strategies – e.g., the Empirical Bayes approach used by Kane and Staiger (2008) or the similar Best Linear Unbiased Predictor used by the Tennessee Value Added Assessment System (Sanders and Horn, 1994) – suggest methods for doing this.

²This aspect of assignments is likely to depend on the accountability metric in place: If teachers are rewarded for their value added and if value added estimates can be biased by systematic student assignment, the pattern of assignments is likely to change so that favored teachers benefit from this bias and disfavored ones are penalized.

a favored teacher or kept away from a disfavored one.

The evaluation challenge in teacher effect modeling is to distinguish teachers' causal effects from the effects of pre-existing differences between the students in their classrooms. If the determinants of classroom assignments are not adequately controlled, teacher effect estimates will be biased. This bias is not averaged away even in large samples, and existing methods for adjusting value added estimates for sampling error will not (absent strong assumptions about, e.g., the across-year stability of teachers' assignments) remove its effects from teacher rankings.

The premise of "value added" models is that differences in the difficulty of the task faced can be controlled by holding teachers responsible not for their students' absolute end-of-year achievement but only for the students' gains over the course of the year. Rothstein (2008) shows that this is false. Students are sorted across classrooms in ways that correlated not just with their score levels but also with their annual gains. Specifically, 4th grade gains are highly non-randomly sorted across 5th grade classrooms, with nearly as much across-class variation as in 5th grade gains. Because annual achievement tends to revert quickly toward a student-specific mean, a student with a 4th grade gain that exceeds the average by one standard deviation can be expected to fall short of the average in 5th grade by about 0.4 standard deviations. Existing value added models attribute this mean reversion to the 5th grade teacher. A teacher assigned students with high 4th grade gains in the previous year will look like a bad teacher through no fault of her own, while a teacher whose students posted poor gains in the previous year will be credited for their predictable reversion to trend.

Although Rothstein (2008) documents substantial non-randomness in teacher assignments that violates the restriction of common value added models (hereafter, VAMs), he does not directly estimate the magnitude of the resulting biases, and he provides little evidence about the prospects for correcting them via more sophisticated controls for students'

past achievement trends.³

This paper attempts to quantify the bias created by non-random assignment in several value added specifications. Three conditions govern the bias. It depends first on the amount of information available for use in the teacher assignment process about students' potential end-of-year achievement or annual gain, second on the importance attached to this information in the formation of teacher assignments, and third on the degree to which the control variables included in the value added specification can proxy for those used in assignments.

I take the classroom effect – the causal effect of being in one classroom as opposed to another in the same school – as the parameter of interest.⁴ This avoids the problem of distinguishing different components of the classroom effect, the most obvious being the effects of teacher quality and of peers. This problem is complex, and is likely made even more difficult by non-random assignments of students to classrooms. But the identification of classroom effects is a necessary precondition for the larger problem of isolating teachers' causal effects, and by focusing on this smaller, first problem I can place a lower bound on the bias in estimates of teachers' effects that is produced by the assignment process.

Bias in classroom effect estimates can be measured directly if and only if classroom assignments are assumed to depend only on variables that are observed by the analyst, with random assignment conditional on these variables. Although I present estimates of this form, the selection-on-observables assumption is unattractive.

The bias created by selection on unobservables cannot be measured directly, but its magnitude can be quantified under assumptions about the amount and nature of information

³Rothstein (2008) does demonstrate that unbiased estimation requires controls for *dynamic* student achievement: Teacher assignments are not governed solely by permanent student characteristics, but respond dynamically to each year's test scores. This rules out fixed effects solutions like those used by Harris and Sass (2006); Koedel and Betts (2007); Jacob and Lefgren (2008); Rivkin et al. (2005); and Boyd et al. (2007).

⁴Some value added studies use multiple cohorts of students assigned to each teacher. If assignments are uncorrelated across cohorts – that is, if a teacher who gets high-potential-gain students this year is no more or less likely than any other teacher to get high-potential-gain students next year – then multiple cohort studies can convert bias in the classroom effect into mere sampling error in the teacher's effect. But this uncorrelated assignments assumption is a strong one, and it does not appear to hold – even approximately – in the North Carolina data used here and in Rothstein (2008).

available to the principal for use in classroom assignments and about the way in which that information is used.⁵ The approach that I take is in the spirit of Altonji, Elder, and Taber's (2005) assumption that sorting on unobserved variables resembles sorting on observables, though the specific assumptions differ: Where Altonji et al. (2005) assume that sorting is incidental and is equally correlated with observed and unobserved determinants of the outcome variable of interest, I assume that the sorting is intentional and that it depends on a limited set of predictors that are observed by the school principal, a subset of which are observed by the researcher as well. Altonji et al.'s assumption represents a limiting case for my analysis, in which the principal can perfectly predict students' end-of-year achievement and gains before making teacher assignments.

Section 2 describes the data. In Section 3, I demonstrate that past test scores and behavioral variables are strongly predictive of future achievement and achievement gains. Section 4 summarizes the evidence from Rothstein (2008) that teacher assignments are importantly correlated with past scores. In Section 5, I compute and summarize the bias that arises in several common value added models if classroom assignments are random conditional on the observed variables. Section 6 develops the methodology for assessing the bias that would arise if the principal had more information about students' potential learning growth than is available in research data sets. Section 7 presents the results of the analysis of bias with selection on unobservables. Section 8 concludes.

2 Data

I work with longitudinal administrative data on students in public elementary schools in North Carolina, assembled and distributed by the North Carolina Education Research Data Center. North Carolina has been a leader in the development of linked longitudinal data on

⁵For simplicity, I discuss class assignments as the outcome of principals' decisions. This is not meant to restrict the principal to be the only determinant of these assignments; the principal's decision might reflect input from parents, teachers, and the student itself.

student achievement, and the North Carolina data have been used for several previous value added analyses (Clotfelter et al., 2006; Goldhaber, 2007).⁶

I focus on the value added of 5th grade teachers in 2000-2001. I use annual end-of-year tests that were given in grades 3-5, as well as “pre-test” scores given at the beginning of grade 3. I treat the pre-tests as 2nd grade tests.

The tests purport to use a so-called “developmental” scale, and the score scale is intended to be meaningful (i.e. scores are cardinal and not simply ordinal measures) both across grades and across the distribution within grades.⁷ I standardize scores so that the population mean is zero and the standard deviation one in 3rd grade; by using the same standardization in all grades I preserve the comparability of scores across grades.

The North Carolina data do not identify students’ teachers directly, but they do identify the person who administered the end-of-grade tests. In the elementary grades, this was usually the regular teacher. I follow Clotfelter et al. (2006) in using a linked personnel database to identify test administrators with regular teaching assignments. I count a match as valid if the test administrator taught a self-contained (all day, all subject) 5th grade class, if that class was not coded as Special Education or Honors, and if at least half of the tests that she administered were to 5th grade students. 73% of 5th grade tests were administered by teachers who are valid by this definition.

My analysis focuses reading scores, though similar results obtain for math scores. My sample consists of students who were in 5th grade in 2000-2001, who had a valid teacher assignment in that year, and for whom I have complete test score data in grades 3-5. Table 1A presents summary statistics and a correlation table for reading scores on the 3rd grade

⁶North Carolina was one of the first two states approved by the U.S. Department of Education to use “growth-based” accountability models in place of the status-based metrics that are otherwise required under No Child Left Behind.

⁷It is not clear that a scale with this property is even possible (Martineau, 2006), or even if it is how one would know whether a test’s scale has the property. Nevertheless, value added modeling as typically practiced is difficult to justify if scores do not have the so-called interval property. See Ballou (2002) and Yen (1986). The analysis here is not sensitive to violations of this property, though if it does not hold the value added estimators considered (here, and elsewhere in the literature) are difficult to justify. See Rothstein (2008).

pretest and on the end-of-grade tests in 3rd, 4th, and 5th grades, as well as for the 5th grade gain score (defined as the difference between the 4th and 5th grade scores). Mean scores in my complete-data sample are about 0.07 standard deviations higher than in the population in every grade. Scores are correlated about 0.80 in adjacent grades (lower for the 3rd grade pre-test, which is substantially shorter), with slightly reduced correlations across longer time spans. 5th grade gains are weakly positively correlated (+0.07) with 5th grade score levels and strongly negatively correlated (-0.52) with 4th grade scores. They are notably negatively correlated (-0.25) with 3rd grade scores as well.

Observed scores are noisy measures of true achievement. The degree of measurement error in test scores is usually measured by the “test-retest” reliability, the correlation between students’ scores on alternative forms of the same test administered a short interval apart.⁸ A 1996 report estimates that the test-retest reliability of the North Carolina 7th grade reading test is 0.86 (Sanford, 1996, p. 45). Unfortunately, test-retest studies have not been conducted for other grades. Under the assumption that individual item reliability is constant across grades and that item responses are independent, the 7th grade reliability can be extended to the shorter tests in earlier grades.⁹ Doing so, I estimate that the grade-3 pre-test has reliability 0.72, the grade-3 end-of-grade test has reliability 0.84, and the tests in grades 4 and 5 have reliability 0.86. I treat these as known, without sampling error.¹⁰

⁸Test makers often report alternative measures of reliability, e.g. internal consistency measures that are based on correlations between a student’s scores on different subsets of questions. The internal-consistency reliabilities for the tests in grades 3, 4, and 5, respectively, are 0.92, 0.94, and 0.93 (Sanford, 1996, p. 45). The corresponding statistic for the grade-3 pre-test used for the cohort under consideration is not reported, but a more recent form of the test has reliability 0.82 (as compared with 0.92 in on the corresponding tests in grades 3-5; see Bazemore, 2004, p. 63). These statistics are computed under the assumption that responses are independent across questions; common shocks (e.g. a cold on test day) would lead these methods to overstate the test’s reliability.

⁹If item responses are not independent, reliability will be less sensitive to test length, and I will most likely understate the reliability of the (relatively short) 3rd grade pretest.

¹⁰The sample for the test-retest study was only 70 students, in 3 classrooms. If the 70 observations are independent, an approximate confidence interval for the grade-7 test reliability is (0.78, 0.91), though within-classroom dependence would imply a wider interval. Note also that a given test will have higher reliability in a heterogeneous population than in a homogeneous one; the likely homogeneity of the test-retest sample suggests that the reliability in the population of North Carolina students is probably higher than was indicated.

A known reliability allows me to compute summary statistics for true achievement, net of measurement error, assuming that errors are independent across grades. These are reported in Table 1B. The correlation between a student's true achievement in grades g and $g+1$ is approximately 0.96. The 5th grade gain is negatively correlated with achievement levels in all grades.

One can examine across-grade correlations in gain scores as well as in score levels. The correlation between observed grade-4 and grade-5 gains is -0.42. Measurement error in the annual test scores biases this downward, but even when corrected the correlation remains negative. Thus, students with above-average gains in grade 4 will, on average, have below-average gains the following year. To the extent that such students are systematically assigned to particular teachers, value added models that fail to account for this mean reversion will be biased against those teachers.

3 Predictions of grade 5 achievement and gains

Table 2 presents several specifications for students' reading scores at the end of grade 5, using prior scores and other predetermined variables as explanatory variables. Because it is almost certainly more difficult to control for the sorting of students across schools than within, and because I focus in this paper in identifying differences in teachers' effects within schools, I consider only specifications for within-school variation in 5th grade scores. The first column shows that 13% of the variance in 5th grade scores is across schools. Column 2 adds the 4th grade reading score. This has a coefficient of 0.680; neither zero (corresponding to a white noise process for individual scores) nor one (corresponding to a martingale) is within the confidence interval. The inclusion of the 4th grade score increases the model's R-squared by 0.55; 4th grade scores explain 63.5% of the within-school variation in 5th grade scores.

Column 3 adds to the specification reading scores from the beginning and end of grade

3. Both are significant predictors of 5th grade scores. Their inclusion lowers the 4th grade score coefficient by about one third, and raises the within-school R-squared by 0.045. Column 4 adds three lagged scores on the math exam. Again, all are significant. The within-school R-squared is 0.058 higher than in the specification with just a single lagged reading score. Column 5 adds 28 additional covariates, measured in grade 4, that might help to predict students' grade-5 achievement. These include race, gender, and free lunch status indicators; measures of parental education; various categories of "exceptionality" and learning disabilities; and measures of the time spent on homework and watching TV. These are jointly highly significant, though their inclusion raises the explained share of variance by only 0.003.

The available variables – all or nearly all of which would be readily observable when classroom assignments are made – explain nearly 70% of the within-school variation in students' grade-5 test scores. Moreover, this substantially understates the predictability of student achievement. Recall from Section 2 that 14% of the variance in observed 5th grade scores is noise that would not even persist into a second administration of the test a week later. This noise is irrelevant to the predictability of achievement, and is uncorrelated with all predictor variables. Table 2 also shows estimates of the explained share of the within-school variance of true achievement, net of this transitory noise. These range from 0.764 with just the 4th grade score to 0.837 with the full set of controls.

Many value added models focus on the gain score rather than the end-of-year level. So long as the grade-4 score is included as a covariate, the coefficients in a prediction equation for (observed) gains are identical to those for levels, save that the grade-4 score coefficient is reduced by 1. The bottom rows of the Table show the R-squared statistics for specifications that take grade-5 gains as the dependent variable. These range from 0.279 to 0.398 within schools. The first-difference transformation reduces but does not eliminate predictability; the principal clearly has substantial information at his disposal for the prediction of student

gain scores.¹¹

Also relevant to the analysis below is the value of past gains for predicting future scores and gains. Table 3 presents specifications using grade-4 gains as explanatory variables. These explain only 0.2% of the within-school variance in 5th grade achievement but 10.3% of the variance in 5th grade gains.

4 Evidence for non-random assignment

The simplest value added model estimates each teacher’s effect as the average gain score of her students. In order to attribute this average gain to the teacher, it must be the case that the information used to make teaching assignments is uninformative about students’ potential gains, conditional on any control variables. As shown in Section 3, prior achievement and gains are strongly predictive of future scores and gains, so correlations between teacher assignments and past gains would violate the simple VAMs identifying assumption. Rothstein (2008) tests for “effects” of grade- g teachers on gains in grade $g - 1$. Given the evidence in Table 3, effects of this sort would indicate that expected grade- g gains are not balanced across grade- g classrooms, and that the simple VAM will be biased.

Let A_{ig} be the test score for student i in grade g . Then the student’s grade- g gain score is $\Delta A_{ig} \equiv A_{ig} - A_{i,g-1}$. The simple value added model specifies gain scores as depending only on school (-by-grade) and teacher effects and random errors:¹²

$$\Delta A_{ig} = S_{ig}\alpha_g + T_{ig}\beta_g + \varepsilon_{ig}, \quad (1)$$

¹¹The fit statistics cannot be directly converted to those that would be seen for the true gain score, net of measurement error, because measurement error in the grade-4 score appears on both sides of the equation for grade-5 gains. I discuss in Section 6.4 how the coefficients of specifications for true gains can be recovered from the estimates in Table 2. True gains are quite predictable as well.

¹²This is essentially the specification used by the Tennessee Value Added Analysis System (Ballou et al., 2004; Bock and Wolfe, 1996; Sanders and Horn, 1994, 1998; Sanders and Rivers, 1996; Sanders et al., 1997). Though TVAAS is estimated through a mixed effects framework, it implies equation (1)’s specification for gains, and it requires the same exclusion restriction.

where S_{ig} and T_{ig} are vectors of indicators for students' grade- g schools and teachers, respectively. The teacher effects β_g are normalized to have mean zero across all teachers in each grade at each school. The estimated effect of teacher j at school s is the average gain in classroom j less the average gain in the school:

$$\begin{aligned}\hat{\beta}_{gj} &= E[\Delta A_{ig} | T_{ig} = j, S_{ig} = s] - E[A_{ig} | S_{ig} = s] \\ &= \beta_{gj} + E[\varepsilon_{ig} | T_{ig} = j, S_{ig} = s] - E[\varepsilon_{ig} | S_{ig} = s].\end{aligned}\quad (2)$$

If the mean of the error term distribution is the same for all teachers in the grade at the school, $E[\varepsilon_{ig} | S_{ig}, T_{ig}] = E[\varepsilon_{ig} | S_{ig}]$, this is unbiased.

This identifying assumption can be evaluated by examining gains in grade $g - 1$. The mean gain in grade $g - 1$ for students who will have teacher j in grade g is

$$E[\Delta A_{i,g-1} | T_{ig} = j, S_{ig} = s] = E[S_{i,g-1}\alpha_{g-1} + T_{i,g-1}\beta_{g-1} + \varepsilon_{i,g-1} | T_{ig} = j, S_{ig} = s]. \quad (3)$$

Setting aside the first two terms, which might be absorbed through controls for the school attended and teacher assigned in grade $g - 1$, the grade- g teacher's "effect" on $g - 1$ gains is $\theta_j \equiv E[\varepsilon_{i,g-1} | T_{ig} = j, S_{ig} = s] - E[\varepsilon_{i,g-1} | S_{ig} = s]$, the average grade- $g - 1$ residual among students in grade- g classroom j less the average in school s .

If this is non-zero, the grade- g effect $\hat{\beta}_{gj}$ will in general be biased. Suppose, for example, that the ε process is autoregressive: $\varepsilon_{ig} = \rho\varepsilon_{i,g-1} + v_{ig}$, where v is serially uncorrelated and v_{ig} is independent of the grade- g teacher assignment. Then

$$E[\varepsilon_{ig} | j(i, g) = j, s(i, g) = s] - E[\varepsilon_{ig} | s(i, g) = s] = \rho\theta_j. \quad (4)$$

Table 3 indicates that $\rho = -0.39$. Thus, any evidence that θ_j is also non-zero would imply that the identifying assumption for the value added model (1) is violated.

I present estimates of 5th grade teachers' coefficients in models for gain scores in grades 5, 4, and 3, using specifications like (1) and a balanced panel of students who attended the same school for all three grades. These are similar to those reported in Table 3 of Rothstein (2008), albeit estimated from a slightly different sample.

Begin with the model for grade-5 gains,

$$\Delta A_{i5} = S_{i5}\alpha_5 + T_{i5}\beta_5 + \varepsilon_{i5}. \quad (5)$$

The 3,013 elements of the $\hat{\beta}_5$ vector (normalized to mean zero at the school) can be summarized by their standard deviation. This, 0.145, is shown in Column 1 of Table 4.¹³ I also report an adjusted standard deviation that subtracts from the across-teacher variance the contribution of sampling error to this variance (Aaronson et al., 2007; Rothstein, 2008). This adjusted standard deviation, which estimates the variability of the true β coefficients net of sampling error, is 0.106: A teacher who is one standard deviation better than average has students who gain 1/10 of a standard deviation (of achievement levels) relative to the average over the course of the year. This resembles existing estimates (Aaronson et al., 2007; Kane et al., forthcoming; Rivkin et al., 2005).

The remaining columns of Table 4 present counterfactual estimates that vary only the dependent variable. Column 2 presents estimates for 4th grade gains:¹⁴

$$\Delta A_{i4} = S_{i5}\tilde{\alpha}_4 + T_{i5}\tilde{\beta}_4 + e_{i4}. \quad (6)$$

We know that there are no causal effects of 5th grade teachers on 4th grade gains (i.e.

¹³Across-teacher means and standard deviations are weighted by the number of students taught, and degrees of freedom are adjusted for the normalization of $\hat{\beta}_5$. Further details of the methods are available in Rothstein (2008).

¹⁴In principle, the omission of controls for 4th grade teachers and schools creates an omitted variables bias in (6). Rothstein (2008) presents estimates that include such controls. In practice, there is little correlation between teacher assignments in different grades, and estimates of the coefficients on T_5 in equations for grade-4 gains are nearly identical in specifications that do and do not control for T_4 .

that $\tilde{\beta}_4 = 0$), so any non-zero coefficients in this specification are indicative of student sorting. The hypothesis that $\tilde{\beta}_4 = 0$ is decisively rejected, and indeed there is nearly as much variation in the elements of $\hat{\beta}_4$ as in those of $\hat{\beta}_5$: The sampling-adjusted standard deviation of 5th grade teachers' normalized "effects" on 4th grade gains is 0.080, nearly as large as that for 5th grade gains. Column 3 presents an analogous model where the dependent variable is the 3rd grade gain, the difference between the student's score on the end-of-grade reading test and the beginning-of-the year pretest. We see even larger apparent effects of 5th grade teachers here.

The lower portion of Table 4 presents correlations between the estimates of the coefficient vectors β_5 , $\tilde{\beta}_4$, and $\tilde{\beta}_3$, first unadjusted for sampling error and then adjusted. Adjacent coefficients are highly negatively correlated, both before and after the adjustment for sampling error, while there is nearly no correlation between β_5 and $\tilde{\beta}_3$.

Two of these correlations are of particular interest here. First, $\text{corr}(\beta_5, \tilde{\beta}_4) = -0.35$. This indicates that 5th grade teachers who appear (by the simple model 5) to have high value added tend to be those whose students experienced below-average gains in grade 4. As noted earlier, gains are negatively autocorrelated at the student level; at least a portion of the variation in estimated 5th grade value added apparently reflects predictable consequences of non-random student assignments.

The second interesting correlation is that between $\tilde{\beta}_4$ and $\tilde{\beta}_3$, -0.36. One hypothesis that could explain the presence of counterfactual "effects" of 5th grade teachers on earlier grades' gains is that students differ systematically in their rate of gain, and that classroom assignments depend in part on that rate. Rothstein (2008) refers to this explanation as "static tracking"—the determinants of classroom assignments are constant across grades, and conditional on these determinants the test score in grade g does not affect the teacher assignment in $g + 1$. In the presence of static tracking, the bias in teacher effects coming from non-random assignment can be absorbed by pooling data on a student's gains across several

grades and including student fixed effects in the specification. This sort of specification is used by Harris and Sass (2006); Koedel and Betts (2007); Jacob and Lefgren (2008); Rivkin et al. (2005); and Boyd et al. (2007), among others.

As Rothstein (2008) notes, static tracking implies that in simple specifications like those in Table 4 the coefficients for the grade- g teacher on gains in grades h and k ($h, k < g$) should be identical, up to sampling error. In other words, $\text{corr}(\tilde{\beta}_4, \tilde{\beta}_3) = 1$.¹⁵ This restriction does not even approximately hold in the data. Classroom assignments are evidently *not* made on the basis of permanent student characteristics, but respond dynamically to annual student performance. This implies that student fixed effects specifications provide inconsistent estimates of teachers' causal effects. The only way to control for non-random classroom assignments while permitting consistent estimation of teachers' effects is to measure the determinants of assignments directly.

Many value added specifications (Gordon et al., 2006; Kane et al., forthcoming; Aaronson et al., 2007; Jacob and Lefgren, 2008) control for the baseline score, in effect modeling the end-of-year score as a function of the beginning-of-year score and the teacher assignment. These specifications are robust to dynamic teacher assignments of a very restricted form: Unless teacher assignments are random conditional on the baseline score, estimates will still be biased. The estimates in Tables 2 and 3 indicate that there is a great deal of information available to principals about students' potential gains above and beyond that provided by the lagged score; there is no reason to expect that the use of this information in forming classroom assignments can be absorbed with simple controls. I show below that the once-lagged-score specification is rejected by the data.

¹⁵Again, this conclusion is supportable only if the correlation between $\tilde{\beta}_4$ and $\tilde{\beta}_3$ is negative in specifications that include controls for 4th and 3rd grade teachers, where those in Table 2 do not. The correlation is nearly identical when these controls are included.

5 Selection on observables

Strategies for isolating causal effects in the presence of non-random assignment of treatment (in this case, of classroom assignments) depend importantly on whether the determinants of treatment are observed or unobserved. In this section, I assume that 5th grade teacher assignments are random conditional on observable variables measured in 4th grade. Under this assumption, bias can be avoided by controlling for the full set of observables in the value added model. But models that use fuller controls may be biased if the included variables are unable to absorb all of the non-randomness of teacher assignments.

Note that no harm is done by controlling for variables that are *not* used in teacher assignments; this merely sacrifices some precision. Accordingly, I assume in this subsection that $X_{i,4}$ is the set of variables included in Column 5 of Table 2 – the history of math and reading test scores plus a set of demographic and behavioral variables as measured in grade 4. My baseline estimator for 5th grade teachers’ causal effects is:

$$\Delta A_{i5} = S_{i5}\alpha + T_{i5}\beta + X_{i,4}\gamma + \varepsilon_{ijs5}. \quad (7)$$

I estimate this by OLS, imposing the normalization that β have weighted mean 0 across teachers at each school. I compare the estimates that it yields to those from four value added models (hereafter, *VAMs*) with less-complete controls:

$$\mathbf{VAM1:} \quad A_{i5} = S_{i5}a + T_{i5}b + e_{i5}$$

$$\mathbf{VAM2:} \quad \Delta A_{i5} = S_{i5}a + T_{i5}b + e_{i5}$$

$$\mathbf{VAM3:} \quad \Delta A_{i5} = S_{i5}a + T_{i5}b + A_{i4}c + e_{i5}$$

$$\mathbf{VAM4:} \quad \Delta A_{i5} = S_{i5}a + T_{i5}b + A_{i4}c_1 + A_{i3}c_2 + A_{i2}c_3 + e_{i5}$$

VAM1 credits each teacher with the average achievement of students in her class (less the school-level average). Few would advocate this “levels” specification. VAM2 credits each

teacher with her students' average 5th grade gain score (again less the school average). This is the basic specification used in most value added policy and above in Section 4. VAM3 controls for students' 4th grade achievement. Though I have written this as a model for the 5th grade gain, it is equivalent to a similar specification for the 5th grade score. VAM4 controls not just for last year's score but for the two prior scores as well. This sort of specification is not widely used, but in principle it could be used in most value added implementations.

For each model, I compute the standard deviation across teachers of b and of the bias relative to the coefficient vector from the richer specification (7). A useful summary statistic is the variance of the bias relative to that of teachers' true effects, $\frac{V(b-\beta)}{V(\beta)}$. I also compute the correlation between the bias and the true effect, $\text{corr}(b-\beta, \beta)$: It is helpful to know whether good teachers (at least as indicated by the baseline model 7) are helped or hurt by the assignment process. A strong positive correlation between true effects and the bias would imply that teacher rankings are not much affected by sorting bias.

Table 5 presents the results. Each statistic is computed first from the estimated coefficients (in the first panel), then adjusted for the influence of sampling error (second panel). The baseline specification indicates that the standard deviation of teachers' effects is 0.096, or 0.124 before the adjustment for sampling error. VAM1 indicates much more variability of teacher effects, though this is primarily bias—the bias in this specification is more than three times as large (in variance terms) as the true variability that we are attempting to measure. The specification for gain scores, VAM2, eliminates much of the bias, but its variance is still half that of the true effects. VAM3, controlling for the 4th grade score, cuts the standard deviation of the bias in half; here, the variance of the bias is 13% of that of the quality signal. This is small in comparison with the previous models, but still substantial enough to represent a problem for policy. In each case, biases are only weakly correlated with true coefficients.

VAM4 eliminates nearly all of the bias relative to the richer selection-on-observables specification. This is unsurprising: Recall that Table 2 indicated that the control variables included in (7) but excluded from VAM4 added only 0.3% to the explained share of variance of grade-5 achievement and 0.6% to the explained share of variance of grade-5 gains. Thus, my assumption that specification (7) permits unbiased estimation of teachers' causal effects implies that omitted variables bias in VAM4 is negligible. To understand the true potential for bias in this specification, we will need to consider the impact of selection on information that is *unobserved* in my sample but is available for use in forming classroom assignments. I develop methods for assessing this in the next Section.

6 A model of tracking on unobservables

There is no good reason to think that classroom assignments depend only on the variables available in my data. Indeed, the presence of noise in the observed test score history strongly suggests otherwise. Even a principal who had no additional information would almost certainly be able to form a less noisy measure of students' achievement each year by combining test scores with other measures (e.g. grades) that I do not observe. In this section I develop a framework in which classroom assignments depend on the observed variables and on unobserved variables that have known correlations with the observables. This permits computation of the variance across teachers of the bias in feasible estimates of β , though not the bias in any individual teacher's estimated effect.

6.1 The sorting process

Let Ω be the information available to the principal for prediction of student outcomes Y , and let $I = E[Y | \Omega]$ be the best prediction that the principal can make given the available information. Y might measure true gains or observed gains; we will see below that this has important consequences for the analysis. The amount of information available to the

principal about future gains can be measured by $V(I)/V(Y)$ or, if we write $Y = I + \varepsilon$, by $\frac{\sigma_I^2}{\sigma_I^2 + \sigma_\varepsilon^2}$.

The principal makes classroom assignments on the basis of an index $\lambda = I + \eta$. I represents the portion of the determinants of classroom assignment that are predictive of future achievement, while η represents the remaining portion. I assume that η is orthogonal to I , to all variables in the principal's information set, Ω , and (by construction) to ε .¹⁶ I also assume that $\{I, \eta, \varepsilon\}$ are jointly normally distributed. The importance of predicted outcomes in assignments is controlled by σ_η^2 : If the principal assigns students to classrooms solely on the basis of predicted outcomes, $\sigma_\eta^2 = 0$, while perfect random assignment can be seen as the opposite limiting case, $\sigma_\eta^2 = \infty$.

Students are sorted perfectly on λ into classes. That is, all of the students assigned to a particular teacher have the same λ value. This is a crude approximation at best. A typical school has three to five classes per grade; even if these classes are perfectly stratified, λ will have considerable heterogeneity within classes. The assumption of perfect sorting is made for reasons of mathematical tractability: With perfect sorting, we have simple expressions for, e.g., the across-class variance of I . With less than perfect sorting, my methods will understate the importance of I (relative to η) in classroom assignments and therefore will understate the bias due to these assignments.¹⁷

6.2 Bias in undercontrolled value added models

The role of I in classroom assignments produces across-classroom differences in student achievement gains that do not reflect teacher quality, biasing value added models with inadequate controls. It is easiest to characterize the bias in VAM2, which does not include any

¹⁶The assumption that η is uncorrelated with observed predictors of Y is central to my strategy: I assume that the principal uses the observed variables solely to predict Y , and does not sort students on the basis of these variables out of proportion to their information about Y . This is required because Ω (and therefore I) is only partially observable; I use the across-classroom share of variance of the observed test score history to recover σ_η^2 .

¹⁷The basic approach could be extended to stratification on λ across a finite number of classes (so that one class has students with $\lambda \in (-\infty, c_1)$, another has $\lambda \in (c_1, c_2)$, etc.), at the cost of considerable additional complexity. I do not pursue this approach here.

controls. Suppose that Y is the gain score, so that the principal makes classroom assignments on the basis of his predictions of the same outcome that is used to measure teacher effects. Then across-classroom differences in I represent biases in b . The across-classroom variance of I is $V(I) - V(I|\lambda) = \text{corr}^2(I, \lambda)V(I) = \frac{\sigma_I^4}{\sigma_I^2 + \sigma_\eta^2}$.

In richer VAMs that include control variables, Z , these variables may absorb some of the bias. Write the regression of I onto T and Z as $I = T\kappa + Z\pi + v$, where $T\kappa$ is the remaining bias. Because λ is assumed to be perfectly sorted across classrooms, and because the teacher's identity is informative about I only through λ , the teacher effects $T\kappa$ solely reflect differences across classes in λ , and we can therefore write $T\kappa = \lambda\xi$ and $I = \lambda\xi + Z\pi + v$.

To obtain the coefficients ξ and π , I assume that any variables available for use in the value added model are used by the principal to form his predictions (i.e., $Z \subseteq \Omega$), and are orthogonal to η . The variance of the triplet (λ, I, Z) is thus

$$V \begin{pmatrix} \lambda \\ I \\ Z \end{pmatrix} = \begin{pmatrix} \sigma_I^2 + \sigma_\eta^2 & \sigma_I^2 & \text{cov}(Z, I) \\ \sigma_I^2 & \sigma_I^2 & \text{cov}(Z, I) \\ \text{cov}(Z, I) & \text{cov}(Z, I) & V(Z) \end{pmatrix} \quad (8)$$

and

$$\begin{aligned} \begin{pmatrix} \xi \\ \pi \end{pmatrix} &= \left(V \begin{pmatrix} \lambda \\ I \\ Z \end{pmatrix} \right)^{-1} \begin{pmatrix} \text{cov}(\lambda, I) \\ \text{cov}(Z, I) \end{pmatrix} \\ &= \begin{pmatrix} \sigma_I^2 + \sigma_\eta^2 & \text{cov}(Z, I) \\ \text{cov}(Z, I) & V(Z) \end{pmatrix}^{-1} \begin{pmatrix} \sigma_I^2 \\ \text{cov}(Z, I) \end{pmatrix}. \end{aligned} \quad (9)$$

There are thus three parameters that determine the variance of the bias in the under-controlled model, two deriving from the sorting process and one from the choice of value added specification. The first, σ_I^2 , concerns the principal's ability to predict students' out-

comes. The second, σ_η^2 , controls the importance that the principal attaches to predicted outcomes in classroom assignments. The last describes the relationship between the control variables included in the value added model and the principal's prediction, $\text{cov}(Z, I)$. (A fourth parameter, $V(Z)$, is readily measured.) With knowledge of these parameters, we can recover ξ and, via it, $V(\kappa) = \xi^2 V(\lambda) = \xi^2 (\sigma_I^2 + \sigma_\eta^2)$.

To fix ideas, it is useful to consider three limiting cases. First, suppose that we control for all of the variables used by the principal (under selection on observables). Then $I = Z\pi$ for some π , $\xi = 0$, and $V(\lambda\xi) = V(T\kappa) = 0$. Second, suppose that the principal places much more weight on variables unrelated to achievement than on predicted achievement in forming assignments, $\sigma_\eta^2 \gg \sigma_I^2$. Then there is little sorting on I , $\xi \approx 0$, and $V(T\kappa) \approx 0$ regardless of the content of Z . Finally, suppose that the principal uses *only* predicted achievement to form assignments, $\sigma_\eta^2 = 0$. Then $\lambda = I$, and bias depends only on the extent to which Z can account for the principal's predictions. If there are no Z variables, bias will be in proportion to the principal's ability to predict performance; with Z variables, bias will depend on the extent to which the econometrician can predict the principal's predictions.

6.3 The principal's prediction

Clearly, results concerning bias depend importantly on the information available to the principal for predictions of students' future growth. In order to impose structure, I parametrise the principal's information and its relationship with observed variables. I consider several scenarios. Intermediate cases between selection-on-observables and perfect predictability of future outcomes are the most realistic and I focus on these, though I also include the limiting cases for comparison. I begin with base cases in which selection is on observables, as in Section 5:

- A. $I = E[Y | A_{g-1}] = A_{g-1}\phi_A$: The principal has no information about future achievement gains beyond that contained in the prior grade's test score.

B. $I = E[Y | A] = A\phi_B$, where $A = (A_1, \dots, A_{g-1})$ is the history of test scores up to grade $g - 1$. The principal observes the test score history, but has no additional information about achievement gains.

Note that scenarios A & B are falsified by the evidence in Table 2: Since the principal can observe all of the grade-4 variables that are available in my data, the fact that these variables are useful in predicting gains indicates that the principal has more information about potential gains than just the score history. Nevertheless, these scenarios provide useful baselines.

One use to which the principal might put his information is to reduce the noise that is contained in the test score history, A . Thus, a useful parametrisation of the principal's information assumes that he has access to k some number of additional series, unobserved by the researcher, that measure the true achievement history with independent, identically distributed error. That is, if the true achievement history through grade $g - 1$ is $A^* = (A_1^*, \dots, A_{g-1}^*)$ and $A = A^* + u$, we can suppose that the principal observes in addition $\{q_1, q_2, \dots, q_k\}$, where $q_j = A^* + v_j$, $E(v_j'v_j) = E(u'u)$, and $E(v_j'v_h) = E(v_j'u) = 0$. The q series can be thought of as representing grades, student evaluations, or classroom observations that are available to the principal but not reported in typical data sets.

C. $I = E[Y | A, q_1, \dots, q_k] = A\phi_C + q_1\tau_{C1} + q_2\tau_{C2} + \dots + q_k\tau_{Ck}$.

In the limit as $k \rightarrow \infty$, this scenario converges to one where the principal observes A^* without error:

D. $I = E[Y | A^*, A_{g-1}] = A^*\chi_D + A_{g-1}\phi_D$.

Note that I retain A_{g-1} as a conditioning variable; when Y is the observed gain, the lagged observed score can provide information about the error component in Y (i.e. about $\Delta A_g - \Delta A_g^*$). By contrast, when Y is the true gain, $\phi_D = 0$.

These scenarios are quite restrictive. As we will see, the true achievement history explains only 34% of the within-school variance of true achievement gains, and it is plausible

that the principal, who knows something of the child's family situation and emotional and cognitive development patterns, has information about the remaining portion. Let W be a proxy for the principal's information about $\Delta A_g^* - E[\Delta A_g^* | A_1^*, \dots, A_{g-1}^*]$. I assume without loss of generality that W is orthogonal to A^* and u .

E. $I = E[Y | A^*, A_{g-1}, W] = A^* \chi_E + A_{g-1} \phi_E + W \psi_E.$

Results here will depend on the quantity of information that W is assumed to contain. I index this by $f = V(W\psi_E)/V(A^*\chi_E)$. The limiting case (as $f \rightarrow V(\Delta A_g^* | A^*)/V(A^*\chi_E)$ ¹⁸) is one in which the principal can predict (true) gains perfectly:

F. $I = Y.$

This is not a particularly plausible scenario for the problem at hand, but it is useful for illustrating the relationship between the methods used here and those used by Altonji et al. (2005). Where in the earlier scenarios the principal observed all of the variables available to the analyst plus a subset of the remaining component of students' gains, here the principal observes both components equally. As a result, both are equally sorted across classrooms. This does not imply perfect sorting on potential gains, as the principal may consider factors other than a student's gain in forming classroom assignments. However, it does imply that selection on unobservables is identical to selection on observables, as in Altonji et al. (2005).¹⁹

The six scenarios are summarized in Table 6.

¹⁸This corresponds to $f \rightarrow \frac{1-R^2}{R^2}$, where R^2 is the explained share of variance from a regression of ΔA^* onto A^* . Since R^2 is empirically about 0.34, this limit is just below 2.

¹⁹Altonji et al. also consider intermediate cases, where the correlation between the unobserved determinants of selection and outcomes lies between zero (no selection) and the value corresponding to scenario F. The above framework can be seen as providing a basis for the choice of this correlation.

6.4 Calibration

Selecting a single scenario characterizing the principal's information, observed covariances can be used to calibrate the model. There are three steps to the calibration: First, the coefficients entering into the principal's prediction are estimated. This takes advantage of the observed relationship between gains and past scores, and of the structure that the various scenarios in Section 6.3 place on the principal's predictions. Second, the degree of sorting of students to classrooms is computed, using as an input the measured between-classroom variance in observed predictor variables. Third, the bias in various value added models is computed.

6.4.1 Estimating the prediction coefficients

Table 2 presented estimates of the ϕ coefficients for scenarios A and B, when Y is the observed gain score. Estimates for predictions of *true* gain scores, measured without error, can be computed using omitted variables formulae. The computation for scenario A illustrates the method. The observed gain, ΔA_g , equals the true gain ΔA_g^* plus the difference between the measurement errors in the grade- g and grade- $g - 1$ scores: $\Delta A_g = \Delta A_g^* + u_g - u_{g-1}$. Because the test measurement error is independent across grades, u_g cannot be predicted based on lagged variables. But u_{g-1} can, and the prediction coefficients can be obtained by viewing the $g - 1$ score, A_{g-1} , as a noisy measure of u_{g-1} . Because the test error is orthogonal to true achievement, $E[A_{g-1}u_{g-1}] = E[A_{g-1}^*u_{g-1} + u_{g-1}^2] = V(u_{g-1})$, and the bias in a regression that takes ΔA_g as the dependent variable relative to one that uses ΔA_g^* is simply $-V(u_{g-1})/V(A_{g-1})$. Thus, the A_{g-1} coefficient from a model for true gains equals the A_{g-1} coefficient from a model for observed gains (Table 2, column 2) plus $V(u_{g-1})/V(A_{g-1})$. A multivariate version of this yields coefficients for scenario B when Y is the true gain.

Similar methods can be used to recover the coefficients in scenarios C and D, for either definition of Y . Begin with scenario D when Y is the true gain. We have already discussed

a method for obtaining ϕ_B , the coefficients of a regression of true gains on the observed score history, A . A standard errors-in-variables formula relates these to the coefficients for predictions of true gains from the true achievement history, χ_D :

$$\phi_B = \left(I - (E[AA'])^{-1} E[uu'] \right) \chi_D. \quad (10)$$

Inversion of this formula provides an expression for χ_D . It is straightforward to extend this to the case where Y is instead the *observed* gain.

Now consider scenario C, where the principal observes $k + 1$ noisy measures of the achievement history but not the history itself. If Y is the true gain, the principal's best prediction will use the average of his measures, $\bar{A} = \frac{1}{k+1} (A + q^1 + \dots + q^k)$. The variance of the measurement error in this average will equal $\frac{1}{k+1}$ times the variance of the error in a single series. Thus, when $k + 1$ series are available the coefficients for each series will be

$$\phi_C = \tau_{C1} = \dots = \tau_{Ck} = \frac{1}{k+1} \left(I - \frac{1}{k+1} (E[AA'])^{-1} E[uu'] \right) \chi_D. \quad (11)$$

(Note that this is identical to (10) when $k = 0$.) When Y is instead the observed gain, the coefficients on the observed history will deviate from those for the k other histories. The correction for the presence of correlated measurement error in the dependent variable and one of the independent variables is again straightforward.

Scenarios E and F differ, in that not all of the coefficients can be estimated directly. Assumptions about f , the ratio of the principal's information about the component of gains that is orthogonal to the achievement history to the information contained in that history, replace estimates of the prediction coefficients ψ . χ and ϕ are as in scenario D.

6.4.2 Recovering the sorting parameters

With estimates of coefficients for predictor variables with known variance, it is trivial to compute σ_I^2 . The next step is to estimate the extent to which students are sorted across classrooms on the basis of I . I assume that there is some uni-dimensional student-level statistic ω that is observable to the researcher and is contained within Ω (so orthogonal to η). Empirically, I use linear combinations of past test scores for ω .²⁰ The various scenarios pin down $\text{cov}(I, \omega)$. We can recover σ_η^2 by analyzing the between-classroom variance of ω . The within-classroom variance is

$$V(\omega|\lambda) = V(\omega)(1 - \text{corr}^2(\omega, \lambda)) \quad (12)$$

$$= V(\omega) - \frac{\text{cov}^2(\omega, \lambda)}{V(\lambda)} \quad (13)$$

$$= V(\omega) - \frac{\text{cov}^2(\omega, I)}{\sigma_I^2 + \sigma_\eta^2}. \quad (14)$$

Rearranging terms, we obtain

$$\sigma_\eta^2 = \frac{\text{cov}^2(\omega, I)}{V(\omega) - V(\omega|\lambda)} - \sigma_I^2. \quad (15)$$

Note that the denominator here is simply the across-class variance of ω .

6.4.3 Computing the bias

I consider value added models VAM2, VAM3, and VAM4 from Section 5. These are distinguished by the control variables that are included in models for the grade- g gain. In each case, the control variables are subsets of the A vector, so it is straightforward to compute the covariance between these variables and the principal's prediction. As indicated by equation

²⁰I weight past scores based on the weights applied to either true or observed achievement in the principal's prediction. That is, I let $\omega = A_{g-1}\phi_A$ (in scenario A); $\omega = A\phi_j$ (for $j = B, C$); or $\omega = A\chi_j$ (for $j = D, E, F$).

(9), this is sufficient to compute the variance of the bias term, $V(T\kappa)$.²¹

7 Results

7.1 The principal's prediction

Table 2 presented prediction models for the grade-5 test score as a function of test scores in earlier grades. As discussed above, these are readily converted into predictions of grade-5 gains given the observed achievement history. Columns 1 and 2 of Table 7 present the prediction coefficients, when the predictor variables are the 4th grade reading score (column 1) or the sequence of three prior reading scores (column 2).

These models overstate the value of prior test scores for predicting true gains, net of measurement error. This is because the noisy 4th grade test score achieves predictive power for the observed 5th grade gain due to the presence of the same measurement error, u_4 , in both variables. As discussed above, standard errors-in-variables formulae can be used to obtain the best prediction equations for true gains. These are presented in columns 3 and 4 of Table 7.²² The within-school R^2 statistics and especially the prediction coefficients themselves are reduced in magnitude from the specifications for observed gains. True achievement gains are negatively correlated with past achievement levels, but not dramatically so.²³ The model for observed gains in column 2 implicitly attaches a coefficient of around -0.81 ($= -0.57 - 0.24$) to the 4th grade gain, while the corresponding model for true gains assigns a weight of only -0.02 ($= -0.07 - -0.05$) to this gain.

Table 8 presents estimates of the coefficients that the principal would apply to the available predictor variables in scenarios C and D. Columns 1-4 show prediction coefficients

²¹Extending the analysis to VAMs that control for non-test variables requires assumptions about the relationship between these variables and the q and W variables seen by the principal. I do not pursue this here.

²²In principle, the coefficients of regressions that include math scores could be recovered as well.

²³Note that this implies that value added models which impose a coefficient of 1 on the lagged achievement level, as in VAM2, are mis-specified.

for observed gains, while columns 5-8 show coefficients for true gains. Columns 1 and 5 repeat the coefficients from scenario B, where only the observed test score history is available. Columns 2 and 6 show coefficients when a second, equally noisy series is available. Columns 3 and 7 show coefficients when two series are available in addition to observed scores (i.e. $k = 2$). Note that the coefficients on the observed and unobserved series are identical in columns 6 and 7, where the dependent variable is the true gain, but that they differ in columns 2 and 3, where the observed series can be used to predict some of the measurement error in the observed gain. Columns 4 and 8 show predictions assuming that the principal is able to observe the history of true achievement. This substantially improves his ability to predict observed gains, as the measurement error portion of the 4th grade score can be perfectly isolated, but adds relatively little to his ability to predict true gains over what could be done with three noisy histories.

7.2 The importance of predictions in classroom assignments

Using the coefficients from Tables 2, 7, and 8 and relying on an observed component of the principal's predictions, I can compute the variance decomposition of the principal's predictions, I , into within- and between-classroom components. For scenario C, I present estimates for $k = 1$ and $k = 2$. In scenario E, I present estimates for $f = 0.25$, $f = 0.5$, and $f = 1$; the scenario of perfect information, F, corresponds to $f = 1.96$.

The first column of Table 9 shows the fraction of the within-school variance in gains that the principal is able to predict (i.e. $\sigma_I^2/V(Y)$) in each scenario, for true gains in the first panel and for observed gains in the second panel. The second column shows the across-school share of variance for the scenarios in which I is perfectly observable. Coefficients for between-school predictions may differ from those for the within-school predictions that I focus on, so I do not compute the across-school component of the incompletely observed indices. Column 3 shows the fraction of the within-school variation that is across class-

rooms. This equals $\frac{\sigma_I^2}{\sigma_I^2 + \sigma_{\eta}^2}$, the weight placed on predicted outcomes relative to other factors in classroom assignments. Column 4 shows the across-classroom standard deviation in predicted gains.

Not surprisingly, the scenarios in which the principal has more information permit him to explain a larger share of the within-school variance in gains. Moreover, the richer prediction scenarios yield larger estimates of the across-classroom share of variance of predicted gains. Thus, the more information that we permit the principal to have about the student's achievement history, the larger is the bias that is implied for value added specifications (like VAM2) that do not allow for across-classroom sorting.

Sorting appears to be substantially more important when the principal is presumed to be using predictions of observed rather than true gains for classroom assignments. But this can be misleading: True gains are much less variable than observed gains (with a standard deviation less than half as large). Disparities between the panels are smaller in column 3, showing the fraction of the variance of predicted gains that is across classrooms. Even in this column, though, scenarios C-E show more sorting in the second panel. This is because observed scores form a smaller share of predicted observed gains than of predicted true gains in these scenarios (compare the R-squared statistic in column 1 of Table 8 to those in columns 2-4, versus that in column 5 and those in 6-8), so the same sorting on observed variables corresponds to more overall sorting in the observed gain scenarios.

7.3 Bias in value added models with controls for observables

Table 9 shows that the standard deviation of across-classroom differences in predicted gain scores ranges from 0.037 to 0.191, depending on the assumptions made about the information used in sorting. This variation is bias in specifications like VAM2 that do not control for classroom assignments. By comparison, the total across classroom standard deviation of observed gain scores is 0.134. Thus, even scenarios that restrict the principal to use little

more than the observed variables in classroom assignments indicate biases in simple value added models that are large relative to the effects that we hope to measure.

Table 10 presents estimates of the standard deviation of the bias in richer models that include controls for students' prior achievement. Columns 1-3 of this table index value added models, corresponding to VAMs 2, 3, and 4, respectively. The bias in the simplest model (VAM2) is substantial in every scenario. Column 2 shows that the inclusion of a control for the prior year's test score eliminates much of the bias in VAM2, though there is important variation across scenarios. If we assume that the principal forms classroom assignments on the basis of his predictions of true gains (rather than observed gains) *and* that he has no information about students' potential gains beyond that contained in their achievement histories (as in scenarios B-D), the remaining bias in VAM3 is negligible. However, if we allow the principal to have additional information or if we assume that he sorts on the basis of predicted *observed* gains – as he likely would if accountability policies condition rewards and punishments on observed gains rather than on unmeasurable true gains – then the bias remains important. If the principal observes even two independent achievement histories (e.g. the test score history plus an additional series, perhaps coming from teacher grades) and uses them in classroom assignments, the standard deviation of the bias in VAM3 is 0.043.

Column 3 shows that much of the bias in VAM3 remains in VAM4, which controls for the full sequence of prior test scores. If the principal is assumed to observe the student's true achievement history plus another set of variables that explain an equal amount of student gains (i.e. scenario E with $f = 1$), the standard deviation of the bias ranges from 0.051 to 0.076, both large relative to the standard deviation of teachers' estimated "effects."

8 Conclusion

Typical value added analyses treat the process by which students are assigned to teachers as ignorable, under the implicit assumption that the statistical model used can absorb any systematic non-random assignment. This would be true if, for example, classroom assignments were random conditional on students' prior-grade test scores. But there is little reason to think that this is an adequate characterization of classroom assignments. Principals have a great deal of information beyond the prior test score that is predictive of students' end-of-year achievement, and this information is unlikely to be ignored in classroom assignments.

This paper attempts to quantify the bias that arises in value added models that fail to control for the determinants of classroom assignments. The task is straightforward if classroom assignments are assumed to be random conditional on observable variables. My analysis indicates that simple VAMs that fail to control for the dynamic process of test scores, simply modeling differences in mean gain scores across classrooms, are substantially biased by student sorting. The bias is reduced – with a variance about 15% as large as that of teachers' true effects – in a VAM that controls for the lagged score, and is further reduced when additional lagged scores are included as controls.

The analysis is more complex if we loosen the unrealistic assumption that all of the information considered by the principal in forming teacher assignments is available in the research dataset. I develop methods for assessing the bias when the principal is assumed to have access to a limited amount of information that the researcher cannot observe. I consider several scenarios for the information set, and estimate the bias in three value added models under each scenario.

A great deal turns out to depend on *how* the principal uses his information: If he weights past achievement to best predict observed gains, even a limited amount of unobserved information generates substantial biases in the sorts of value added models that are commonly used. Richer models that control the full test score history rather than just a single lagged

score reduce these biases, but only if the principal has very limited information about students' potential. With less restrictive assumptions, biases remain quantitatively important even in rich value added models.

Three recent studies have provided evidence that appears to validate observational value-added estimates. On closer examination, however, all are consistent with the presence of substantial bias in these estimates. Jacob and Lefgren (2008) and Harris and Sass (2007) compare value added estimates with principals' subjective assessments of teacher quality, which might be assumed to reflect unbiased estimates of teachers' causal effects. Both papers find that the two measures are correlated, though far from perfectly. This indicates that there is at least some signal in the value added estimates. But the weak correlations leave plenty of room for non-causal factors in the VAM estimates.

Kane and Staiger (2008) compare estimates of teacher effects from a randomized experiment with observational estimates based on data prior to the experiment. They test the hypothesis that the (appropriately shrunken) observational estimate is an unbiased prediction of the causal estimate, and obtain estimates consistent with this hypothesis. There are three important sources of slippage here. First, Kane and Staiger test a statistical hypothesis about the joint distribution of the true coefficients and the bias; while zero bias is consistent with the null hypothesis, so are large biases that are negatively correlated with teachers' true causal effects.²⁴ Second, Kane and Staiger's sample provides low power. Their standard errors are consistent with substantial attenuation of the prediction coefficient due to bias in the observational estimates. While their confidence intervals might rule out my scenario F (if biases are assumed to be uncorrelated with true quality), my more realistic scenarios are wholly consistent with the Kane and Staiger estimates but are nevertheless extremely troubling regarding the potential for bias in value added estimates. Finally, the Kane and

²⁴They test the hypothesis that $\frac{\text{cov}(\beta, b)}{V(b)}=1$, where β is the vector of causal effects and b is the best linear predictor of $\beta + \kappa$, the sum of causal effects and any bias, based on the coefficients from the value added model. This equality will hold either if $V(\kappa) = 0$ – i.e., there is no bias – or if $\text{corr}(\beta, \kappa) = -\sqrt{V(\kappa)/V(\beta)}$.

Staiger analysis is based on a carefully selected sample of pairs of teachers for which principals consented to random assignment. One might expect that principal consent was more likely when the two teachers would have been given similar students in any case. If so, the results cannot be generalized beyond the sample, even to other teachers at the same schools.

The results here suggest that it is hazardous to interpret typical value added estimates as indicative of causal effects. Although some assumptions about the assignment process permit nearly unbiased estimation, other plausible assumptions yield large biases. Further evidence on the process by which students are assigned to classrooms is needed before it will be clear which types of assumptions are closest to reality. The most recent such study, Monk (1987), is now more than twenty years old. More recent evidence, from studies more directly targeted at the assumptions of value added modeling, is badly needed, as are richer VAMs that can account for real world assignments. In the meantime, causal claims will be tenuous at best.

References

- Aaronson, Daniel, Lisa Barrow, and William Sander**, “Teachers and Student Achievement in the Chicago Public High Schools,” *Journal of Labor Economics*, January 2007, 24 (1), 95–135.
- 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.
- Ballou, Dale**, “Sizing Up Test Scores,” *Education Next*, Summer 2002, 2 (2), 10–15.
- , **William Sanders, and Paul Wright**, “Controlling for Student Background in Value-Added Assessment of Teachers,” *Journal of Educational and Behavioral Statistics*, Spring 2004, 29 (1), 37–65.
- Bazemore, Mildred**, “North Carolina Reading Comprehension Tests,” Technical Report (Citable Draft), Office of Curriculum and School Reform Services, North Carolina Department of Public Instruction. September 22, 2004.

- Bock, R. Darrell and Richard Wolfe**, “Audit and review of the Tennessee Value-Added Assessment System (TVAAS),” Final Report, Office of Education Accountability, Comptroller of the Treasury. March 1996.
- Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff**, “The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools,” Working Paper 10, Center for Analysis of Longitudinal Data in Education Research. September 2007.
- 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.
- Goldhaber, Dan**, “Everyone’s Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?,” Working Paper 9, Center for Analysis of Longitudinal Data in Education Research. April 2007.
- Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger**, “Identifying Effective Teachers Using Performance on the Job,” Discussion Paper 2006-01, The Hamilton Project. April 2006.
- Harris, Douglas N. and Tim R. Sass**, “Value-Added Models and the Measurement of Teacher Quality,” April 2006. Unpublished manuscript.
- and —, “What Makes for a Good Teacher and Who Can Tell?,” July 2007. Unpublished manuscript.
- Jacob, Brian A. and Lars Lefgren**, “Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education,” *Journal of Labor Economics*, January 2008, 25 (1), 101–136.
- Kane, Thomas J. and Douglas O. Staiger**, “Are Teacher-Level Value-Added Estimates Biased? An Experimental Validation of Non-Experimental Estimates,” March 17, 2008. Unpublished manuscript.
- , **Jonah E. Rockoff, and Douglas O. Staiger**, “What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City,” *Economics of Education Review*, forthcoming.
- Koedel, Cory and Julian R. Betts**, “Re-Examining the Role of Teacher Quality in the Educational Production Function,” Working paper 07-08, University of Missouri Department of Economics April 2007.
- Martineau, Joseph A.**, “Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability,” *Journal of Educational and Behavioral Statistics*, Spring 2006, 31 (1), 35–62.

- Monk, David H.**, “Assigning Elementary Pupils to Their Teachers,” *Elementary School Journal*, November 1987, 88 (2), 167–187.
- Rivkin, Steven G., Eric A. Hanushek, and John F. Kain**, “Teachers, Schools, and Academic Achievement,” *Econometrica*, March 2005, 73 (2), 417–458.
- Rothstein, Jesse**, “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” May 2008. Unpublished manuscript, Princeton University.
- Sanders, William L. and June C. Rivers**, “Cumulative and Residual Effects of Teachers on Future Student Academic Achievement,” Research Progress Report, University of Tennessee Value-Added Research and Assessment Center, November 1996.
- **and Sandra P. Horn**, “The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment,” *Journal of Personnel Evaluation in Education*, October 1994, 8 (3), 299–311.
- **and —**, “Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research,” *Journal of Personnel Evaluation in Education*, September 1998, 12 (3), 247–256.
- **, Arnold M. Saxton, and Sandra P. Horn**, “The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment,” in Jason Millman, ed., *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure*, Thousand Oaks, CA: Corwin, 1997, pp. 137–162.
- Sanford, Eleanor E.**, “North Carolina End-of-Grade Tests: Reading Comprehension, Mathematics,” Technical Report #1, Division of Accountability/Testing, Office of Instructional and Accountability Services, North Carolina Department of Public Instruction 1996.
- Yen, Wendy**, “The Choice of Scale for Educational Measurement: An IRT Perspective,” *Journal of Educational Measurement*, Winter 1986, 23 (4), 299–325.

Table 1A: Summary statistics and correlations for reading test scores and gains

	Score levels				5th grade
	Grade 3 pretest	Grade 3	Grade 4	Grade 5	gain
	(1)	(2)	(3)	(4)	(5)
Mean	-0.82	0.07	0.42	1.05	0.63
SD	0.87	0.96	0.95	0.82	0.55
Correlations					
Grade 3 pretest	1	0.70	0.69	0.65	-0.23
Grade 3 end-of-grade	0.70	1	0.80	0.77	-0.25
Grade 4 end-of-grade	0.69	0.80	1	0.81	-0.52
Grade 5 end-of-grade	0.65	0.77	0.81	1	0.07
Grade 5 gain	-0.23	-0.25	-0.52	0.07	1

Notes: N=49,453

Table 1B: Summary statistics and correlations for reading achievement levels and growth, net of sampling error

	Score levels				5th grade
	Grade 3 pretest	Grade 3	Grade 4	Grade 5	gain
	(1)	(2)	(3)	(4)	(5)
Mean	-0.82	0.07	0.42	1.05	0.63
SD	0.74	0.88	0.88	0.75	0.27
Correlations					
Grade 3 pretest	1	0.91	0.89	0.84	-0.56
Grade 3 end-of-grade	0.91	1	0.96	0.92	-0.57
Grade 4 end-of-grade	0.89	0.96	1	0.96	-0.59
Grade 5 end-of-grade	0.84	0.92	0.96	1	-0.33
Grade 5 gain	-0.56	-0.57	-0.59	-0.33	1

Notes: N=49,453

Table 2: Predictability of grade 5 reading scores from prior information

	(1)	(2)	(3)	(4)	(5)
Grade 4 reading score		0.680 (0.003)	0.430 (0.004)	0.356 (0.005)	0.347 (0.005)
Grade 3 reading score			0.245 (0.004)	0.196 (0.004)	0.186 (0.004)
Grade 3 pretest score, reading			0.082 (0.003)	0.066 (0.003)	0.063 (0.003)
Grade 4 math score				0.120 (0.005)	0.109 (0.005)
Grade 3 math score				0.045 (0.005)	0.041 (0.005)
Grade 3 pretest score, math				0.020 (0.005)	0.017 (0.005)
Non-test covariates	n	n	n	n	y
N	49,453	49,453	49,453	49,409	49,285
Goodness-of-fit measures					
Models for G5 achievement					
R2	0.131	0.683	0.722	0.733	0.736
R2, within school	n/a	0.635	0.680	0.693	0.696
R2, within school, for true achievement	n/a	0.764	0.819	0.834	0.837
Models for G5 gains					
R2	0.047	0.313	0.397	0.421	0.427
R2, within school	n/a	0.279	0.367	0.392	0.398

Notes: All columns include fixed effects for 838 schools, and standard errors are clustered at the school level. "Non-test covariates" in column (5) include indicators for gender, for race/ethnicity, for learning disabilities in reading or in any area, for Title 1 participation, for each possible "exceptionality" (gifted, hearing impaired, mentally handicapped, etc.), for parental years of education, for free and for reduced-price lunch participation, for reporting never doing any homework; and a linear control for the number of hours of TV watched each school day (plus a dummy for missing values for this variable).

Table 3: Prediction models with past gains as predictors

	Dependent variable					
	Grade 5 reading score			Grade 5 reading gain		
	(1)	(2)	(3)	(4)	(5)	(6)
Grade 4 reading gain	0.051 (0.007)	0.082 (0.007)	0.430 (0.004)	-0.394 (0.005)	-0.410 (0.005)	-0.570 (0.004)
Grade 4 math gain		-0.130 (0.008)			0.067 (0.005)	
Grade 3 reading gain			0.675 (0.004)			-0.325 (0.004)
Grade 3 pretest score, reading			0.757 (0.003)			-0.243 (0.003)
Goodness-of-fit measures						
R2	0.132	0.140	0.722	0.221	0.225	0.397
R2, within school	0.002	0.010	0.680	0.182	0.186	0.367

Table 4. Simple models for 5th grade teachers' "effects" on gains in grades 3, 4, and 5

	Gain score measured in:		
	Grade 5	Grade 4	Grade 3
	(1)	(2)	(3)
<i>Standard deviation of normalized teacher coefficients</i>			
Unadjusted for sampling error	0.152	0.142	0.170
Adjusted for sampling error	0.107	0.080	0.097
<i>Correlations, unadjusted for sampling error</i>			
	Grade 5	Grade 4	Grade 3
Grade 5	1	-0.39	-0.06
Grade 4	-0.39	1	-0.40
Grade 3	-0.06	-0.40	1
<i>Correlations, adjusted for sampling error</i>			
	Grade 5	Grade 4	Grade 3
Grade 5	1	-0.35	-0.08
Grade 4	-0.35	1	-0.36
Grade 3	-0.08	-0.36	1

Notes: All specifications include fixed effects for 5th grade schools and for 5th grade teachers, normalized to mean zero at each school; only the dependent variable changes. Sample excludes 111 teachers with fewer than 10 sample students each. The remaining sample has 49,235 students, 2,733 teachers, 784 schools. Correlations are between teacher coefficients in the three specifications, weighted by the number of students taught and adjusted for the degrees of freedom absorbed by the school-level normalization.

Table 5. Bias in simple value added specifications if classroom assignment is random conditional on observables

	SD of teacher coefficients	SD of bias	Bias variance / total (correct) variance	corr(bias, true effect)
	(1)	(2)	(3)	(4)
<i>Panel 1: Unadjusted for sampling error</i>				
Control for all observables	0.124	0		
Levels, no controls (VAM1)	0.251	0.208	279%	0.09
Gain scores, no controls (VAM2)	0.153	0.095	59%	-0.05
Control for lagged score (VAM3)	0.137	0.050	16%	0.06
Control for score history (VAM4)	0.128	0.025	4%	0.06
<i>Panel 2: Adjusted for sampling error</i>				
Control for all observables	0.096	0.000		
Levels, no controls (VAM1)	0.208	0.171	318%	0.14
Gain scores, no controls (VAM2)	0.114	0.070	53%	-0.09
Control for lagged score (VAM3)	0.106	0.035	13%	0.11
Control for score history (VAM4)	0.100	0.018	3%	0.11

Notes: Specification that controls for "all observables" includes controls for math and reading scores in grades 2, 3, and 4; indicators for gender, for race/ethnicity, for learning disabilities in reading or in any area, for Title 1 participation, for each possible "exceptionality" (gifted, hearing impaired, mentally handicapped, etc.), for parental years of education, for free and for reduced-price lunch participation, and for reporting never doing any homework; and a linear control for the number of hours of TV watched each school day (plus a dummy for missing values for this variable).

Table 6. Scenarios for the principal's information about student gains

Scenario	Principal's information set	
Selection on observables		
A	A_{g-1}	Principal observes only the prior test score
B	$A=\{A_1, \dots, A_{g-1}\}$	Principal observes full history of test scores
Selection on observed and some unobserved variables		
C	$\{A, q^1, \dots, q^k\}$	Principal observes history of test scores plus k addl. sequences, each a noisy measure of true achievement in grades 1, ..., g-1.
D	$\{A^*, A_{g-1}\}$	Principal observes true achievement history (without measurement error) plus observed prior test score
E	$\{A^*, A_{g-1}, W\}$	Principal observes true achievement history, observed prior score, and an additional measure that is predictive of $A^* - E[\Delta A^* A^*]$
Selection on unobservables is like selection on observables		
F	$\{Y\}$	Principal is able to perfectly predict student outcomes

Table 7: Models for observed and true (measured without error) grade-5 reading gains

Scenario:	Dependent variable			
	Observed gains		True achievement gains	
	A	B	A	B
	(1)	(2)	(3)	(4)
Grade 4 reading score	-0.320 (0.003)	-0.570 (0.004)	-0.150 (0.003)	-0.073 (0.004)
Grade 3 reading score		0.245 (0.004)		-0.055 (0.004)
Grade 3 pretest score, reading		0.082 (0.003)		-0.051 (0.003)
R2	0.313	0.397	0.449	0.484
R2, within school	0.279	0.367	0.273	0.312

Notes: See text for computational details. Standard errors treat the test reliability as known perfectly. In practice, this is estimated, likely with substantial sampling and non-sampling error.

Table 8. Prediction weights if principal has more information than just the observed test score history

Scenario:	Predictions of observed gains				Predictions of true gains			
	B	C	C	D	B	C	C	D
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Observed test score history								
Grade 4	-0.57	-0.74	-0.81	-1.00	-0.07	-0.04	-0.03	0.00
Grade 3	0.24	0.11	0.07	0.00	-0.06	-0.03	-0.02	0.00
Grade 2	0.08	0.02	0.00	0.00	-0.05	-0.03	-0.02	0.00
Grade 4 (math)								
Grade 3 (math)								
Grade 2 (math)								
Second noisy achievement history								
Grade 4		0.26	0.19			-0.04	-0.03	
Grade 3		0.11	0.07			-0.03	-0.02	
Grade 2		0.02	0.00			-0.03	-0.02	
Third noisy achievement history								
Grade 4			0.19				-0.03	
Grade 3			0.07				-0.02	
Grade 2			0.00				-0.02	
History of true achievement								
Grade 4				0.90				-0.10
Grade 3				0.00				0.00
Grade 2				-0.10				-0.10
R2 (within school)	0.37	0.44	0.46	0.54	0.31	0.33	0.33	0.35

Notes: All coefficients are for within-school predictions. See text for details of computations.

Table 9: Variance decompositions for actual and predicted grade-5 gains

Predicted variable Scenario Predictor variables	Explained share of within- school variance (1)	ANOVA for predicted gains		
		Across- school share (2)	Fr. of within- school variance that is across classrooms (3)	SD of across- class, within- school component (4)
<i>True gain</i>				
A Using grade 4 score	27.3%	12.3%	7.3%	0.037
B Using 3 prior scores	31.2%	12.3%	7.5%	0.040
C Using 2 independent achievement histories	32.7%	--	7.8%	0.041
C Using 3 independent achievement histories	33.2%	--	7.9%	0.041
D Using true achievement history	34.5%	--	8.4%	0.043
E Using true history & W variable (f=0.25)	43.2%	--	10.5%	0.054
E Using true history & W variable (f=0.5)	51.8%	--	12.5%	0.065
E Using true history & W variable (f=1)	69.1%	--	16.7%	0.087
F Using perfect information (f=1.96)	102%	--	24.8%	0.129
<i>Observed gain</i>				
A Using grade 4 score	27.9%	12.3%	7.3%	0.079
B Using 3 prior scores	36.7%	9.3%	5.9%	0.082
C Using 2 independent achievement histories	43.5%	--	9.8%	0.111
C Using 3 independent achievement histories	46.2%	--	11.2%	0.123
D Using obs. & true achievement histories	54.0%	--	14.1%	0.149
E Using true history & W variable (f=0.25)	55.9%	--	14.6%	0.155
E Using true history & W variable (f=0.5)	57.9%	--	15.1%	0.160
E Using true history & W variable (f=1)	61.7%	--	16.1%	0.171
F Using true gains (f=1.96) plus obs scores	69.1%	--	18.1%	0.191
Grade 5 gain (observed)	1	4.7%	5.8%	0.134

Table 10. Bias in value added measures if information is used in teacher assignments that is not observed by the researcher

	Value added model includes controls for:		
	Teachers only (VAM2)	Lagged score (VAM3)	Score history (VAM4)
	(1)	(2)	(3)
	SD of teachers' estimated effects		
Unadjusted for sampling error	0.153	0.137	0.128
Adjusted for sampling error	0.114	0.106	0.100

If classroom assignments depend on predictions of true gains

Scenario	SD of bias		
B Using observed achievement history	0.039	0.005	0.000
C Using 2 independent achievement histories	0.041	0.007	0.002
C Using 3 independent achievement histories	0.041	0.008	0.003
D Using true achievement history	0.043	0.010	0.004
E Using true history & W variable (f=0.25)	0.054	0.021	0.016
E Using true history & W variable (f=0.5)	0.065	0.033	0.028
E Using true history & W variable (f=1)	0.087	0.056	0.051
F Using perfect information (f=1.96)	0.126	0.098	0.094

If classroom assignments depend on predictions of observed gains

Scenario	SD of bias		
B Using observed achievement history	0.080	0.020	0.000
C Using 2 independent achievement histories	0.111	0.043	0.019
C Using 3 independent achievement histories	0.123	0.052	0.028
D Using true achievement history + obs. scores	0.149	0.078	0.053
E Using true history & W variable (f=0.25)	0.155	0.084	0.059
E Using true history & W variable (f=0.5)	0.160	0.089	0.065
E Using true history & W variable (f=1)	0.171	0.101	0.076
F Using true gains (f=1.96) plus observed scores	0.191	0.123	0.099