Stacy, Brian

Preprint

Ranking Teachers when Teacher Value-Added is Heterogeneous Across Students

Ranking Teachers when Teacher Value-Added is Heterogeneous Across Students

Brian Stacy

December 8, 2014

Abstract

The typical measure used by researchers and school administrators to evaluate teachers is based on how the students' achievement increases after being exposed to the teacher, or based on the teacher's "value-added". When teacher value-added is heterogeneous across her students, the typically used measure reflects differences in the average value-added the teacher provides. However, researchers, administrators, and parents may care not just about the average value-added, but also its dispersion. In this paper, I examine the robustness of typical teacher quality measures to alternate ranking systems factoring in the variance of value-added. Encouragingly, ranking systems factoring in the variance produce similar rankings as the ranking system based only on the mean. I also examine whether classroom characteristics and teacher experience affect a teacher's value-added variance and find that they explain little of the variation in value-added variances.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305D100028 and R305B090011 to Michigan State University. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education.

1 Introduction

Teacher quality measures based on student achievement data are increasingly being utilized by researchers in topics ranging from the impact of teacher quality on later life outcomes, to the impact of teacher quality on housing prices, to the quality of teachers who transfer or leave the teacher labor force (see, e.g., Chetty, Friedman, & Rockoff (2014b), Imberman & Lovenheim (2013), or Boyd, Grossman, Lankford, Loeb, & Wyckoff (2008) for examples of each). Additionally, federal education policies, such as the Teacher Incentive Fund and the Race to the Top, have sparked substantial demand for rigorous measures of teacher quality by administrators who wish to identify the most and least effective teachers. The most commonly used measures of teacher quality are value-added measures that attempt to isolate a teacher's contribution to student learning in a year.

Some studies make the simplifying assumption that teacher value-added is identical for all students.[1] With this assumption, a "teacher effect" can be estimated for each teacher, which reflect differences in the value-added provided. Other studies explicitly explore heterogeneity in teacher value-added and find evidence that teacher value-added is different for different students.[2] With het-

[1]The assumption of a constant value-added is explicitly stated in Chetty, Friedman, & Rockoff (2011) for instance, but implicitly assumed in many structural models of achievement used in value-added estimation.

[2]For instance, Dee (2004) examines whether assigning a student to a teacher of the same race improves student achievement using experimental project STAR data, and finds an increase for both black and white students. One year with same race teacher increases achievement 2 to 4 percentile points. Aaronson, Barrow, & Sander (2007) computes teacher value-added separately for students with high and low prior year test scores and finds that the correlation between the two is .39. A similar exercise is done by Condie, Lefgren, & Sims (2014). Loeb, Soland, & Fox (2014) examines whether teachers quality depends on whether a student is an English learner. Lockwood & McCaffrey (2009) examine heterogeneity in teacher value-added by interacting value-added

erogeneity, the "teacher effects" that are typically estimated reflect differences in the mean value-added provided. From here on I will refer to these measures as "value-added means".

Despite the recognition that teacher value-added can be heterogeneous, little work has been done examining teacher quality beyond the value-added means.[3] Teachers may differ in the variance of the value-added they provide, and this information may be important for the researchers and administrators using teacher quality ratings. For example, an individual may view a teacher that produces large learning gains for a few students and small gains for the rest differently from a teacher that produces moderate gains for all students. Examining the variance of value-added in addition to the mean can distinguish between these two cases.

In this paper, I examine the sensitivity of teacher rankings to alternate rankings that factor in the variance of teacher value-added. I estimate "value-added variances", which reflect differences across teachers in the variance of the value-added a teacher provides. These can be identified using the same assumptions made to identify value-added means. I then use this additional information to create alternate rankings, which I compare to the rankings based solely on value-added means. I also examine whether classroom characteristics, teacher experience, and student dissimilarity within classrooms have an effect on the value-added variance for a teacher. These effects are identified using within teacher variation in these characteristics.

with predicted achievement and find modest interaction effects with the interactions explaining around 10% of the total variation in teacher effects across teachers.

[3] Some exceptions include the papers listed in the footnote above.

Using administrative data linking students to teachers from a large, diverse, anonymous state, I find little evidence of a systematic mean-variance trade-off in teacher value-added. The value-added means and variances are in fact negatively correlated (math: -.328, p<.001, reading: -.206, p<.001). I find that there are larger differences across teachers in terms of the mean than the variance. As a result, teacher rankings systems incorporating both value-added means and value-added standard deviations are highly correlated (above .9 in most cases) with a system only comprised of value-added means. I also find that classroom characteristics, teacher experience, and student dissimilarity explain little of the variation in value-added variances across teachers.

2 Framework for Evaluating Teacher Quality

A convenient framework for analyzing teacher quality is the potential outcomes framework.[4] For our purposes, the potential outcomes are the potential achievement outcomes if a student is assigned to any of the possible teachers. Let $A_i(j)$ be the achievement level of student i if they are assigned to a particular teacher j.

Administrators and researchers are typically interested in identifying how students would perform if they were assigned to one teacher compared to another. The primary difficulty in making this type of causal inference is that it is possible to observe only one of the potential outcomes, the outcome for the teacher that the student is actually assigned to.

[4]See Rosenbaum & Rubin (1983), Rubin (1974), Rubin, Stuart, & Zanutto (2004), or Imbens & Wooldridge (2008) for further background.

The key assumption used to make causal inferences about teachers is that assignment of teachers to student is random conditional on X_i, which is a set of observable characteristics of students. With this assumption, even though we do not observe each of the potential outcomes, $A_i(j)$, for a student, we can use the observed outcome, A_i, to estimate teacher effects. This assumption of selection only on observables is sometimes referred to as ignorability (or unconfoundedness) conditional on X_i.[5] This assumption implies that principals base assignment on observable characteristics of students, such as prior year test scores, but do not assign on unobservable factors that affect achievement.

The ignorability assumption has been hotly debated in the value-added literature.[6] Important for my purposes, this assumption is necessary for estimating value-added means. And without further identifying assumptions, we can estimate value-added variances.

A typical way of estimating teacher effects is to estimate the parameters in the

[5] See Imbens (2000) for further discussion.

[6] The assumption is not directly testable. However, Rothstein (2010) develops an indirect falsification test based on the idea that future teachers cannot impact contemporaneous test scores, so evidence of a relationship is evidence of a violation of the assumptions. Rothstein finds that the falsification test rejects, suggesting estimates of teacher effects may be biased. However, Goldhaber & Chaplin (2012) and Guarino, Reckase, Stacy, & Wooldridge (forthcoming) both find that such falsification tests may over reject. Also, Guarino, Reckase, & Wooldridge (2015) produce simulation evidence that estimators flexibly controlling for prior year test scores and teacher fixed effects are fairly robust across a variety of nonrandom assignment scenarios. Chetty, Friedman, & Rockoff (2014a) find that value-added measures controlling for prior year achievement are unbiased using a test that examines bias by including previously unobserved parental characteristics and another test based on examining predicted changes in achievement as teachers switch schools. Also, Chetty et al. (2014b) find that value-added measures predict long term outcomes such as earnings and college attendance. Finally, Kane, McCaffrey, Miller, & Staiger (2013) examine whether value-added estimates are biased using a large randomized experiment in which students were randomly assigned to teachers within schools. The authors find no evidence of bias in estimators that control for a student's prior achievement scores and demographics.

following equation for the conditional mean of achievement:[7]

$$E(A_i|X_i,T_i) = (X_i - \mu_X)\beta + T_{i1}\gamma_1 + \cdots + T_{iJ}\gamma_J. \quad (1)$$

The vector X_i is a set of control variables. The variable T_{i1} is an indicator variable equal to 1 if assigned to teacher 1 and 0 otherwise, T_{i2} is an assignment indicator for teacher 2, and so on, and γ_j is the teacher effect for teacher j.[8] Under the ignorability assumption, the estimates of γ_j are consistent estimates of the value-added means.

In the case where a teacher's value-added varies across students, γ_j does not fully characterize the impact of assigning students to a teacher. There are numerous reasons why teacher value-added may vary. For instance, a teacher's pedagogical style may work well with some students and not others. Also, some teachers may relate better with some students than others, for instance if they are of the same race or gender, which could lead to differences in the value-added provided. Teachers may also deploy more resources at some students than others.[9] Some others may be less able to cater their instruction to the needs of all students in a classroom.

Define σ_j^2 as the value-added variance for teacher j. With γ_j and σ_j^2 we can

[7]For instance see Rothstein (2009) or Harris, Sass, & Semykina (2010). This achievement model is sometimes motivated using the education production function framework. For more details, see Hanushek (1979) or Todd & Wolpin (2003).

[8]In my parameterization, X_i is centered around its mean, μ_x. With this parameterization, γ_j can be interpreted as teacher j's mean level of achievement produced for the average student. A value of zero for γ_j indicates that a teacher produces a mean achievement level of zero for the average student. With this parameterization, no intercept is included.

[9]Neal & Schanzenbach (2010) find evidence that teachers may target resources at students in the middle of the achievement distribution because of proficiency requirements.

get a more complete measure of teacher quality than looking at the mean alone. In order to estimate σ_j^2, assume that the conditional variance of achievement has the following function form:

$$Var(A_i|X_i,T_i) = exp((X_i - \mu_X)\delta + T_{i1}\psi_1 + \cdots + T_{iJ}\psi_J). \quad (2)$$

Again, X_i is centered around its mean, μ_X, so the estimates of the value-added means and variances, γ_j and $\sigma_j^2 = exp(\psi_j)$, can be interpretted as estimates of teacher j's mean and variance of achievement produced for the average student.[10]

In order to estimate γ_j and σ_j^2, I use the following procedure. I first estimate the parameters in Equation (1) by regressing the student achievement scores on $X_i - \bar{X}$ and T_i. Then I form residuals from this initial regression and estimate the parameters in Equation (2) by regressing these squared residuals on $X_i - \bar{X}$ and T_i using non-linear least squares.[11]

[10]For clarity, a value of zero for γ_j indicates that a teacher produces a mean achievement level of zero for the average student, and a value of zero for σ_j^2 indicates that a teacher produces a variance of achievement of zero for the average student.

The exponential function is chosen to model the conditional variance instead of a linear function, because a linear function would not guarantee that the predicted conditional variance is positive. Using the exponential function to model a conditional variance dates back in the econometrics literature to Harvey (1976).

[11]To see why the non-linear least squares regression using the squared residuals can consistently estimate the parameters in the conditional variance, note that $Var(A_i|X_i,T_i) = E(\varepsilon_i^2|X_i,T_i)$ by definition, where $\varepsilon_i = A_i - E(A_i|X_i,T_i)$. Because the OLS residuals converge in distribution to ε_i, as noted in Harvey (1976), using the squared residuals in place of ε_i^2 in the NLS regression still produces consistent estimates of the parameters in $E(\varepsilon_i^2|X_i,T_i)$.

3 Data

The data come from an administrative data set in a large and diverse anonymous state. Basic student information such as demographic, socio-economic, and special education status are available. The data contain 3,341,109 student year observations in grades 3-6 from years 2001-2007. The data include achievement scores in reading and math on a state criterion referenced test. Students and teachers can be linked in the data. The test scores are vertically scaled. The benefit of the vertical scale is that if, for instance, a student scores a 500 in 4th grade and a 550 in 5th grade, then this indicates that the student made a 50 point learning gain from 4th to 5th grade.

The analysis focuses on mathematics and reading student achievement in grade 6. Grade 6 is chosen for two reasons. First, conditioning on a larger number of previous test scores increases the plausibility that assignment of students to teachers is unrelated to student unobservables. Second, teachers in grade 6 often teach multiple sections in a given year, which increases the number of student observations. The larger number of student observations is important for the precision of the estimates. Although I focus on grade 6 as an ideal case, I also present results for grade 5 in the sensitivity checks section.

I impose some restrictions on the data. Students that cannot be linked with a teacher are dropped, as are students linked to more than one teacher in a school year in the same subject. The analysis focuses on traditional public school students, so students in charter schools are dropped. I also drop teachers with less

than 12 student observations because accurately estimating value-added means and variances requires a large number of student observations. In an analysis below, I directly assess how accuracy improves when more students are available. In all around one third of the student observations are not used in the analysis. Student level characteristics of the final data set are reported in the Table 1. The students in the final sample tend to be somewhat higher achieving, more white, and less likely to be free-and-reduced price lunch or limited English proficient than the students in the original sample.

Table 2 reports summary statistics aggregated to the teacher level. There are 5,987 math and 6,606 reading teachers in the sample. There are on average 114.6 and 105.0 student observation per teacher for math and reading teachers respectively. Student characteristics aggregated to the teacher level are also reported.

4 Results

The controls included are similar to other papers in the literature (e.g. Chetty et al. (2014a)). The set of covariates, X_i, includes cubic functions of lagged and twice lagged math and reading scores, indicators for whether the student is a minority, the student's free-and-reduced price lunch status, the student's limited English proficiency status, and gender.

In order to increase the precision of the estimates, I pool student observations across all available years and include year dummies as additional controls. Estimation is done separately for math and reading teachers.

Similar to Rothstein (2009), I standardize test scores so that grade 6 test scores have a population mean of zero and a standard deviation of one. Using the same standardization in each grade keeps the vertical scale intact.[12] Therefore, one test score unit translates into an increase of one standard deviation in achievement for sixth graders.

Based on this, I estimate the measures for the value-added means (γ_j) and the value-added variances (σ_j^2) for the 6,249 mathematics teachers and 6,836 reading teachers. As reported in Table (3), the standard deviation of the estimates of γ_j across teachers is .207 in mathematics and .155 in reading.[13] Additionally, going from the teacher at the 50th percentile in the estimated distribution of γ_j to a teacher at the 75th percentile in mathematics increases mean value-added by .13 test score standard deviations. Going from the 50th to 75th percentile in reading increase mean value-added by .092 standard deviations.

The differences across teachers in σ_j^2 are more modest. The standard deviation of σ_j^2 across teachers is .086 in mathematics and .106 in reading. Going from a teacher at the 50th percentile in the estimated distribution of σ_j^2 to a teacher at the 75th percentile increases the variance of value-added by .043 test score standard deviation units. This is 4.3% of the variance of overall achievement. Going from

[12] With this standardization, grade 5 math test scores have a mean of -.152 and standard deviation of .928. Grade 4 math test scores have a mean of -.763 and standard deviation of .979. Grade 5 reading test scores have a mean of -.209 and standard deviation of .981. Grade 4 reading test scores have a mean of -.413 and standard deviation of .960.

[13] These estimates are in line with what other researchers have found for the standard deviation across teachers for the mean. Kane & Staiger (2010) find a standard deviation adjusted for sampling variation of .143 for mathematics teachers. Aaronson et al. (2007) find an adjusted standard deviation of .193 for mathematics teachers and .113 for reading teachers. Rothstein (2009) finds an adjusted standard deviation of .107 for reading teachers.

the 50th percentile to the 75th percentile in reading means increasing variance of value-added by .054.

In order to provide some information about the precision of the estimates, I report estimates of γ_j and σ_j^2 along with their standard errors for select teachers in Table 4.[14] Estimates and standard errors for teachers at the 10th, 25th, 50th, 75th, and 90th percentiles of γ_j (top panel) and σ_j^2 (bottom panel) are reported. Additionally, Figures 1 and 2 show the 95% confidence intervals and standard errors plotted on the number of student observations for a randomly selected subsample of teachers for math and reading.[15] The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right.[16] An average standard error at each number of student observations for the OLS estimates of γ_j is displayed in the bottom left, and the standard errors for the NLS estimates of σ_j^2 are in the bottom

[14]I use a bootstrapping technique to produce standard errors for the estimates of γ_j and σ_j^2. In order to keep the number of student observations per teacher fixed for every bootstrap replication, I do sampling with replacement within teachers. To be clear, if there are N observations and N_j observations corresponding to teacher j in the original data set, to produce N observations for each bootstrap sample, draw N_j observations for teacher j, where the N_j observations are randomly drawn with replacement from the set of students assigned to the teacher, and repeat this procedure for all teachers. 100 bootstrap replications were performed. Since estimation of σ_j^2 involves two steps (first forming residuals after an OLS regression of current achievement on the covariates and teacher indicators then NLS of the squared residuals on the covariates and teacher indicators) each bootstrap iteration involves estimation of both steps. The sampling with replacement of teachers done in this paper is similar to bootstrapping approach done in Winters, Dixon, & Greene (2012).

[15]The randomly selected subsamples of 584 mathematics teachers and 743 reading teachers were used instead of the entire sample, because the bootstrapping procedure was very time intensive.

[16]I also try a procedure based on Normal quasi-MLE to estimate γ_j and σ_j^2. I parameterize the mean and variance of the normal distribution so that

$$D(A_i|X_i,T_i) = Normal((X_i-\mu_X)\beta + \mathbf{T_i}\gamma, exp(\mathbf{T_i}\psi + (X_i-\mu_X)\delta)) \quad (3)$$

The estimates were similar in the two approaches, although the QMLE results were slightly more efficient. Since the QMLE is more complex and more computationally difficult to implement, I chose to present the results for the simpler two-step estimator.

right.[17]

As expected, the estimates become more precise as more student observations are available for each teacher. This is evident in Figures 1 and 2 by noticing that both the confidence intervals and standard errors shrink as the number of student observations increase. Also, the magnitudes of the standard errors do not differ much for γ_j and σ_j^2.

In the bottom panels of Figures 1 and 2, I also include cutoffs for whether the measures are accurate enough to distinguish top and bottom performing teachers, which is often a goal of forming teacher quality measures. The upper blue line represents the standard error necessary to say with 95% accuracy that a teacher ranked in the bottom 10% is not in the top 10%. The lower blue line represents the standard error necessary to say that a teacher ranked in the bottom 25% is not in the top 75%.[18] The average standard errors should be below the cutoffs. 12 student observations are enough typically to distinguish teachers at the top 10% and the bottom 10% for both γ_j and σ_j^2 in mathematics and reading. This can be seen by noting that the average standard error is below this cutoff in all cases at 12 students. When going to the tougher requirement of distinguishing teachers at the 75th and 25th percentiles, 12 student observations is only enough in the

[17] The average standard error at each number of student observations was formed using a polynomial smoother.

[18] I form the blue lines by calculating the difference in γ_j and σ_j^2 at the 90th and 10th percentiles and the 75th and 25th percentiles. For math and γ_j, the 90-10 difference is .48 and the 75-25 difference is .254. In reading and γ_j, the 90-10 difference is .374 and the 75-25 difference is .183. In math and σ_j^2, the 90-10 difference is .176 and the 75-25 difference is .079. In reading and σ_j^2, the 90-10 difference is .223 and the 75-25 difference is .147. I then form the standard error necessary at each of the gaps, by dividing the gap by 1.96.

case of mathematics for γ_j. Approximately 40 student observations is enough in reading at the 75-25 difference for γ_j, and approximately 66 observations is enough in reading for σ_j^2. Around 316 are necessary in mathematics to distinguish between the 75th and 25th percentiles in σ_j^2. This is partially due to the smaller difference in the value-added variances for mathematics teachers at the 75th and 25th percentiles compared to for instance the value-added variances for reading teachers (a gap of .079 versus .147 in reading). Overall, 12 student observations are enough to distinguish top and bottom performing teachers for both γ_j and σ_j^2, but in some cases it may be difficult to distinguish teachers toward the center of the distribution without large numbers of student observations.

4.1 Correlation between γ_j and σ_j^2

A worry in using only estimates of γ_j in rankings is that teachers that produce high value-added means may be leaving some students behind, producing small gains for these students. In order to examine whether this is the case, in Table (3) I report the correlation between $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$, which is -.328 for mathematics and -.206 for reading. Scatterplots for the estimates of γ_j and σ_j^2 are also shown in Figures (3) and (4). In both cases the correlation is statistically different from 0 at the 1% level.[19] This indicates that teachers with higher levels of mean value-added tend also to have a lower variance in value-added. Therefore, if having a low variance is a good thing, teachers rated favorably along one dimension are more likely to

[19]The standard errors for the significance test for the correlations are calculated by bootstrapping.

be rated favorably along the other. Moreover, because there are fewer differences across teachers in σ_j^2 than in γ_j, then rankings that incorporate information on the teacher's effect on the variance may not differ much from a ranking based solely on the mean effect.

4.2 Do Teacher Rankings Change When We Add Information on Value-Added Variances under Plausible Teacher Ranking Functions?

Principals or administrators may be interested in ranking teachers at least in part on the variance of value-added. A teacher that produces a given mean level of value-added, but with a high variance, may generate more complaints from parents than a teacher that produces a similar mean level and a lower variance. Administrators may also have asymmetric payoffs, for instance if they are penalized for having a certain number of students fall below basic proficiency levels, that may make them rate the slightly lower mean, lower variance teacher more highly.[20]

In the following section I produce teacher rankings under a variety of ranking schemes. I use value-added standard deviations in the ranking function rather than variances, because standard deviations are expressed in the same units as the mean, whereas the variance is expressed in squared units.[21] I use the following

[20]There may be cases where individuals would prefer a higher variance. For instance, if a school's sole focus was to produce a few super star students, they would want teachers to have a large variance.

[21]The value-added standard deviations are estimated by taking the square root of the estimated value-added variances.

simple ranking function:[22]

$$r_j = q\hat{\gamma}_j - (1-q)\hat{\sigma}_j.$$

where r_j is teacher j's ranking and q is a weight put on the value-added mean and value-added standard deviation. I will compare three alternate ranking systems to the rankings based only on γ_j:

Baseline Ranking: Teacher rankings are based solely on the estimate of γ_j

25% on σ_j: Teacher rankings based 75% on estimate of γ_j and 25% on estimate of σ_j

33% on σ_j: Teacher rankings based 67% on estimate of γ_j and 33% on estimate of σ_j

50% on σ_j: Teacher rankings based equally on estimate of γ_j and σ_j

I produce Spearman rank correlations between the baseline ranking system and the three alternate rankings systems in in Table (5). The rank correlations are above .94 in mathematics, and above .88 in reading. All rankings are above .96 when less than 1/3 of the weight is placed on the value-added standard deviation

[22]There are many other potential objective functions, which may not translate exactly into a mean-variance trade off. For instance, a principal may want to maximize the number of students that pass a proficiency level, and suppose that principal wants to assign a teacher to a classroom of students that is initially far below the proficiency level. The principal in this case may want a teacher that produces a large variance in value-added to get more students up to that proficiency level. However, I chose the ranking function in this paper for its simplicity.

and above .98 when less than 25% of the weight is placed on σ_j. Thus, incorporating σ_j into teacher rankings isn't likely to dramatically alter the rankings for most teachers under a variety of alternative ranking systems compared to ranking teachers solely on their value-added mean. This result is likely driven by the negative correlation between the value-added mean and standard deviation and the more modest variation in σ_j compared to γ_j.

To add some comparison to the numbers, Goldhaber, Walch, & Gabele (2013) compare teacher rankings, based on value-added means, under alternate sets of control variables. The authors find that the correlation between estimates that control for student test scores and demographics and estimates that control for additional peer characteristics is around .99. The correlation between estimates that control for school fixed effects and estimates that do not is only .65. This suggests that the decision to include information on the value-added standard deviation is slightly more consequential than the decision to include peer variables, and much less important than the decision to include school fixed effects.

One caveat is that, even though the correlations are strong, for particular teachers changing the ranking system can have a large impact. In order to provide a rough sense of how far a teacher may be moving using the different rankings, in the bottom panel of Table (5), I report the fraction of teachers that move in the rankings $\pm$ 10% of teachers. This corresponds to a move of 625 spots in the rankings for math teachers and 684 spots for reading teachers. Particular teachers can move quite a bit in the rankings in some of the alternate ranking schemes. 22% of teachers move in the rankings $\pm$ the equivalent of 10% of teachers in the case

where 50% of the weight is put on the standard deviation in math. However, in the case where 25% of the weight is put on the standard deviation, only 2% of teachers move the equivalent of $\pm$ 10% of teachers.

4.2.1 Sensitivity Analyses

I perform a number of sensitivity checks for the analysis, which are reported in Table (6) for mathematics and Table (7) for reading. I discuss the results for mathematics in detail below, but the results for reading are similar. Overall, the results for the sensitivity checks are similar to the baseline results.

In the first sensitivity check, I examine the results using fifth grade teachers rather than sixth grade teachers.[23] A large number of school districts use value-added models to evaluate elementary school teachers, so understanding how rankings change when information on the value-added standard deviations is added is important as well. In row 2 of the tables, I report the correlation between the estimates of γ_j and σ_j^2, the Spearman rank correlations between a system were teachers are ranked only on the mean and a system where 50% of the weight is put on the estimate of σ_j, and the percentage of teachers that move $\pm$ 10% of teachers in the rankings under the alternate ranking system. The correlation between the estimates of γ_j and σ_j^2 is somewhat weaker in fifth grade than sixth grade, with a correlation of -.146 rather than -.328, and the rank correlation when

[23] I have also examined results for fourth grade teachers. In this case, only one prior year score is available as a controls, because testing started in grade three. However, the results are similar to fifth grade.

50% of the weight is put on the estimate of σ_j is .884 rather than .947.[24] The fraction moving more the 10% also increases to 39% in fifth grade compared to teachers in sixth grade. It appears that adding information on the value-added variances is more important in fifth grade than sixth grade, but this could also be explained by the greater imprecision in the measures of value-added means and variances in fifth grade.[25] With greater imprecision, the correlation between any two measures will tend to be lower than than two more precise measures, all else the same.

As discussed in Goldhaber et al. (2013), there is considerable disagreement about the conditioning variables that are needed for ignorability. It is common to include classroom level peer characteristics or school indicator variables in the regressions. In row 3 of Table (6), results for the estimates of γ_j and σ_j^2 when classroom level peer variables are included.[26] The correlation between the estimates of γ_j and σ_j^2 when the classroom level variables are included is -.244. The spearman rank correlation between the alternate ranking system and the ranking based only on the mean is .906, and the percent moving more than 10% is 34%. The correlation is slightly lower, and the percent moving 10% is slightly higher than the baseline case. This may be due to the additional noise in the estimates

[24]The rank correlation when 25% of the weight is put on σ_j (not reported in the table) is still a very high .982 in fifth grade compared .993 in sixth grade.

[25]Fifth grade teachers tend to have fewer student observations than sixth grade teachers, which affects the precision of the estimates. In my data, fifth grade teachers have on average 42.8 student observations, while sixth grade teachers have 109.7 student observations.

[26]The peer variables I include are: average prior year math and reading scores, proportion free and reduced-price lunch, and proportion limited English proficient. These coefficients are identified using within teacher variation in classroom composition.

created by trying to identify the coefficients on the classroom peer variables.

In row 4, I show results from when I estimate value-added variances using a linear functional form rather than an exponential functional form and while keeping the covariate set identical to the baseline specification.[27] I estimate σ_j^2 in an OLS regression of squared residuals from the regression to estimate (1) on $X_i - \bar{X}$ and teacher indicator variables. The correlation between the estimate of σ_j^2 and the estimate of γ_j is -.348. The rank correlation is .919, which is similar to the rank correlation from the baseline specification of .947.

In the final row, I report results from a specification with school dummy variables. Due to computation issues related to finding convergence in the non-linear least squares algorithm when school and teacher indicator variables were both included, I again change the functional form for the variance from an exponential function of the parameters to a linear function. I estimate σ_j^2 in an OLS regression of squared residuals from the regression to estimate (1), which also had school indicator variables included, on $X_i - \bar{X}$, school indicator variables, and teacher indicator variables.[28] In this case, the correlation between the estimates of γ_j and σ_j^2 is -.310, and the Spearman correlation drops slightly to .860 compared to .947 in the baseline specification. The percent that move $\pm$ 10% also increases to 35%.

[27] Note that the estimates of σ_j^2 are not guaranteed to be positive using this approach. However, in practice there are only a few instances where σ_j^2 is estimated to be negative for a teacher. In the case with the linear variance but no school fixed effects, only .2% teachers have negative estimates. In the specification, reported below, with school dummy variables and linear variance only .1% teachers have negative estimates.

[28] I used the user written felsdvreg package in Stata to estimate the coefficients on the teacher and school indicator variables. The coefficients are identified by teachers switching schools.

4.3 Do Classroom and Teacher Characteristics Explain Value-Added Variances? Exploiting Within Teacher Variation in Characteristics

Researchers, parents, and administrators may be interested in identifying attributes of classrooms and teachers that result in lower value-added variances.[29] In this section, I use within teacher variation in classroom and teacher characteristics to identify the impact of these characteristics on the value-added variances. Specifically, I examine the effects of classroom composition, teacher experience, and the dissimilarity of students in the classroom, all of which could conceivably affect the value-added variance produced by a teacher. Classroom composition could affect a teachers value-added variance, if for instance, it were easier to produce more equal learning gains for initially higher performing students than for initially lower performing students. The variances could also depend on the experience level of a teacher, for instance if teachers are able to learn how to better meet the needs of all students in a classroom. Finally, having a classroom of initially more dissimilar students could conceivably result in a large value-added variance, for instance if teachers are less able to cater to the needs of all students in a classroom in these conditions.

[29] Multiple studies have examined the attributes of classrooms and teachers that result in high value-added means. For instance, see Wiswall (2013), Harris & Sass (2011), and Goldhaber & Hansen (2013). Wiswall (2013) finds that teacher experience increases value-added means in mathematics, but finds little evidence that is does so for reading. Harris & Sass (2011) also find that value-added means increase with experience, and find little evidence that professional development training, undergraduate training, or college entrance exam scores raise value-added means. Goldhaber & Hansen (2013) find that only a small portion of the variation in value-added means can be explained by observable factors of teachers and classrooms.

In order to identify the effects of these variables, I use an approach similar to Wiswall (2013) and Goldhaber & Hansen (2013), who examine the effects of teacher observables on a teacher's value-added mean. I exploit naturally occuring variation in student composition and teacher experience for a teacher over time to identify the effects of the average incoming achievement levels of students, the fraction of classroom that is white, teacher experience, and the standard deviation of incoming achievement levels among students in a classroom, which I use as a measure of dissimilarity between students within a teacher's classroom. Specifically, I first form yearly value-added variances for teachers, then I perform a within teacher (teacher fixed effects) regression of these yearly value-added variances on yearly classroom characteristics, teacher experience, and year dummies.[30] Under the assumption that the naturally occuring variation in these characteristics is unrelated to unobserved differences in value-added variances over time, the estimates consistently estimate the effects of these variables.

Estimates are reported for mathematics and reading in Table 8. Standard er-

[30]The procedure to estimate the yearly value-added variances is nearly identical to the procedure to estimate value-added variances described in section 2, except instead of teacher indicator variables, I include teacher-year indicator variables. Specifically, in a first step I regress student test scores on $X_i - \bar{X}$ and teacher-year indicator variables. Then I form residuals from this initial regression and estimate yearly value-added variances for teachers using non-linear least squares of the squared residuals on $X_i - \bar{X}$ and teacher-year indicators. Teacher j then has a value-added variance estimate for each year they are in the data set. To examine the effects of classroom composition and experience on the value-added variances, I then do a within teacher (teacher fixed effect) regression of the value-added variance estimates on the average prior year achievement level of the teacher's students in a year, the proportion white, experience, experience squared, the standard deviation of the prior year achievement level of the teacher's students, and year dummies. There are 12,607 teacher-year observations for math teachers and 5,475 unique teachers that are included in the data set for multiple years. There are 13,540 teacher-year observations for reading teachers and 6,049 unique teachers that are included in the data set for multiple years.

rors, clustered at the teacher level, are reported in parenthesis. Overall, the variables do not explain much of the variation in the value-added variances. For mathematics, only the classroom dissimilarity measure, which is the standard deviation of the prior year test scores of the students, has a statistically significant effect on the value-added variance. The coefficient is .020, which is the expected sign, suggesting that going from having a classroom at the median in terms of the standard deviation of prior year scores (.684) to the 75th percentile (.804) increases the value-added variance for a teacher by only .0024. This would move a teacher with a value-added variance at the median to the 52nd percentile. For reading, experience, experience squared, and the standard deviation of the prior year scores are statistically significant predictors. Experience has the largest impact. The estimates suggest that going from 5 years of experience to 15 years of experience would actually *increase* the value-added variance by .024, which would move a teacher with a value-added variance at the median to the 62nd percentile. This suggest that more experienced reading teachers are actually worse at producing more equal achievement gains for their students. The standard deviation of prior year achievement has a coefficient of .035, which suggests that going from median of the standard deviation of prior year achievement (.789) to the 75th percentile (.919) increases the value-added variance by .0046, which would move a teacher with a value-added variance at the median to the 52nd percentile.

5 Summary and Conclusions

Researchers and administrators interested in teacher quality typically produce a single measure of teacher quality. If teachers are having heterogeneous impacts on their students, this measure reflects differences across teachers in the mean value-added they provide, but only examining the effect for the mean may offer an incomplete characterization of a teacher's quality. This paper offers an empirical strategy for identifying measures of value-added variances, and examines how rankings change when this information is added and whether classroom and teacher characteristics affect the value-added variances.

There are several important findings in this paper. I find evidence that there are modest to moderate differences across teachers in the size of the value-added variance, but the differences across teachers for σ_j^2 are smaller than differences across teachers for γ_j. Teacher rankings based on the mean and the variance are negatively correlated, with a correlation around -.25. As a result, teacher rankings that include value-added variances tend to be highly correlated with rankings that only include value-added means under some plausible ranking schemes. Typically the correlation is above .9. A positive conclusion from this paper is that rankings using measures of value-added means are fairly robust to adding information on the value-added variance. I additionally find that observable characteristics of teachers and their classrooms explain only a small amount of the variation in value-added variances.

This paper also shows that value-added variances can be calculated at fairly

low cost. Researchers already computing value-added means by regressing test scores on covariates and teacher indicator variables can estimate value-added variances using the two step approach used in the paper. These estimates could be useful for researchers who wish to study the factors that affect the variance in teacher value-added for instance. More research could be done on this topic. The methods and findings in this paper can serve as a starting point.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95–135.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). Who leaves? Teacher attrition and student achievement. National Bureau of Economic Research Working Paper 14022.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. National Bureau of Economic Research Working Paper 17699.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of

teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–79.

Condie, S., Lefgren, L., & Sims, D. (2014). Teacher heterogeneity, value-added and education policy. *Economics of Education Review*, *40*(0), 76 – 92.

Dee, T. S. (2004). Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics*, *86*(1), 195–210.

Goldhaber, D., & Chaplin, D. (2012). Assessing the Rothstein falsification test: Does it really show teacher value-added models are biased? *Center for Education Data & Research Working Paper*.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589 – 612.

Goldhaber, D., Walch, J., & Gabele, B. (2013). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy*, *1*(1), 28–39.

Guarino, C. M., Reckase, M. D., Stacy, B., & Wooldridge, J. M. (forthcoming). Evaluating specification tests in the context of value-added estimation. *Journal of Research on Educational Effectiveness*.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, *10*(1).

Hanushek, E. A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of human Resources*, (pp. 351–388).

Harris, D., Sass, T., & Semykina, A. (2010). Value-added models and the measurement of teacher productivity. CALDER Working Paper 54.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of public economics*, *95*(7), 798–812.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, (pp. 461–465).

Imbens, G. M., & Wooldridge, J. M. (2008). Recent developments in the econometrics of program evaluation. National Bureau of Economic Research Working Paper 14251.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, *87*(3), 706–710.

Imberman, S. A., & Lovenheim, M. F. (2013). Does the market value value-added? Evidence from housing prices after a public release of school and teacher value-added. National Bureau of Economic Research Working Paper 19157.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. research paper. met project. *Bill & Melinda Gates Foundation*.

Kane, T. J., & Staiger, D. O. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *Bill & Melinda Gates Foundation*.

Lockwood, J., & McCaffrey, D. F. (2009). Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy*, *4*(4), 439–467.

Loeb, S., Soland, J., & Fox, L. (2014). Is a good teacher a good teacher for all? Comparing value-added of teachers with their English learners and non-English learners. *Educational Evaluation and Policy Analysis*.

McCaffrey, D. F., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of educational and behavioral statistics*, *29*(1), 67–101.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263–283.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247–252.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rothstein, J. (2009). Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of educational and behavioral statistics*, *29*(1), 103–116.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3–F33.

Winters, M. A., Dixon, B. L., & Greene, J. P. (2012). Observed characteristics and teacher quality: Impacts of sample selection on a value added model. *Economics of Education Review*, *31*(1), 19–32.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61–78.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed.

Table 1: Student Level Summary Statistics

Variable	Mean	Std. Dev.
Original Sample		
Number of Student Obs	923,247	
Math Standardized Scale Score	0	1
Reading Standardized Scale Score	0	1
White	0.492	.5
Free and Reduced Price Lunch	0.486	0.5
Limited English Proficiency	0.18	0.384
Female	0.508	0.5
Sample After Restrictions		
Number of Student Obs	685967	
Math Standardized Scale Score	0.074	0.962
Reading Standardized Scale Score	.09	0.956
White	0.497	0.5
Free and Reduced Price Lunch	0.479	0.5
Limited English Proficiency	0.177	0.382
Female	0.512	0.5

Table 2: Teacher Level Summary Statistics

Variable	Mean	Std. Dev.
Math Teachers		
Number of Mathematics Teachers	5987	
Student Obs for Math Teachers	114.58	126.303
Student and Teacher Characteristics Aggregated to Teacher level		
Average Prior Year Math Score	-.203	.547
Fraction Free Reduced Price Lunch	0.527	0.257
Fraction Limited English Proficient	0.186	0.215
Fraction White	0.462	0.299
Teacher Experience	7.826	8.85
Reading Teachers		
Number of Reading Teachers	6606	
Student Obs for Reading Teachers	105.013	119.82
Student and Teacher Characteristics Aggregated to Teacher level		
Average Prior Year Reading Score	-0.337	0.611
Fraction Free Reduced Price Lunch	0.521	0.256
Fraction Limited English Proficient	0.181	0.22
Fraction White	0.471	0.299
Teacher Experience	7.711	8.832

Table 3: Standard Deviation and Correlations for γ_j and σ_j^2

Statistic	Mathematics	Reading
Std Dev $\hat{\gamma}_j$	0.207	.155
Std Dev $\hat{\sigma}_j^2$	0.086	.106
Correlation $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$	-.328	-.206
Number of Teachers	6249	6836

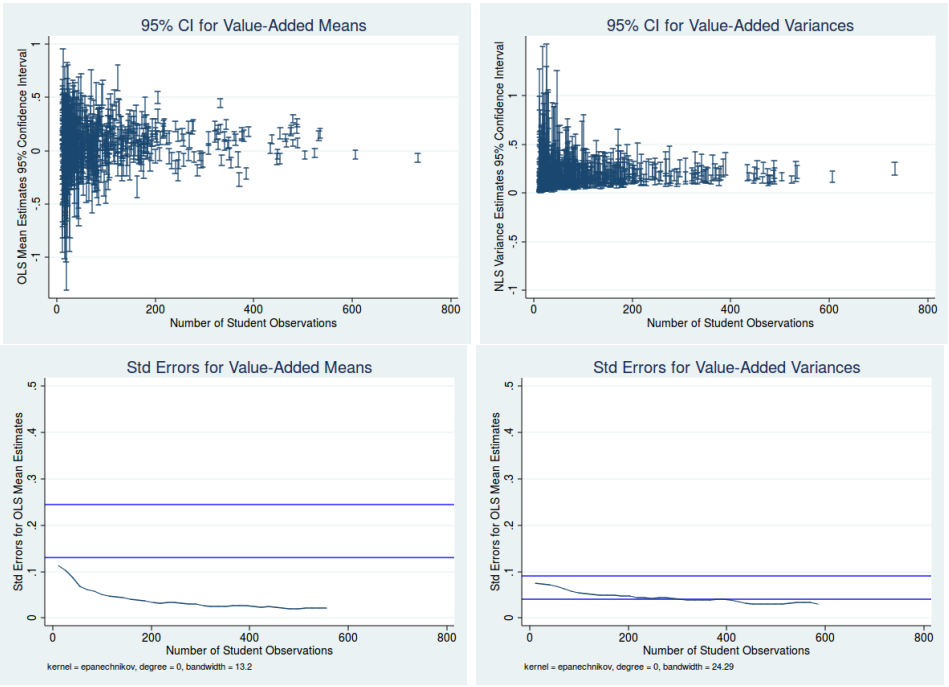
Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 4: Estimates and Standard Errors of γ_j and σ_j^2 for Select Teachers

γ_j				
Select Teachers	Mathematics		Reading	
10th Pctl	-.165	(.106)	-.084	(.094)
25th Pctl	-.061	(.081)	.006	(.061)
50th Pctl	.066	(.056)	.092	(.091)
75th Pctl	.194	(.066)	.187	(.066)
90th Pctl	.314	(.086)	.272	(.109)
σ_j^2				
Select Teachers	Mathematics		Reading	
10th Pctl	.096	(.035)	.171	(.049)
25th Pctl	.133	(.049)	.220	(.059)
50th Pctl	.181	(.043)	.270	(.073)
75th Pctl	.238	(.059)	.329	(.079)
90th Pctl	.301	(.072)	.405	(.122)
Observations	584		743	

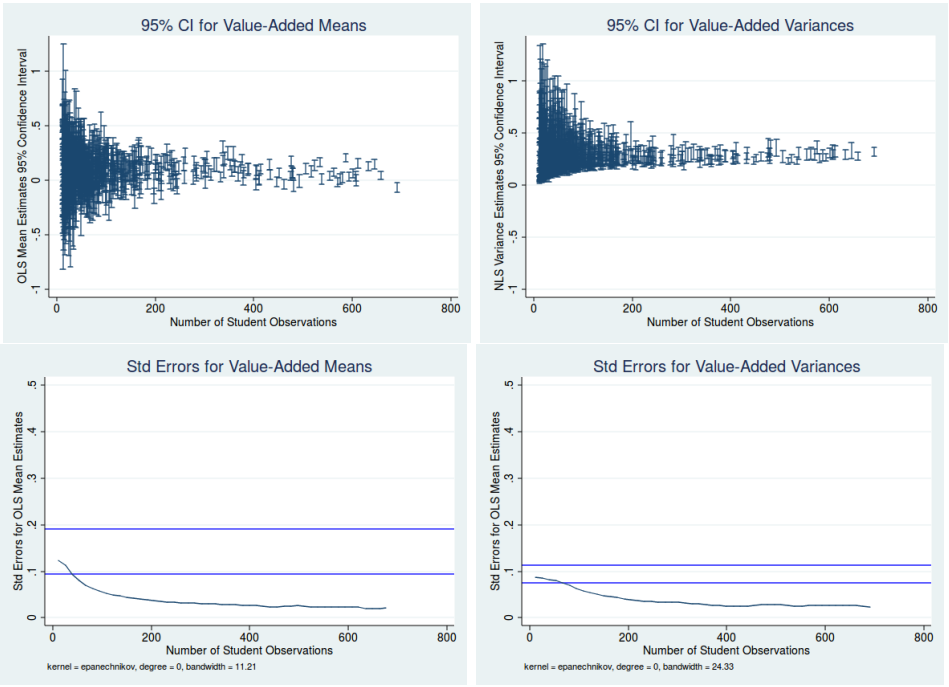
10th Pctl refers to a teacher at the 10th percentile. 25th Pctl refers to a teacher at the 25th percentile and so on. Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Figure 1: Plots of 95% CI and Standard Errors on the Number of Student Observations for Math Teachers



The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right. Average standard errors at each number of student observations, formed using a polynomial smoother, for the OLS estimates of γ_j are in the bottom left, and the average standard errors for the NLS estimates of σ_j^2 are in the bottom right. The blue lines represent the standard error necessary to statistically reject at the 5% level that a teacher at the 25th percentile is not above the 75th percentile, and that a teacher in the 10th percentile is not above the 90th.

Figure 2: Plots of 95% CI and Standard Errors on the Number of Student Observations for Reading Teachers



The OLS estimates of γ_j are in the top left. The NLS estimates of σ_j^2 are in the top right. Average standard errors at each number of student observations, formed using a polynomial smoother, for the OLS estimates of γ_j are in the bottom left, and the average standard errors for the NLS estimates of σ_j^2 are in the bottom right. The blue lines represent the standard error necessary to statistically reject at the 5% level that a teacher at the 25th percentile is not above the 75th percentile, and that a teacher in the 10th percentile is not above the 90th.

Figure 3: Scatterplot of Estimates of γ_j and σ_j^2 for Mathematics

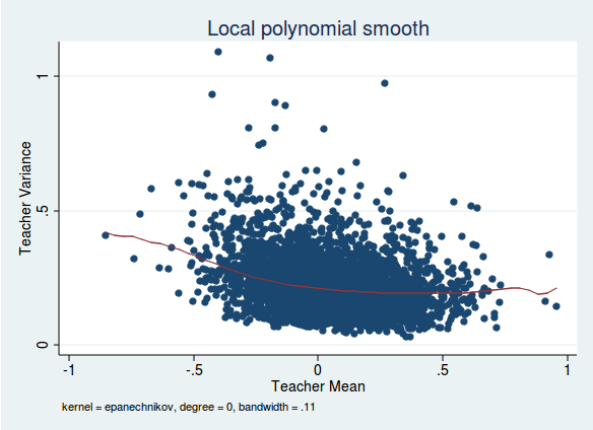


Figure 4: Scatterplot of Estimates of γ_j and σ_j^2 for Reading

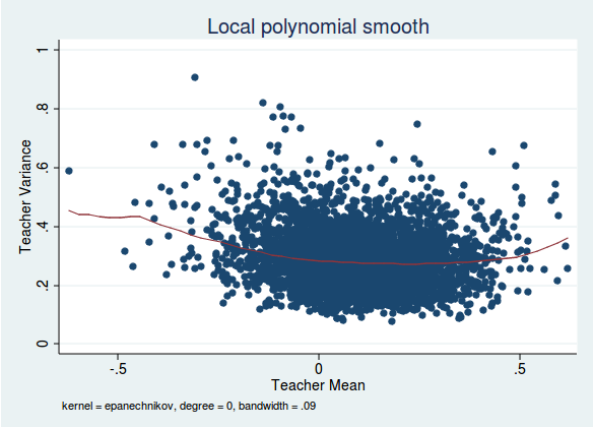


Table 5: Comparison of Ranking System Composed of $\hat{\gamma}_j$ and Alternative Ranking Systems Including σ_j

Subject	**25% on σ_j**	**33% on σ_j**	**50% on σ_j**
Spearman Rank Correlation with $\hat{\gamma}_j$			
Mathematics	.993	.985	.947
Reading	.982	.964	.881
Percentage Moving in Rankings 10% of Teachers			
Mathematics	2%	6%	22%
Reading	8%	16%	37%
Math Observations	6249		
Reading Observations	6836		

Controls included in estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 6: Sensitivity Checks for Mathematics Teachers

Specification	Corr $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$	Spearman 50% on $\hat{\sigma}_j$	Moving $\pm$ 10%
Baseline	-.328	.947 N=6249	22%
Grade 5 Teachers	-.146	.884 N=16726	39%
Classroom Level Variables	-.244	.906 N=6249	34%
Linear Variance	-.348	.919 N=6249	28%
School Dummy Variables with Linear Variance	-.310	.860 N=5987	35%

Controls included in baseline estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 7: Sensitivity Checks for Reading Teachers

Specification	Corr $\hat{\gamma}_j$ and $\hat{\sigma}_j^2$	Spearman 50% on $\hat{\sigma}_j$	Moving $\pm$ 10%
Baseline	-.206	.881 N=6836	37%
Grade 5 Teachers	-.128	.812 N=16827	50%
Classroom Level Variables	-.172	.858 N=6836	40%
Linear Variance	-.221	.844 N=6836	42%
School Dummy Variables with Linear Variance	-.166	.823 N=6608	39%

Controls included in baseline estimation of γ_j and σ_j^2 include a year dummy, cubic functions of lagged and twice lagged math and reading scores, indicators for minority status, free-and-reduced price lunch status, limited English proficiency status, gender, and teacher indicator variables.

Table 8: Effects of Changes in Student Characteristics and Experience on Value-Added Variances

VARIABLES	Mathematics	Reading
Average Prior Year Scores	0.008	-0.001
	(0.005)	(0.006)
Proportion White	-0.006	0.017
	(0.018)	(0.021)
Experience	0.002	0.004**
	(0.001)	(0.002)
Experience Squared	-2.64e-05	-7.98e-05*
	(3.70e-05)	(4.41e-05)
Std Dev Prior Year Scores	0.020**	0.035***
	(0.009)	(0.011)
Observations	12,607	13,540
Number of Teachers	5,475	6,049
R-squared	0.059	0.015

Estimates based on within teacher (teacher fixed effects) regression of yearly value-added variances on the yearly average prior year achievement level, proportion of students white, experience, experience squared, the yearly standard deviation of the student's prior year scores and year fixed effects. Robust standard errors in parentheses clustered at teacher level.

*** p<0.01, ** p<0.05, * p<0.1