The Education Policy Center
AT MICHIGAN STATE UNIVERSITY

WORKING PAPER #35

Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?

Brian Stacy
Mark Reckase
Jeffrey Wooldridge
Michigan State University

Cassandra Guarino
Indiana University

September 2013

Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?

Author Information

Brian Stacy
Mark Reckase
Jeffrey Wooldridge
Michigan State University

Cassandra Guarino
Indiana University

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 authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

This paper investigates how the precision and stability of a teacher's value-added estimate relates to the characteristics of the teacher's students. Using a large administrative data set and a variety of teacher value-added estimators, it finds that the stability over time of teacher value-added estimates can depend on the previous achievement level of a teacher's students. The differences are large in magnitude and statistically significant. The year-to-year stability level of teacher value-added estimates are typically 25% to more than 50% larger for teachers serving initially higher performing students compared to teachers with initially lower performing students. In addition, some differences are detected even when the number of student observations is artificially set to the same level and the data are pooled across two years to compute teacher value-added. Finally, the paper offers a policy simulation which demonstrates that teachers who face students with certain characteristics may be differentially likely to be the recipient of sanctions in a high stakes policy based on value-added estimates and more likely to see their estimates vary from year-to-year due to low stability.

Does the Precision and Stability of Value-Added Estimates of Teacher Performance Depend on the Types of Students They Serve?

Brian Stacy
Michigan State University

Cassandra Guarino
Indiana University

Mark Reckase
Michigan State University

Jeffrey Wooldridge
Michigan State University

September 17, 2013

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.

Abstract

This paper investigates how the precision and stability of a teacher's value-added estimate relates to the characteristics of the teacher's students. Using a large administrative data set and a variety of teacher value-added estimators, it finds that the stability over time of teacher value-added estimates can depend on the previous achievement level of a teacher's students. The differences are large in magnitude and statistically significant. The year-to-year stability level of teacher value-added estimates are typically 25% to more than 50% larger for teachers serving intially higher performing students compared to teachers with initially lower performing students. In addition, some differences are detected even when the number of student observations is artificially set to the same level and the data are pooled across two years to compute teacher value-added. Finally, the paper offers a policy simulation which demonstrates that teachers who face students with certain characteristics may be differentially likely to be the recipient of sanctions in a high stakes policy based on value-added estimates and more likely to see their estimates vary from year-to-year due to low stability.

1 Introduction

Teacher value-added estimates are increasingly being used in high stakes decisions. Many districts are implementing merit pay programs or moving toward making tenure decisions based at least partly on these measures. It is important to understand the chances that a teacher will be misclassified in a way that may lead to undeserved sanctions.

Misclassification rates depend on the precision of teacher effect estimates, which is related to a number of factors. The first is the number of students a teachers is paired with in the data. Teachers that can be matched with more student observations will tend to have more precise teacher effect estimates.

Another factor that can affect the precision of a teacher effect estimate is the error variance associated with students in the teacher's classroom. If the error variance is large, perhaps because the model poorly explains the variation in achievement or because the achievement measures themselves poorly estimate the true ability level of a student, then the precision of a teacher effect estimate will be low.

A question that seems to have lacked much attention is whether the precision varies by the characteristics of the students a teacher faces. Tracking of students into classrooms and sorting of students across schools means that different teachers may face classrooms that are quite different from one another. If it is found that teachers serving certain groups of students have less reliable estimates of value-added than other teachers serving other students, then all else the same,

the probability that a teacher is rated above or below a certain threshold will be larger for teachers serving these groups. High stakes policies that reward or penalize teachers above or below a certain threshold will then, again all else the same, impose sanctions or rewards on teachers serving these groups with a higher likelihood.

There are some reasons for suspecting that the characteristics of students in a classroom relates to the precision of teacher effect estimate. First, there could be a relationship between the characteristics of a classroom and the number of students linked to a teacher. This could be true because of a relationship between class size and student characteristics, because of poor data management for schools serving certain groups, or because of low experience levels for teachers serving certain groups, which limit the number of years that can be used to estimate the teacher's value-added.

Also, heteroskedastic student level error can imply that teachers paired with those students with large error variances may have less reliable teacher effect estimates. There is strong theoretical reason for supposing that the student level error is heteroskedastic. Item response theory suggests that because test items are typically targeted towards students in the center of the achievement distribution, achievement tends to be measured less precisely for students in the tails. The heteroskedasticity is also quite substantial, and suggests that teachers paired with particularly high achieving or low achieving students may have less reliable teacher effect estimates. In addition to heteroskedasticity caused by poor measurement, it is also conceivable that the error variance for true achievement is different for

different students.

In the remainder of the paper, we test for heteroskedasticity in the student level error term. In addition, year-to-year stability coefficients, which are very similar to year-to-year correlations, using a variety of commonly used value added estimators are computed for teachers serving different groups of students. Year to year stability coefficients for teachers with students in the bottom quartile, top quartile, and middle two quartiles in classroom level prior achievement are compared to one another.

A test of the homoskedasticity assumption easily rejects. Also, large and statistically significant differences in the stability coefficients among sub groups of teachers are detected, and the differences persist even after the number of student observations for all teachers is artificially created to be the same and when two years of data are used to compute value added. In many cases, the year-to-year stability coefficients are 25 to more than 50% larger in size for teachers serving initially higher achieving students compared to teachers serving lesser achieving and disadvantaged students.

This finding has several implications. For practitioners implementing high stakes accountability policies, teachers serving certain groups of students may be unfairly targeted for positive or negative sanctions simply because of the composition of their classroom and the variability this creates for their estimates. In this paper, we produce simulation evidence that bears this out. In addition, the heteroskedasticity makes it important for researchers and practitioners to make standard errors heteroskedasticity robust. Also, heteroskedasticity is a potential

source of bias for those using empirical Bayes value-added estimates, which assume homoskedasticity.

2 Previous Literature

A few studies have examined the stability and precision of teacher effect estimates. Aaronson et al. (2007) examined the stability of teacher effect estimates using three years of data from the Chicago public school system. They find that there is considerable inter-year movement of teachers into different quintiles of the estimated teacher quality distribution, suggesting that teacher effect estimates are somewhat unstable over time. They also find that teachers associated with smaller number of student observations are more likely to be found in the extremes of the estimated teacher quality distribution.

Koedel and Betts (2007) perform a similar analysis as Aaronson et al. (2007) using two years of data from the San Diego public school system and also find that there is considerable movement of teachers across quintiles.

McCaffrey et al. (2009) found year-to-year correlations in teacher value added to be .2 to .5 for elementary school teachers and .3 to .7 for middle school teachers using data from 5 county level school districts from the state of Florida from the years 2000-2005. They find that averaging teacher effect estimates over multiple years of data improves the year-to-year stability of the value-added measures.

This paper adds to the previous literature by specifically looking at whether the stability of teacher effect estimates is related to the characteristics of the students

assigned to the teacher.

3 Data

The data come from an administrative data set in large and diverse anonymous state. It consists of 2,985,208 student year observations from years 2001-2007 and grades 4-6. Student-teacher links are available for value-added estimation. Also, basic student information, such as demographic, socio-economic, and special education status, are available. Teacher information on experience is also available. The data include vertically scaled achievement scores in reading and math on a state criterion referenced test. The analysis will focus on value-added for mathematics teachers.

We imposed some restrictions on the data in order to accurately identify the parameters of interest. Students who 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. Students in schools with fewer than 20 students are dropped, and students in classrooms with fewer than 12 students are dropped. Districts with fewer than 1000 students are dropped to avoid the inclusion of charter schools in the analysis, which may employ a set of teachers who are somewhat different from those typically found in public schools. Characteristics of the final data set are reported in Table 1[1].

The analysis presented later is done separately for 4th grade and 6th grade.

[1]These restrictions eliminated about 31.2% of observations in 4th grade and 19% in 6th grade

This is done because the degree of tracking may be different in 6th grade from 4th grade, which may cause differences in the year-to-year stability of value-added estimates.

4 Model

The model of student achievement will be based on the education production function[2], which is laid out in , Todd and Wolpin (2003), Harris et al. (2011), and Guarino et al. (2012), among other places. Student achievement is a function of past achievement, current student and class inputs, along with a teacher effect.

$$\begin{aligned} A_{igt} &= \tau_t + \lambda_1 A_{ig-1t} + \lambda_2 A^{alt}_{ig-1t} + X_{igt}\gamma_1 \\ &\quad + \bar{X}_{igt}\gamma_2 + T_{igt}\beta + v_{igt} \end{aligned} \tag{1}$$

with

$$v_{igt} = c_i + \epsilon_{igt} + e_{igt} - \lambda_1 e_{ig-1t} - \lambda_2 e^{alt}_{ig-1t}$$

where A_{igt} is student i's test score in grade g and year t. τ_t is a year specific intercept. A^{alt}_{ig-1t} is the test score in the alternate subject, which in the analysis presented below is the reading score. X_{igt} is a vector of student level covari-

[2]The model shown includes a lagged score of the alternate subject, which isn't necessary under the assumptions typically made in deriving the regression model based on the education production function. However, including this variable is common in practice, so we chose to include it as well.

Table 1: Summary statistics

4th Grade				
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Math Scale Score	1543.377	240.699	581	2330
Reading Scale Score	1591.033	291.045	295	2638
Math Standardized Scale Score	0.103	0.947	-3.957	3.409
Reading Standardized Scale Score	0.105	0.928	-4.578	3.753
Black	0.208	0.406	0	1
Hispanic	0.224	0.417	0	1
Free and Reduced Price Lunch	0.486	0.5	0	1
Limited English Proficiency	0.173	0.378	0	1
Avg. Lag Math Score	1413.075	142.139	686.75	2066.737
Prop. FRL	0.496	0.28	0	1
Prop. LEP	0.17	0.213	0	1
Prop. Hispanic	0.218	0.245	0	1
Prop. Black	0.216	0.248	0	1
Students/Teacher	49.008	38.534	12	412
Teacher Years of Experience	8.902	8.887	0	47
# of Teachers	14,820			
# of Schools	1,768			
N		726,299		

6th Grade				
Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
Math Scale Score	1701.841	232.71	569	2492
Reading Scale Score	1704.809	294.454	539	2758
Math Standardized Scale Score	0.092	0.913	-4.163	3.354
Reading Standardized Scale Score	0.071	0.928	-4.049	3.526
Black	0.224	0.417	0	1
Hispanic	0.223	0.416	0	1
Free and Reduced Price Lunch	0.476	0.499	0	1
Limited English Proficiency	0.174	0.379	0	1
Avg. Lag Math Score	1647.707	131.958	866	2097
Prop. FRL	0.496	0.259	0	1
Prop. LEP	0.172	0.205	0	1
Prop. Hispanic	0.214	0.234	0	1
Prop. Black	0.24	0.245	0	1
Students/Teacher	145.378	165.685	12	1036
Teacher Years of Experience	9.571	9.362	0	40
# of Teachers	9 5,323			
# of Schools	796			
N		773,849		

ates including free and reduced price lunch and limited English proficiency status, gender, and race. $\bar{X}_{igt}$ consists of class level covariates, including lagged achievement scores, class size, and demographic composition. T_{igt} is a vector of teacher indicators. The teacher effects are represented in the β vector. c_i represents a student fixed effect. ϵ_{igt} represents an idiosyncratic error term affecting achievement. e_{igt} is measurement error in the test scores with e^{alt}_{igt} representing the measurement error in the alternate subject score.

4.1 Estimation Methods

Teacher effects were estimated using two commonly used value-added estimators.[3]

The first is a dynamic OLS estimator (DOLS)[4], which includes teacher indicators in an OLS regression based on equation (1). The estimator is refered to as dynamic because prior year achievement is controled for on the right hand side. The coefficients on the teacher indicator variables are interpreted as the teacher effects. We run our models using one year of data and again using two years of data. Because the effects of class average covariates are not properly identified in a teacher fixed effects regression with only one year of data, these variables are dropped from the DOLS regressions[5]. Additionally, when one year of data is used

[3]We have studied two more estimators based on a gain score equation. One estimator based on teacher fixed effects, and another based on empirical Bayes. The patterns for these two other estimators are similar to those reported for DOLS and EB Lag.

[4]This estimator was found to be the most robust of all the estimators evaluated in Guarino et al. (2012)

[5]We have tried a two step method that can identify the effect of class average covariates in a teacher fixed effects regression as a sensitivity check, and the results are similar. First, using the

to estimate value-added, the year specific intercepts are dropped.

The second is an empirical Bayes estimator (EB Lag) which treats teacher effects as random. The estimator follows closely the approach laid out in Kane and Staiger (2008). The parameters of the control variables are estimated in a first stage using OLS, then unshrunken teacher effect estimates are formed by averaging the residuals from the first stage among the students within a teacher's class. The shrinkage term is the ratio of the variance of persistent teacher effects to the sum of the variances of persistent teacher effects, idiosyncratic classroom shocks, and average of the individual student shocks[6]. Teacher effects are interpreted as the shrunken averaged residuals for each teacher.

5 Heteroskedastic Error

There is good reason to suspect that the error in the student achievement model is heteroskedastic. We will present some basic theory suggesting that measurement error in test scores is heteroskedastic. Also, we will offer some possible reasons

pooled data with multiple years, equation (1) is estimated using OLS with teacher fixed effects included. Then, a residual is formed.

$$\begin{aligned} w_{igt} &= A_{igt} - \hat{\tau}_t - \hat{\lambda}_1 A_{ig-1t} - \hat{\lambda}_2 A^{alt}_{ig-1t} - X_{igt}\hat{\gamma}_1 - \bar{X}_{igt}\hat{\gamma}_2 - \hat{f}(exper_{igt}) \\ &= T_{igt}\beta + \hat{v}_{igt} \end{aligned}$$

which is then used in a second stage regression to form teacher effects using a sample based on 1 year of data.

[6]It is common to treat the variance of the individual student shocks as uniform across the population of students. In an effort to evaluate commonly used estimators, we also computed the shrinkage term by using the same variance term for the student level shocks for all teachers. Under heteroskedasticity, this shrinkage term would not be the shrinkage term used by the BLUP.

why the error variance of actual achievement may be heteroskedastic.

5.1 Heteroskedastic Measurement Error

Item response theory is typically the foundation for estimating student achievement. A state achievement test is typically composed of 40-50 multiple choice questions, or items. Each student can either answer a question correctly or incorrectly, and the probability of answering any individual question is assumed to be a function of the item characteristics and the achievement level of the student. The typical model of a correct response to an item assumes (See Reckase (2009) for more details):

$$Prob(u_{ij} = 1|a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)G(a_i(\theta_j - b_i)$$

where u_{ij} represents an incorrect or correct response to item i by student j. a_i is a discrimination parameter, b_i is a difficulty parameter, and c_i is a guessing parameter for item i. θ_j is the achievement level of student j. Often, a logit functional form is assumed for $G(\cdot)$, although the probit functional form is also used. In the case of the logit form we have:

$$Prob(u_{ij} = 1|a_i, b_i, c_i, \theta_j) = c_i + (1 - c_i)\frac{e^{(a_i(\theta_j - b_i))}}{1 + e^{(a_i(\theta_j - b_i))}}$$

Parameters can then be estimated using maximum likelihood or alternatively using a Bayesian estimation approach. To illustrate why heteroskedasticity ex-

ists, we will focus on maximum likelihood estimation. Lord (1980), under the assumption that the answer to each test item by each respondent is independent conditional on θ, showed that the maximum likelihood estimate of θ has a variance of:

$$\sigma^2(\hat{\theta}|\theta) = \left(\sum_{i=1}^{n} (c_i a_i)^2 \frac{e^{(a_i(\theta_j - b_i))}}{(1+e^{(a_i(\theta_j - b_i))})^2} \right)^{-1}$$

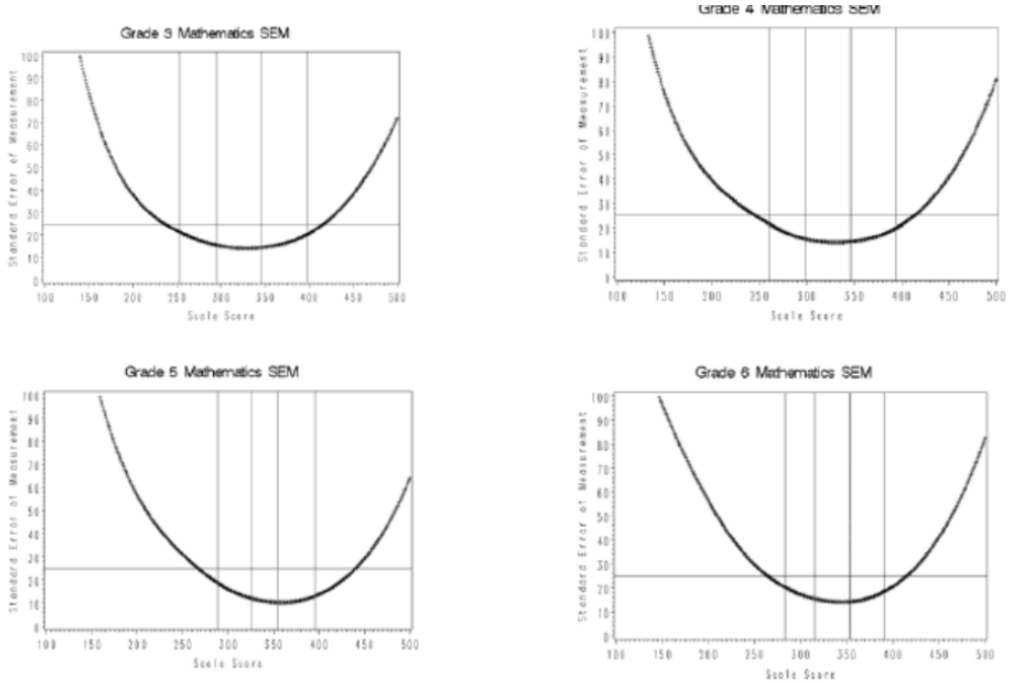
where n is the number of items. As can be seen, the variance would be minimized with respect to θ if $\theta_j - b_i = 0$ for all items, and as $\theta_j - b_i$ approaches $\pm\infty$, the variance grows large.

Since test items are often targeted toward students near the proficient level, in the sense that $\theta_j - b_i$ is near 0 for these students, students in the lower and upper tail often have noisy estimates of their ability. The intuition is that the test is aimed at distinguishing between students near the proficiency cutoff, and so the test offers little information for students near the top or bottom of the distribution.

Plots of the estimated standard deviation of the measurement error (SEM) on the student's test score level are shown below in Figure 1. The SEMs are on the vertical axis and the student's test score are on the horizontal axis for grades 3 through 6 for mathematics. The plots are from the 2006 State X Technical Report on Test Characteristics. The measurement error variance is a function of the test score level. Students in the extreme ranges of the test score distribution have a measurement error variance that is substantially larger than in the center.

Also, it may be the case that some groups of students may be less likely to answer all questions on the exam. As described in State X technical reports, test

Figure 1: Standard Error of Measure Plots for Mathematics Grades 3- 6



scores are computed for all students who answer at least 6 questions in each of 2 sessions. Students who answer only a fraction of the total number of questions on the exam will tend to have less precisely estimated test scores.

A prediction of the theory presented above is that the error variance will be related to all variables that predict current achievement. This is because the variance of the measurement error is directly related to the current achievement of the student, so all variables that influence the current achievement level of the student should also be related to the measurement error variance. In the test of heteroskedasticity that follow, this is the pattern that emerges.

5.2 Other Possible Causes of Heteroskedastic Student Level Error

In addition to heteroskedasticity generated from measurement, it is possible that other sources of heteroskedasticity exist. Little literature exists on this topic, but there are many potential causes, and we can only speculate on what they may be. Some groups of students, such as those with low prior year achievement, may have more variation in unobserved factors such as motivation, classroom disruptions, neighborhood effects, family effects, or learning disabilities. In addition, students who perform pooly on tests may tend to leave many questions blank or guess at answers, and thus their scores from test to test may be more variable.

In the following sections, we test for heteroskedasticity empirically, and look for possible differences in the error variance among groups. This serves to demon-

strate that the theoretical worries are justified and can motivate some predictions about how the precision of teacher effect estimates may depend on certain characteristics of the their students.

6 Testing for Heteroskedasticity

Under homoskedasticity:

$$E(v_{ig}^2|Z_{ig}) = \sigma_v^2$$

where Z_{ig} are the covariates in the regression model. We implemented a simple test of the homoskedasticity assumption examining whether squared residuals are related to student characteristics.

The first test simply grouped students into three groups: those with prior year test scores in the bottom 25%, the middle 50%, and the top 25%. We then calculated the average squared residuals for each group of students. We used the residuals from the DOLS regressions, which made use of teacher indicators. Results are included in Table 2. One thing to note is that the average squared residuals for the group of students in the bottom 25% in terms of prior year achievement are much larger than those for the group of students in the top 25%. The average squared residuals are around 45% larger for the bottom 25% compared with the top 25% for 4th grade and more than twice as large for 6th grade, even though under homoskedasticity, we would expect them to be similar. This is suggestive that more unexplained variation exists for the group of students in the bottom 25%

of the prior year achievement score.

Table 2: Average Squared Residuals for DOLS based on Subgroups of Prior Year Class Average Achievement

Grade	**Overall**	**Bottom 25%**	**Middle 50%**	**Top 25%**
4th Grade	18644.722	28091.514	13665.164	19352.092
N	709302	174780	356821	177701
6th Grade	16395.069	29825.119	11574.907	12670.438
N	723292	179894	357843	185555

Next we regressed the squared residuals on the covariates as well as on their squares and cubes. Results for grades 4 and 6 are reported in Table 3. We found that several of the variables including the lagged test scores, as well as the indicators for the student being African-American, free and reduced priced lunch, and limited English proficiency status were statistically significant predictors at the 10% level.

Since the precision and stability of a teacher's value-added measure depends in part on how much unexplained variation there is in the student's test scores, as will be explained below, this suggests that teachers paired with large numbers of disadvantaged or low achieving students may have less precise teacher value-added estimates. In the following sections, we will present evidence of this. Specifically, we will show that teachers of these types of students tend to have less stable teacher effect estimates over time.

Table 3: Tests for Heteroskedasticity

VARIABLES	Grade 4 DOLS Squared Residuals	Grade 6 DOLS Squared Residuals
Math Lag Score	-91.60***	-176.5***
	(5.357)	(15.28)
Math Lag Score Squared	0.00705*	0.00709
	(0.00396)	(0.00939)
Math Lag Score Cubed	1.05e-05***	1.57e-05***
	(9.45e-07)	(1.90e-06)
Reading Lag Score	-45.76***	-55.68***
	(2.772)	(5.663)
Reading Lag Score Squared	0.0161***	0.0173***
	(0.00195)	(0.00341)
Reading Lag Score Cubed	0.79e-07	-8.73e-07
	(4.43e-07)	(6.65e-07)
Black	293.8*	473.6**
	(177.3)	(205.2)
Hispanic	-265.9*	-272.0*
	(154.7)	(159.8)
FRL	540.6***	1,104***
	(114.9)	(120.3)
LEP	1,249***	711.8***
	(190.7)	(183.3)
Female	-1,436***	-2,609***
	(97.34)	(114.0)
Class Size	-54.90	41.06**
	(35.78)	(19.59)
Class Size Squared	1.124	-0.720***
	(0.724)	(0.266)
Class Size Cubed	-0.00426	0.00300***
	(0.00307)	(0.000923)
Teacher Experience	63.93	-210.4*
	(86.67)	(107.4)
Teacher Experience Squared	-0.758	23.46
	(15.24)	(18.33)
Teacher Experience Cubed	-0.228	-1.205
	(.975)	(1.168)
Constant	134,184***	262,361***
	(2,472)	(8,485)
Observations	709,302	723,292
R^2	0.050	0.079
Joint Test	886.6	862.4
p-value	0	0

Standard errors clustered at school level in parentheses

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

Joint Test refers to F test statistic that all coefficients equal to 0

In addition to the regressions presented in table 3, we performed the traditional Breusch-Pagan test, using fitted values, for heteroskedasticity separately for grade 4 and 6 and using the DOLS estimators. The test easily rejects the null hypothesis that the error is homoskedastic, with p-values for all grades and estimators less than .0001.

7 Evidence of Differences in Classroom Compositions

For there to be differences in the stability or the precision of teacher effect estimates due to student level heteroskedastic error, it is necessary for variation in classroom compositions to exist. For particular districts or states with little variation in classroom composition, it is unlikely that there will be large differences in the stability and precision of estimates due to heteroskedasticity. Also, there are some variables, such as gender, in which there may be a relationship with the error variance, but don't impact the precision and stability of teacher effect estimates, since there is little variation across classrooms with respect to the variables.

To show that there is variation in classroom composition with respect to certain variables across the state, we included a set of summary statistics in the middle panels of table 1 on classroom characteristics, which show that classrooms vary in their characteristics along a number of dimensions. The average past year math score of students in a class ranges from a score of 686.75 to 2066.737 for grade 4 and 866 to 2097 for grade 6. The interval between classrooms 2 stan-

dard deviations above the mean and 2 standard deviations below the mean is [1128.797,1697.353] for grade 4 and [1383.791,1911.623] for grade 6. Additionally, the proportion free and reduced priced lunch, limited English proficiency status, Hispanic, and African-American variables all range from 0 to 1.

8 Effects of Heteroskedastic Student Level Error on Precision of Teacher Value-Added Estimates

8.1 Simple Model of Heteroskedasticity

This model is designed to show, in the simplest case, how heteroskedasticity in the student level error can produce heteroskedasticity in teacher effect estimates. In the model there are two types of students and two teachers that students can be assigned to. The student types differ in the size of the student's error variance.

The achievement equation model is:

$$A_i = T_{0i}\beta_0 + T_{1i}\beta_1 + \epsilon_i$$

where A_i is the achievement level of student i, T_0 and T_1 are teacher assignment indicator variables for the two teachers, teacher 0 and teacher 1, β_0 and β_1 are teacher effects for teacher 0 and teacher 1, and ϵ_i is an error term assumed to be independent of teacher assignment.

Let the variable S_i indicate which of the two student types the student belongs

to and $v_0 < v_1$.

$$\begin{aligned} Var(\epsilon_i) &= v_0 \qquad\qquad if \quad S_i = 0 \\ Var(\epsilon_i) &= v_1 \qquad\qquad if \quad S_i = 1 \end{aligned}$$

$$v_0 < v_1$$

In this simple case, an OLS estimate of the teacher effect for teacher k produces:

$$\begin{aligned} \hat{\beta}_k - \beta_k &= \left(\sum_{i=1}^{N} T_{ki}^2 \right)^{-1} \left(\sum_{i=1}^{N} T_{ki}\epsilon_i \right) \\ &= \frac{\sum_{i=1}^{N} T_{ki}\epsilon_i}{N_k} \\ &= \bar{\epsilon}_k \end{aligned}$$

where $\bar{\epsilon}_k$ is the average error for the students that teacher k receives and N_k is the number of student observations for teacher k.

Let's suppose that each teacher has some students from S=0 and some from S=1. And also that teacher 0 tends to get more students from group 0, and teacher 1 tends to get more students from group 1.

We can use the Central Limit Theorem for inference. According to Greene (2008) (pg 1051, Lindeberg-Feller Central Limit Theorem with Unequal Variances) a central limit theorem result is possible as long as the random variables

are independent with finite means and finite positive variances. Also, the average variance, $\frac{1}{N_k}(\sum_{i=1}^{N_k} \sigma^2_{\epsilon_{ik}})$, where N_k is the number of students for teacher k, must not be dominated by any single term in the sum and this average variance must converge to a finite constant, $\bar{\sigma}^2_{\epsilon_k}$ as the number of students per teacher goes to infinity.

$$\bar{\sigma}^2_{\epsilon_k} = \lim_{N_k \to \infty} \frac{1}{N_k}(\sum_{i=1}^{N_k} \sigma^2_{\epsilon_{ik}})$$

Assume that all of those conditions hold. In that case,

$$\sqrt{N_k}(\hat{\beta}_k - \beta_k) \xrightarrow{d} Normal(0, \bar{\sigma}^2_{\epsilon_k})$$

and

$$Avar(\hat{\beta}_k) \approx \frac{\bar{\sigma}^2_{\epsilon_k}}{N_k}$$

In this simple example the average variance, $\bar{\sigma}^2_{\epsilon_k}$, for teacher 1 will tend to be larger than teacher 0, since they have more students from S=1. Therefore the asymptotic variance of the teacher effect estimate for teacher 1 will tend to be larger.

8.2 Including other Covariates in Achievement Model

Adding in covariates along with the teacher indicator variables complicates the result. In this case the achievement model is:

$$A_i = T_{0i}\beta_0 + T_{1i}\beta_1 + X_i\gamma + \epsilon_i$$

where X_i is a vector of covariates.

A well known result (see Wooldridge (2010)), is that the OLS estimate of the teacher fixed effect for teacher k is:

$$\begin{aligned} \hat{\beta}_k - \beta_k &= \bar{A}_k - \bar{X}_k\hat{\gamma}_{FE} - \beta_k \\ &= \bar{\epsilon}_k - \bar{X}_k(\hat{\gamma}_{FE} - \gamma) \end{aligned}$$

where $\bar{A}_k$ and $\bar{X}_k$ are the class averages of achievement and the covariates, and $\hat{\gamma}_{FE}$ is the fixed effects estimator of γ. It's straight forward to show that

$$Avar(\hat{\beta}_k) \approx \frac{\bar{\sigma}^2_{\epsilon_k}}{N_k} + \bar{X}_k Avar(\hat{\gamma}_{FE})\bar{X}'_k$$

$\frac{\bar{\sigma}^2_{\epsilon_k}}{N_k}$ will tend to be larger for teacher 1 than teacher 0. However, because of the additional terms in the $Avar(\hat{\beta}_k)$, it is not theoretically clear which teacher will have the less precise teacher effect estimate when the relationships between the covariates and the student types are unknown. Ultimately, whether teacher effect estimates are less precise for some teachers is an empirical question. The important point is that it is possible for some teachers to have less precise estimates due to student characteristics, so it is worthwhile to check whether that is the case.

9 Inter-year Stability of Teacher Effect Estimates by Class Characteristics

Imprecision of teacher effect estimates has some important implications, especially for policies that use teacher value-added estimates to make inferences about teacher quality.

The precision of a teacher effect estimate will affect how well that estimate can predict the true teacher effect. If the estimated teacher effect is quite noisy, then the estimate will tend to poorly predict the true teacher effect. This section explains how examining the year to year stability of value-added estimates can reveal important information about the measures for those intending to use them for high stakes policies. The year to year stability is calculated by regressing the value-added measure in year t on a value-added measure in a previous year. We calculate seperate stability coefficients for teachers with classrooms in the bottom 25%, middle 50%, and top 25% in terms of their students incoming average achievement. Those wishing to skip the technical details may move on to the next section.

Following McCaffrey et al. (2009), we can model a teacher effect estimate for teacher j in year t as:

$$\hat{\beta}_{jt} = \beta_j + \theta_{jt} + v_{jt}$$

where $\hat{\beta}_{jt}$ is the teacher effect estimate, β_j is the persistent component of the

teacher effect, θ_{jt} is a transitory teacher effect that may have to do with a special relationship a teacher has with a class or some temporary change in a teacher's ability to teach, and v_{jt} is an error term due to sampling variation. The variance of v_{jt} will be related to the number of student observations used to estimate a teacher effect and the error variance associated with the students in the particular teacher's class.

An important coefficient for predicting the persistent component of the teacher effect using an estimated teacher effect, which is essentially what a policy to deny tenure to teachers based on value added scores would be doing, is the stability coefficient, as termed by McCaffrey et al. (2009). The stability coefficient for teacher j is:

$$S_j = \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}$$

Note that the stability depends on the variance of the error term v_{jt}.

Assuming that the expectation of β_j conditional on $\hat{\beta}_{jt}$ is linear[7] and that β_j, θ_{jt}, and v_{jt} are uncorrelated [8], then:

[7]If the conditional expectation function isn't linear, then the algebra shown works for the linear projection, which is the minimum mean squared error predictor among linear functions of the estimated teacher effect

[8]This essentially implies that the teacher effect estimates are unbiased. There is some empirical support for this assumption at least for the DOLS and EB Lag estimators. Kane and Staiger (2008), Kane et al. (2013), and Chetty et al. (2011) both find that similar value-added estimators are relatively unbiased. If the estimates are biased, then we are effectively evaluating the stability of reduced form coefficients and not the causal effects of teachers on achievement. The estimators evaluated are commonly used in practice and conceivably will be used as the basis for high stakes policies, so it still may be of interest to know how they vary from year-to-year.

$$E(\beta_j|\hat{\beta_{jt}}) = \alpha + \frac{Cov(\hat{\beta_{jt}}, \beta_j)}{Var(\hat{\beta_{jt}})}\hat{\beta}_{jt} = \alpha + \frac{\sigma^2_\beta}{\sigma^2_\beta + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}\hat{\beta}_{jt} = \alpha + S_j\hat{\beta}_{jt}$$

and then also assuming that θ_{jt} and v_{jt} are mean zero, we get:

$$E(\beta_j|\hat{\beta_{jt}}) = (1 - S_j)\mu_{\beta_j} + S_j\hat{\beta_{jt}}$$

where μ_{β_j} is the mean of β_j. So the weight that $\hat{\beta_{jt}}$ receives in predicting β_j is related to the stability coefficient. If the stability coefficient is small, then the estimated teacher effect receives little weight in the conditional expectation function and is of little use in predicting β_j.

The stability coefficient can be estimated by an OLS regression of current year teacher value-added estimates on past year estimates of teacher value-added and a constant. This does impose the additional assumption that the variances of θ_{jt} and v_{jt} are constant over time and that the transitory teacher effect and error terms are uncorrelated over time. In that case the OLS estimates are estimating the population parameter:

$$\frac{Cov(\hat{\beta}_{jt-1}, \hat{\beta}_{jt})}{Var(\hat{\beta}_{jt-1})} = \frac{\sigma^2_{\beta_j}}{\sigma^2_{\beta_j} + \sigma^2_{\theta_{jt-1}} + \sigma^2_{v_{jt-1}}} = S_j$$

Since the variance of the teacher effect estimates tends to be constant over time, the regression coefficient is nearly identical to the inter-year correlation coefficient.

The stability coefficient will be estimated for different subgroups of teachers based on the characteristics of the students a teacher receives. Specifically, the stability will be computed for teachers that received classes in the bottom 25%, middle 50% and top 25% of classroom average prior test score in both years t and $t-1$. If the variance of v_{jt} differs across subgroups of teachers, then the stability and the degree to which the estimate predicts the true teacher effect will also differ.

Another ratio may be of interest. Following McCaffrey et al. (2009) once again, the reliability of a teacher effect estimate, denoted as R_{jt}, is:

$$R_{jt} = \frac{\sigma^2_\beta + \sigma^2_{\theta_{jt}}}{\sigma^2_\beta + \sigma^2_{\theta_{jt}} + \sigma^2_{v_{jt}}}$$

It may be of interest to know how much a teacher affected student learning in a given year. This may be the case in a merit pay system, for instance. In this case, we would be interested in the expected value of $\beta_j + \theta_{jt}$ conditional on the estimated teacher effect in year t. Using similar assumptions as before:

$$E(\beta_j + \theta_{jt}|\hat{\beta}_{jt}) = (1 - R_{jt})\mu_\beta + R_{jt}\hat{\beta}_{jt}$$

Under an additional assumption that variance of β_j and θ_{jt} do not vary across subgroups, then the stability of teacher value added estimates will be proportional to the reliability. This is simply because:

$$R_{jt} = \frac{\sigma^2_\beta + \sigma^2_{\theta_{jt}}}{\sigma^2_\beta} S_j$$

9.1 Brief Overview of the Analysis

Given that there may be differences in the degree of tracking or sorting in elementary and middle schools, the analysis is done separately by grade. Additionally, since it may be that teachers of certain types of classrooms are less experienced, and this may affect the year-to-year stability of the teacher's value-added estimate, the teacher's level of experience is controlled for in the regressions by creating seperate dummy variable for each possible year of experience and including each of those variables in the regressions.

The estimates for the different subgroups were computed by an OLS regression of the current year value-added estimate on the lagged teacher value-added estimate interacted with a subgroup indicator variable, a subgroup specific intercept, and an indicator for the teacher's level of experience. The regression equation is:

$$\begin{aligned}\hat{\beta}_{jt} &= \sum_{g=1}^{3} \alpha_g 1\{subgroup_{jt} = g\} + \sum_{g=1}^{3} \gamma_g \hat{\beta}_{jt-1} 1\{subgroup_{jt} = g\} \\ &\quad + \sum_{\tau=1}^{M} \zeta_\tau 1\{exper_{jt} = \tau\} + \phi_{jt}\end{aligned}$$

where $\hat{\beta}_{jt}$ is teacher j's value added estimate in year t, $subgroup_{jt}$ is a variable indicating the teacher's subgroup, and $exper_{jt}$ is the teacher's experience level. The γ_g parameters are the parameters of interest in the analysis. One way to think about them is as a group specific autoregressive coefficient for a teacher's value-

added score, and they are quite similar to group specific year-to-year correlations in value-added.

The advantage of the regression based approach over calculating year-to-year correlations is that it is much easier to calculate test statistics using conventional regression software. In the following sections, we will test whether the year-to-year stability of teacher value-added estimates for different subgroups are statistically different from one another.

The analysis is also repeated for each grade with the number of student observations artificially set to be equal. Since the precision of estimates for a teacher depends on both the number of student observations and the degree of variation in the student level error, it is of interest to identify the separate effects of these two sources of variability in teacher effect estimates. In order to make the number of student observations equal for all teachers, first all teachers with less than 12 student observations were dropped. Then for those teachers with more than 12 student observations, students are randomly dropped from the classroom until the the number of student observations is 12 for all teachers [9]. To give an example, suppose a teacher has 20 students in a class, then 8 of the students are randomly dropped, so that the teacher's value-added estimate is based on the scores of only 12 students.

First, results will be reported in which all teacher effects are estimated using only one year of data. Then, the analysis will be reported using two years of

[9] We have also done the analysis where the number of observations is set to 15 and 20, and the general patterns reported are the same.

data for each teacher. When two years of data are used to compute value-added the groupings into bottom 25%, middle 50%, and top 25% are based on the two year average of prior year test score within the teacher's classrooms. This then averages over the same sample of students used to compute the two year value-added measures.

In the case of the estimates based on two years of data, the teacher effect estimate for year t will be estimated using years t and $t-1$. The stabilities are computed by regressing the value-added estimate for year t on year $t-2$. This is done so that the years in which teacher effects are estimated do not overlap, which will avoid sampling variation or class level shocks affecting both estimates.

10 Results on the Stability of Teacher Effect Estimates by Subgroup

The inter-year stabilities for subgroups of teachers based on the average past year score of the students in the class are reported below[10]. We perform separate tests for whether the estimates for the middle 50% and top 25% statistically differ from the bottom 25%. Also, a joint test that the estimates for the middle 50% and top 25% are both statistically different from the bottom 25% is reported.

Although there is variation in what is statistically significant across grades

[10]We have also examined whether the inter-year stability differs when classrooms are grouped according to proportion free-and-reduced price lunch, proportion Hispanic, and proportion African-American. We found that teachers in classrooms with high proportions of minority and low-income students also have lower inter-year stabilities. Results are available upon request.

and estimators, a few patterns do emerge. The stability ratio tends to be highest for teachers facing classrooms in the middle 50% and top 25% in average lagged score compared to teachers in the bottom 25%. The stability ratio is typically 25 to over 50% larger for teachers with classrooms in the middle 50 and top 25%. This pattern is true even after the number of student observations is fixed at 12 and in some cases when 2 years of data are used to compute value-added.

10.1 DOLS Stabilities

Tables 4 shows the results for the DOLS estimator. Results for 4th grade and 6th grade are shown separately. The left panels show the DOLS teacher value-added estimates when the data is based on only one year of data. The right panel are based on estimates with two years of data. Within each panel, results labeled "Unrestricted Obs" are based on teacher value-added estimates that use all the available student observations in a year. Results labeled "12 Student Obs" are based on only 12 randomly chosen student observations in each year. For the two year results, the results reported under the "12 Student Obs" column are based on 12*2=24 student observations. Standard errors are clustered at the school level [11]. A "+" symbol indicates that the middle 50% (or top 25% as the case may be) coefficient is statistically different from the bottom 25% at the 5% level.

[11] We have also tried clustering at the teacher level, but the school level standard errors were more conservative, so we chose to report those.

Table 4: Estimates of Year to Year Stability for DOLS by Subgroups of Class Achievement

DOLS 4th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.359***	0.308***	0.551***	0.465***
	(0.0277)	(0.0266)	(0.0437)	(0.0449)
Middle 50%	0.483***+	0.392***+	0.646***	0.578***+
	(0.0181)	(0.0180)	(0.0325)	(0.0315)
Top 25%	0.555***+	0.471***+	0.730***+	0.660***+
	(0.0255)	(0.0246)	(0.0495)	(0.0485)
Observations	8,124	7,650	2,735	2,527
R^2	0.227	0.165	0.357	0.298
Joint Test	14.70	10.14	3.677	4.436
p-value	4.81e-07	4.27e-05	0.0257	0.0121

DOLS 6th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.534***	0.356***	0.812***	0.574***
	(0.0452)	(0.0476)	(0.0588)	(0.0756)
Middle 50%	0.619***	0.401***	0.717***	0.560***
	(0.0209)	(0.0247)	(0.0447)	(0.0485)
Top 25%	0.665***+	0.479***+	0.711***	0.575***
	(0.0263)	(0.0310)	(0.0403)	(0.0508)
Observations	4,290	3,772	1,506	1,359
R^2	0.481	0.288	0.642	0.445
Joint Test	3.684	3.233	1.193	0.0274
p-value	0.0256	0.0401	0.304	0.973

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited english proficiency, female, and year dummies. Standard errors clustered at school level in parentheses

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

+ Indicates value statistically different from Bottom 25% at 5% level

Joint Test: F-test statistic that Middle 50 % and Top 25 % coefficients different from Bottom 25%

10.1.1 4th Grade Results

In 4th grade, the stability for teachers with classes in the bottom 25% of prior year achievement is .359, and the stabilities for the middle 50% and top 25% are .483 and .555 respectively when the number of student observations is unrestricted. The coefficients for the middle 50% and top 25% statistically differ from the coefficient for the bottom 25% at the 5% level. The patterns are quite similar once the number of student observations is fixed at 12, although predictably the estimates are somewhat smaller, since in the unrestricted case each teacher's value-added estimate is based on at least 12 observations. The stability for the bottom 25% is .308 while the stabilities for middle and top are .392 and .471 respectively and are statistically different from the bottom. Additionally, in both the unrestricted and restricted to 12 observations cases, the joint test that both the middle 50% and top 25% coefficients differ from the bottom rejects comfortably at the 5% level.

For the cases in which two years of data are used, the stability is calculated using four years of data. The teacher effect estimate in year t, which uses data from year t and $t-1$, is regressed on the teacher effect estimate from year $t-2$, which uses years $t-2$ and $t-3$. For a teacher to be included in one of the quartile groupings, the teacher had to have a two-year average prior year achievement score in that quartile range for years t and $t-2$. This dramatically reduced the sample of teachers available to compare.

When two years of data are used to estimate teacher value-added in 4th grade the stability for teachers with classes in the bottom 25% increase to .551 and to .646 and .730 for the middle and top, respectively, in the unrestricted observations

case. The difference between the coefficients for the top and bottom is statistically significant at the 5% level. The point estimate for the middle 50% is larger than the bottom 25%, but the difference between the two is not statistically significant at the 5% level. The joint test that top or the middle coefficient differs from the bottom is significant at the 5% level. When the number of student observations per year is fixed at 12, the point estimates in the case of the middle and top are larger than the bottom, and both are statistically different from the bottom. The joint test that either the middle or top is different from the bottom also rejects.

10.1.2 6th Grade Results

The results for 6th grade are broadly similar to 4th grade using one year of data. With one year of data and unrestricted observations the stabilities tend to be higher than in 4th grade. This is likely due to 6th grade teachers having more student observations per year. In this case, the stabilities are .534, .619, and .665 for the bottom, middle, and top respectively. The tests for whether the top stabilities are different from the bottom rejects, while the test for the middle 50% does not. The joint test also rejects. When 12 student observations are used, the stabilities are .356, .401, and .479, respectively, for the bottom, middle, and top. Once again the test that the top and bottom differ and the joint test rejects, while the test that the middle differs from the bottom does not.

In the case of two years of data, none of the estimates statistically differ from one another in either the case of unrestricted observations or the case restricted to 12 student observations.

10.2 EB Lag Stabilities

The results for the empirical Bayes estimates can be found in Table 5 and are quite similar to those for the DOLS estimates. One difference between the empirical Bayes and DOLS specifications is that the regressions corresponding to the empirical Bayes estimates include classroom aggregates of the individual covariates, since this is often one of the justifications for using this approach over DOLS[12].

In the case of one year of data and 4th grade, the stability estimates are .361, .483, and .551 for the bottom 25%, middle 50%, and top 25%, respectively, in the unrestricted observations case. In the case where the number of student observations is set to 12, the stability estimates are .309, .391, and .461 respectively. In both cases, the middle 50% and top 25% estimates are statistically significantly different from the bottom 25%. The estimates are very similar to the DOLS case.

In the two year case in 4th grade, the pattern is again fairly similar to the DOLS results. When the number of observations is unrestricted, only the top and bottom 25% stabilities are statistically from one another. The p-value of the joint test is .0511, however. When the number of observations is restricted to 12, the estimates are .476, .584, and .657, respectively. The difference between the top 25% and bottom 25% coefficients is statistically at the 5% level. The joint test rejects at the 5% level as well.

In 6th grade with one year of data, the only statistically significant difference

[12]We have also included class aggregates in the DOLS regressions, and the results do not change much. Estimates of the class level aggregates were identified for DOLS using the two step approach described previously.

Table 5: Estimates of Year to Year Stability for EB Lag by Subgroups of Class Achievement

EB Lag 4th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.361***	0.309***	0.571***	0.476***
	(0.0278)	(0.0269)	(0.0445)	(0.0459)
Middle 50%	0.483***+	0.391***+	0.659***	0.584***
	(0.0183)	(0.0180)	(0.0341)	(0.0318)
Top 25%	0.551***+	0.461***+	0.733***+	0.657***+
	(0.0254)	(0.0246)	(0.0497)	(0.0491)
Observations	8,124	7,650	2,735	2,527
R^2	0.220	0.157	0.352	0.291
Joint Test	13.80	8.813	2.985	3.697
p-value	1.16e-06	0.000158	0.0511	0.0252

EB Lag 6th grade

	1 Year of Data		2 Years of Data	
	Unrestricted Obs	12 Student Obs	Unrestricted Obs	12 Student Obs
Bottom 25%	0.548***	0.354***	0.814***	0.583***
	(0.0433)	(0.0482)	(0.0497)	(0.0702)
Middle 50%	0.614***	0.385***	0.717***	0.551***
	(0.0199)	(0.0247)	(0.0432)	(0.0481)
Top 25%	0.650***+	0.457***	0.714***	0.561***
	(0.0267)	(0.0318)	(0.0405)	(0.0529)
Observations	4,290	3,772	1,506	1,359
R^2	0.437	0.224	0.610	0.387
Joint Test	2.492	2.402	1.558	0.0715
p-value	0.0835	0.0913	0.212	0.931

All regressions include lagged math and ELA test scores, indicators for Black, Hispanic, free and reduced price lunch, limited english proficiency, female, class averages of all preceding varianbles, class size, a quadratic function of experience, and year dummies. Standard errors clustered at school level in parentheses.
*** p<0.01, ** p<0.05, * p<0.1
+ Indicates value statistically different from Bottom 25% at 5% level
Joint Test: F-test statistic that Middle 50 % and Top 25 % coefficients different from Bottom 25%

at the 5% level is between the top 25% and bottom 25% in the unrestricted case, with point estimates of .650 for the top and .548 for the bottom. In the case of 2 years of data, no statistically significant differences are detected.

11 Sensitivity Checks

We performed a number of sensitivity checks. All of them support the conclusion that differences exist in the inter-year stabilities across sub-groups.

We performed the analysis using English language arts scores and found similar patterns as mathematics. The teachers assigned to students in the bottom 25% tended to have less stable value-added scores from year to year. One thing interesting to note is that English language arts value-added scores tended to be less stable from year-to-year overall compared to mathematics. This finding is consistent with the findings reported in the MET project reports.

Since it conceivable that teachers of students with low average prior achievement scores are inexperienced and inexperienced teachers also have lower inter-year stabilities, the analysis was repeated dropping all teachers with less than 5 years of experience. However, the teacher's experience was controlled for in the regression of the teacher's current value-added score on their prior value-added score specifically to account for this issue, and the patterns described above were very similar to those seen in this sensitivity check as expected.

As an additional sensitivity check, we repeated the analysis with school dummies. We were still able to detect statistically significant differences in inter-year

stabilities across sub-groups.

We tried estimating the empirical Bayes estimates using an alternate estimator. In the alternate estimator, we estimated the model parameters using a mixed effects estimator that treated the teacher effects as random. These results were very similar to the empirical Bayes approach outlined above that was based on the approach taken in Kane and Staiger (2008).

Also, we used twice lagged reading and math scores as instruments for the once lagged reading and math scores to help account for measurement error in these variables as another sensitivity check. Again, statistically significant differences were found in the stabilities across sub-groups.

Finally, we performed the analysis separately for the six largest school districts in the state. The general patterns held. In a majority of the cases, the stability coefficient was estimated to be the smallest in the case of the bottom 25%. In no case was the stability coefficient of the middle 50% or top 25% statistically significantly smaller than the bottom 25%. In some districts, the teachers with classrooms in the middle 50% had the largest year-to-year stability, while in others the top 25% had the largest year-to-year stability. In one case the year to year stability of the bottom 25% was the largest, but it wasn't statistically significantly so. The estimates were quite noisy when the sample was separated in this way, so it is not clear whether this reflected real differences across districts or not. It seems possible that in different context the group of teachers that has the largest year-to-year stability could differ. However, our main takeaway is that some groups of teachers have less stable value added estimates from year-to-year.

Tables for all of these sensitivity checks are available upon request.

12 High Stakes Policy Simulation

There is an increasing push to use value-added estimates for high stakes decisions such as tenure or merit pay bonuses. Since the precision and stability of a teacher's value-added estimate is related to the makeup of the teacher's class, it may be the case that the teachers serving certain groups of students may be more likely receive a sanction or bonus.

In order to examine this, we produced a simulation in which high stakes decisions are made based upon value-added scores, and teachers differ in the stability of their value-added estimates. We base the stability level of the measure of value-added on the results we found in the previous sections. Each teacher is ranked and flagged if they are in the bottom or top 10% according to their teacher value-added score. We then calculate the proportion of teachers associated with each stability level that are labeled as either in the bottom or top 10%.

The simulation consists of 300 teachers and 3 stability levels. 100 teachers are assigned to each stability level. The true teacher effects are normally distributed and have a mean of 0 and a variance of 1. The "estimated" teacher effects have estimation error added that is normally distributed with mean 0, and the variance depends on the stability level of the teacher.

Two sets of stability levels were chosen. The first corresponds to the DOLS estimates in 4th grade with 12 student observations and one year of data, with

stabilities of .308, .392, and .471. The second corresponds to the DOLS estimates in 4th grade with 12 student observations and 2 years of data, with stability levels of .465, .578, and .660.

We calculate the average proportion of teachers associated with each stability level over the 5000 reps. Results are included in Table 6. The results from the simulation using the DOLS estimates in 4th grade with 12 student observations and one year of data can be found in the upper panel. For teachers associated with the stability of .308, which was the stability associated with teachers of classrooms in the bottom 25% in the analysis above, the proportion found in the bottom or top 10% was .249. When the stability level was .392 the proportion dropped to .195, and when the stability went to .471, the proportion fell to .156. This last drop was nearly a 10 percentage point change from the lowest stability. The results using two years of data show a similar pattern and can be found in the bottom panel. Teachers associated with the lowest stability have a proportion of .243. Teachers associated with stabilities of .578 and .660, which were associated with students in the middle 50% and top 25%, respectively, were found in the bottom or top 10% of the estimated teacher quality distribution at a proportion of .193 and .164 respectively. This represents an almost 8 percentage point drop for the latter.

The simulation results indicate that the differences in stability levels found in this analysis can have a large impact on the likelihood that a teacher finds his or herself in the top or bottom of the estimated teacher quality distribution.

Table 6: High Stakes Policy Simulation

Simulation 1: DOLS Stability, 4th Grade, 12 Student Obs, 1 year of Data

Stability	Error Variance of VAM Estimate	Proportion Found in Bottom or Top 10%
.308	2.247	.249
.392	1.551	.195
.471	1.123	.156

Simulation 2: DOLS Stability, 4th Grade, 12 Student Obs, 2 years of Data

Stability	Error Variance of VAM Estimate	Proportion Found in Bottom or Top 10%
.465	1.151	.243
.578	.730	.193
.660	.515	.164

Simulations results are based on 5000 Monte Carlo repetitions. There are 100 teachers per type. True teacher effects are distributed Normal(0,1). Error in the value-added measures is normally distributed with mean 0 and a variance listed in the "Error Variance of VAM Estimate" column.

13 Conclusion

This paper provides evidence that the variability and stability of teacher effect estimates depends on the characteristics of a teacher's class. Policies to deny tenure to teachers and policies designed to reward teacher performance in a given year, which are based on teacher value-added estimates, may differentially impact teachers with certain types of students.

The relationship between the stability of estimates and the classroom characteristics of students extends beyond the number of student observations. There is a strong theoretical reason for suspecting that a student's error term is heteroskedastic and statistical tests bear this out. As a consequence of this and student tracking and sorting into schools, teachers will serve different groups of students and have differences in the precision of their teacher effect estimates as a result. The differences in the stability ratios are large in magnitude and statistically significant even after fixing the number of student observations to a constant.

Also, some evidence is presented that the relationships remain even as more observations are added. When two years of data are used, there still exist statistically significant and large differences for different subgroups of teachers.

The heteroskedasticity is likely due in part to heteroskedastic measurement error variance. Assuming the item response model is correct, heteroskedastic measurement error is a direct result of the maximum likelihood estimation procedure which produces estimates of the achievement level of each student. The patterns that teachers of students with lagged achievement scores in the middle of

the achievement distribution tend to have the highest inter-year stabilities is consistent with heteroskedasticity caused by the measurement error, although teachers with students in the top 25% also tend to have more stable estimates. One reason the top and bottom may be different is that there may be greater potential for guessing or item non-response for students at the bottom of the distribution. It may be possible to reduce the heteroskedasticity by improving measurement. Future work will hopefully explore how much of the heteroskedasticity is attributable to measurement.

Heteroskedastic student level error also has other implications for researchers and policymakers. Empirical Bayes estimators are commonly computed assuming homoskedastic student level error. This assumption does not seem to be true, and since there are large differences in stability ratios that appear to be driven by heteroskedasticity, the violation of this assumption may impact the teacher rankings that are created using the empirical Bayes estimators. Allowing heteroskedasticity in the student level error should be done if possible.

Additionally, it is quite common for standard errors and the corresponding confidence intervals to be based on a homoskedasticity assumptions[13]. It is important that the confidence intervals accurately reflect imprecision caused by all sources of variability, not just the number of student observations, so standard errors should at least be made heteroskedasticity robust. This is particularly important since the teacher value-added estimates are being made publicly available

[13]Ballou et al. (2004) assume homoskedasticity in computing standard errors, as does the value-added estimator employed by the NYC school district

in some school districts.

It is important to understand the limitations of any measure of performance. The analysis presented here does suggest that for all subgroups value-added measures do have positive inter-year stabilities, so information can be gathered for all subgroups of teachers. However, teachers of certain groups of students will tend to have less precise and less stable teacher value-added estimates. As a result of this, it is the opinion of the authors that care should be used in evaluating teachers using value-added estimators and value-added estimates should not be used as the sole basis of any high stakes policy involving teachers.

14 Work Cited

References

Aaronson, Daniel, Lisa Barrow, and William Sander, "Teachers and student achievement in the Chicago public high schools," *Journal of Labor Economics*, 2007, *25* (1), 95–135.

Ballou, Dale, William Sanders, and Paul Wright, "Controlling for student background in value-added assessment of teachers," *Journal of Educational and Behavioral Statistics*, 2004, *29* (1), 37–65.

Center, Value-Added Research, "NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model 2010," Technical Report 2010.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, "The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood," Technical Report, National Bureau of Economic Research 2011.

Greene, William H, *Econometric Analysis*, Pearson, 2008.

Guarino, Cassandra, Mark D Reckase, and Jeffrey M Wooldridge, "Can value-added measures of teacher performance be trusted?," Technical Report, Discussion Paper series, Forschungsinstitut zur Zukunft der Arbeit 2012.

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

Harris, Douglas, Tim Sass, and Anastasia Semykina, "Value-added models and the measurement of teacher productivity," 2011.

Jacob, Brian A and Lars Lefgren, "Can principals identify effective teachers? Evidence on subjective performance evaluation in education," *Journal of Labor Economics*, 2008, *26* (1), 101–136.

Kane, Thomas J and Douglas O Staiger, "The promise and pitfalls of using imprecise school accountability measures," *The Journal of Economic Perspectives*, 2002, *16* (4), 91–114.

_ and _ , "Estimating teacher impacts on student achievement: An experimental evaluation," Technical Report, National Bureau of Economic Research 2008.

_ , Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment. Research Paper. MET Project.," *Bill & Melinda Gates Foundation*, 2013.

Koedel, Cory and Julian Betts, "Re-Examining the Role of Teacher Quality In the Educational Production Function," Technical Report, Department of Economics, University of Missouri 2007.

Lord, Frederic M, *Applications of Item Response to Theory to Practical Testing Problems*, Lawrence Erlbaum, 1980.

McCaffrey, Daniel F, JR Lockwood, Daniel Koretz, Thomas A Louis, and Laura Hamilton, "Models for value-added modeling of teacher effects," *Journal of educational and behavioral statistics*, 2004, *29* (1), 67–101.

_ , Tim R Sass, JR Lockwood, and Kata Mihaly, "The intertemporal variability of teacher effect estimates," *Education*, 2009, *4* (4), 572–606.

Morris, Carl N, "Parametric empirical Bayes inference: theory and applications," *Journal of the American Statistical Association*, 1983, *78* (381), 47–55.

Reckase, Mark, *Multidimensional Item Response Theory*, Springer, 2009.

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

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2010.