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

Brian Stacy Cassandra Guarino Mark Reckase
Jeffrey Wooldridge

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 based on one year of data can depend on the previous achievement level, racial characteristics, and saucy-economic status of a teacher's students. The differences are large in magnitude and statistically significant. In some cases the year to year correlation of teacher value-added estimates is twice as large for teachers serving certain groups compared to other teachers serving other groups. In addition, some differences are detected even when the number of student observations are artificially set to the same level and the data is pooled across two years to compute teacher value-added. This implies that teachers who face students with certain characteristics may be differentially likely to be the recipient of negative or positive 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 is poorly explaining 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 will test for heteroskedasticity in the student level error term. In addition, inter-year correlations based upon one year of achievement data using a variety of commonly used value added estimators will be computed for teachers serving different groups of students. Inter-year correlations for teachers with students in the bottom quartile, top quartile, and middle two quartiles in classroom level prior achievement, race, and free and reduced price lunch status will be compared to one another.

A test of the homoskedasticity assumption will easily reject. Also, large and statistically significant differences in the correlations 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 correlations are roughly double in size for teachers serving higher achieving or socially advantaged students compared to teachers serving lesser achieving and disadvantaged students.

This has 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 addition, the heteroskedasticity makes it important for researchers and practitioners to make standard errors heteroskedasticity robust. Also, the heteroskedasticity is a potential source of bias for those using empirical Bayes value-added estimates, that assume homoskedasticity, as right hand side variables in an effort to circumvent attenuation bias caused by measurement error in value-added estimates.

2 Previous Literature

A few studies have examined the stability and precision of teacher effect estimates. Aaronson, Barrow, and Sander (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 inter-year correlations 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 received by the teacher.

3 Data

The data come from an administrative data set in large and diverse anonymous state. It consists of 2,372,528 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, sauce-economic, and special education status, are available. Teacher information on experience is also available. The data includes 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 restrictions on the data in a number of ways. First, the data set is too large to estimate teacher value-added for every teacher included in the data. This led to choice on how to drop observations so that the analysis is computationally feasible. One possibility was to choose certain school districts within the state to perform the analysis. However, in order to maintain the large diversity of the state in the sample, we avoided doing the analysis on only one or a few of the districts and chose a cluster sampling scheme[1]. Sampling was done at the school level, so that we kept all observa-

[1]The cluster sampling was done using the user written sample2 program in Stata, while

tions contained within the schools selected in the data set used for analysis. Summary statistics are available in the appendix to show that the cluster sampling maintained the balance found within the entire state (Tables 1 and 2).

We imposed some additional restrictions in order to accurately identify the parameters of interest. 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. Students in schools with less than 20 students are dropped, and students in classrooms with less 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 that are somewhat different from those typically found in public schools. Characteristics of the final data set are reported in the appendix (Table 2).

The analysis presented later will be done separately for 4th grade and 6th grade. This is done because tracking may be much more common in 6th grade than it is in 4th grade, which may cause differences in the inter-year correlations.

4 Model

The model of student achievement will be based on the education production function [2], which is laid out in Hanushek (1979) Todd and Wolpin (2003), Harris, Sass, and Semykina (2010) and Guarino, Reckase, and Wooldridge (2011), among other places. Student achievement is a function of past achievement, current student and class inputs, along with a teacher and

clustering by school id

[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.

school effect.

$$\begin{aligned} A_{ig} &= \tau_g + \lambda_1 A_{ig-1} + \lambda_2 A^{alt}_{ig-1} + X_{ig}\gamma_1 + \bar{X}_{ig}\gamma_2 \\ &\quad + f(exper_{ig}) + T_{ig}\beta + S_{ig}\xi + v_{ig} \end{aligned} \tag{1}$$

with

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

where A_{ig} is student i's test score in grade g. A^{alt}_{ig-1} is the test score in the alternate subject, which in the analysis presented below is the reading score. X_{ig} is a vector of student level covariates including free and reduced price lunch and limited English proficiency status, and race. $\bar{X}_{ig}$ consists of class level covariates, including lagged achievement scores, class size, and demographic composition. $f(exper_{ig})$ is a quadratic function of teacher experience. T_{ig} is a vector of teacher indicators. S_{ig} is a vector of school indicators. c_i represents a student fixed effect. ϵ_{ig} represents an idiosyncratic error term affecting achievement. e_{ig} is measurement error in the test scores with e^{alt}_{ig} representing the measurement error in the alternate subject score.

The teacher effects are represented in the β vector. Since a later part of the analysis focuses on year to year correlations between teacher effect estimates, experience is netted from the teacher effect. The teacher effects can be thought of as innate teaching ability, or teaching ability unchanging over time.

4.1 Estimation Methods

Teacher effects were estimated using four commonly used value-added estimators.

The first is a dynamic OLS estimator (DOLS), which includes teacher

indicators in an OLS regression based on equation (1). Teacher effects are interpreted as the coefficients on the teacher indicator variables.

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. [3] Teacher effects are interpreted as the shrunken averaged residuals for each teacher.

The next set of estimators make an assumption on equation (1) that there is no decay of past inputs, so the model reduces to a gain score model. The model is then:

$$A_{ig} - A_{ig-1} = \tau_g + X_{ig}\gamma_1 + \bar{X}_{ig}\gamma_2 + f(exper_{ig}) + T_{ig}\beta + S_{ig}\xi + v_{ig}$$

with X_{ig} no longer containing any lagged test scores. All other covariates are identical to those included in model (1).

First, a pooled OLS estimator (POLS) is used which includes teacher indicators to capture teacher effects. Alternately, an empirical Bayes estimator (EB Gain) is used. OLS is used in a first stage to estimate the parameters of the other covariates. Then, the averaged residuals for each teacher are shrunken using the same shrinkage formula as before.

[3] 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. Also, it's useful to note that the shrinkage term is related to the stability coefficient, which will be defined later.

5 Heteroskedastic Error

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

A plot of the estimated standard deviation of the measurement error on the student's test score level is available in the appendix. The figure is for 7th grade reading, but is representative of the relationship between the measurement error variance and ability level for all subjects and grades. It shows that the measurement error variance is a function of the test score level, and that for student's in the extreme ranges of the distribution, that the measurement error variance is substantial.

A prediction of the theory presented above is that the heteroskedasticity will be with respect 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's possible that other sources of heteroskedasticity exists. It's entirely possible that some groups have more variation in unobserved factors, such as motivation, neighborhood effects, family effects, or learning disabilities. In the following sections, we test for heteroskedasticity empirically, and look for possible differences in the error variance among groups. This serves to demonstrate 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 teacher's 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 by regressing squared residuals on a cubic polynomial of student and class covariates.

Results for 6th grade are reported in the appendix, but results for 4th grade are quite similar (Table 3). The regressions for each variable were done separately to help form clear predictions on the precision of teacher value-added estimates for teacher's with large numbers of certain types of students. As an example for clarity, we regressed the squared residuals on a

cubic polynomial of the math lagged score for instance, then we separately regressed the squared residuals on an indicator for whether the student was African-American, etc. We used the residuals from the DOLS and POLS regressions, which made use of teacher indicators.

Two general patterns are evident in the table. Students with lagged scores near the bottom of the distribution tend to have larger error variances than students in the middle and top. The coefficients for the terms of the polynomial for the math lagged score are such that variance is decreasing over the range of test score values. The coefficients are also statistically significant at the 1% level. Also, African-American, Hispanic, and free and reduced priced lunch status students tend to have larger error variance than other students. These coefficients are also statistically significant at the 1% level. In addition, we repeated the analysis using limited English proficiency status, lagged reading scores, and class average lagged scores and found similarly that economically and socially disadvantaged as well as those who have or are associated with lower levels of achievement tend to have higher error variances. These variables were statistically significant predictors of squared residuals.

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.

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 and POLS estimators. The test easily rejects the null hypothesis that the error is homoskedastic, with p-values for all grades and estimators less than .0001. Also, to check the prediction from item response theory that the heteroskedasticity is with respect to all the variables that predict achievement, we performed a multivariate

regression using all the covariates together. We found that several of the variables including the indicator for the student being African-American, the lagged scores, free and reduced priced lunch, and limited English proficiency status were statistically significant predictors at the 5% level, which was consistent with the theoretical prediction that all variables that predict student achievement will also be related to the error variance.

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's necessary for variation in classroom compositions to exist. For particular districts or states with little variation in classroom composition, it's 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 bottom panel of table 2 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 569 to 2442, and the interval between classrooms 2 standard deviations above the mean and 2 standard deviations below the mean is [1415.311,1898.539]. For proportion free and reduced priced lunch, limited English proficiency status, Hispanic, and African-American, the variables all range from 0 to 1. The intervals for values between two standard deviations below and above the

mean are [0,1],[0,.581],[0,.704], and [0,.705] for proportion free and reduced priced lunch, limited English proficiency, Hispanic, and African-American respectively.

8 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.

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:

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

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[4] and that β_j, θ_{jt}, and v_{jt} are uncorrelated [5], then:

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

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

[4] 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

[5] 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) and Chetty, Rockoff, and Friedman (2012) both find that the DOLS and EB Lag 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.

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[6]. Specifically, the stability will be computed for teachers that received classes in the bottom 25%, middle 50% and top 25% for each of the variables: classroom average prior test score, proportion receiving free or reduced price lunch, proportion Hispanic, and proportion African-American 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:

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

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:

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

[6]The estimates for the different subgroups were computed by interacting the lagged teacher effect estimate with a subgroup indicator variable and allowing different intercepts for each group in an OLS regression.

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:

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

8.1 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.

In addition, the analysis is 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. 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 from using two years of data for each teacher.

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.

9 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, proportion free and reduced price lunch, proportion Hispanic, and proportion African-American are reported in the appendix. The stabilities for the bottom 25% for each of these variables are reported along with the difference between the stability for those in the middle 50% and those in the bottom 25% and the difference between stabilities for those in the top 25% and those in the bottom 25%. Also, a joint test that the differences for the middle 50% and top 25% are both zero is reported for each variable.

Although there is variation in what is statistically significant across grades 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%. Also, teachers serving classrooms in the top 25%, meaning that they have the highest proportions, in proportion free and reduced price lunch, proportion Hispanic, or proportion African-American tend to have lower stability ratios than teachers in bottom 25% and middle 50%. This pattern is true even after the number of student observations is fixed at 12.

9.1 DOLS Stabilities

Tables 4 and 5 show the results for the DOLS estimator. In 4th grade, the stability for the bottom 25% of average lag score is .137 and the difference between the bottom 25% and middle 50% of average lag score is .139 and statistically significant at the 5% level. This means that the inter-year stability for teachers in the middle 50% is roughly double that of teachers with students in the bottom 25%. This difference holds even after the number of student observations is fixed at 12. The difference between teachers receiving students in the top 25% and bottom 25% of lagged student achievement is .110 with an unrestricted number of student observations and .122 with the number of student observations fixed at 12, which is also nearly double that stability for the bottom 25%. The difference between the top 25% and bottom 25% ranges from -.107 to -.180 for proportion free and reduced priced lunch and Hispanic and is statistically significant at least at the 5% level. There isn't a statically significant difference for proportion African-American until the number of student observations is set to 12. In that case the difference between the top 25% and bottom 25% is -.134 and significant at the 1% level. In most cases, the inter-year stability is roughly half as large for those teachers in the top 25% compared to those in the bottom 25% for proportion free and reduce priced lunch, Hispanic, and African-American.

The results for 6th grade, show less variables that are statistically significant, but this could be due to somewhat smaller sample size. The only variable statistically significant is proportion Hispanic once the number of student observations is set to 12. The stability for those in the bottom 25% of proportion Hispanic is .238 and the difference between those in the top and bottom 25% is -.211. This point estimate suggests that the year to year stability for teachers found in the top 25% in proportion Hispanic is close to zero, although the stability isn't estimated very precisely.

9.2 EB Lag Stabilities

It should be noted that the sample size for the empirical Bayes estimators (EB Lag and EB gain) are somewhat larger than for POLS and DOLS. This is because regression packages that estimate teacher fixed effects by including teacher indicators arbitrarily drop a teacher in each school as a reference teacher. It would take extra programming to recover the teacher effects for the reference teachers.

The empirical Bayes estimator is designed to trade bias for efficiency. As explained in Kane and Staiger (2008), the empirical Bayes estimator minimizes the mean squared prediction error between the true teacher effect and the unshrunken estimate. We would expect that the stability ratios to be higher overall than those for POLS and DOLS, but since the empirical Bayes estimators in this analysis don't allow for heteroskedastic error, there is reason to suspect differences in stabilities across subgroups for the empirical Bayes estimators as well.

The inter-year stabilities do tend to be slightly higher overall for the empirical Bayes estimator using lagged scores (EB Lag) compared to DOLS (Tables 6 and 7).

There also appear to be large differences across subgroups for grades 4 and 6 and with the number of student observations fixed at 12 and without.

In 4th grade, the pattern is quite similar to DOLS for the average lagged score variable. The difference for the middle 50% and bottom 25% is .187 and statistically significant at the 1% with the number of student observations fixed at 12, and the difference between the top 25% and bottom 25% is .0828 and statistically significant at the 10% level. However, none of the variables for proportion free and reduced price lunch, Hispanic, and African-American are statistically different from zero.

The stabilities for 6th grade differ from the DOLS estimator. There are many more variables that are statically significant. With the number of student observations unrestricted, there is a statistically significant at the 1%

level difference of .144 for the difference in average lagged score for the middle 50%. Once the number of student observations is fixed at 12, the relationship is no longer statistically significant, but the difference between the top and bottom 25% does become statistically significant at the 5% level. There are also statistically significant differences at the 5% level for proportion free and reduced price lunch, Hispanic, and African-American for the top 25% once the number of student observations is fixed at 12.

9.3 POLS and EB Gain Stabilities

The stabilities for the POLS estimator are very similar to the DOLS estimator. Results can be viewed in the appendix of supplemental tables (Tables 9, 10, and 11) The relationships for POLS are strongest for 4th grade, which has the larger sample size. In 4th grade, there are statistically significant differences for multiple variables, with and without the number of student observations fixed. Once again in 6th grade, the only statistically significant difference is for the difference for the top 25% for proportion Hispanic with the number of student observations set to 12.

The relationships for the EB Lag and EB Gain estimators are also nearly identical (Table 10). Only 6th grade is reported, but 4th grade shows the same basic patterns for EB Gain as EB Lag.

9.4 An Additional Check of Whether Differences for 6th Grade are Weaker

As a check of whether the differences for 6th grade truly seemed to be near zero for DOLS and POLS, We changed the seed for the program that randomly drew the schools that were included in the sample. This generated a different data set from which we ran the analysis. The analysis was done

with the number of student observations set to 12. Results can be viewed in the appendix of supplemental tables (Table 13). In that case, we found a difference, statistically significant at the 10% level, of .121 for DOLS between those in the top 25% in lagged achievement and the bottom 25%, which is similar in size to those found in 4th grade in the prior sample. The difference for POLS between the top and bottom 25% in average lagged achievement is .112 and has a p-value of .113, so it just missed the 10% significance level. In addition there was a statistically significant difference at the 5% level with the proportion free and reduced priced lunch of -.154 for the middle 50% and -.159 for the top 25% compared to the bottom 25% for DOLS. The differences for free and reduced priced lunch for POLS were similar in magnitude to the DOLS estimates and statistically significant at the 5% level. Although these results are stronger and better make our case of differences in stabilities across subgroups, we reported the weaker results for full disclosure. It also is suggestive that the lack of statistical significance in the first set of results may be related to lack of power.

9.5 Inter-year Stabilities using Two Years of Data

Tables 8 shows the inter-year stabilities using two years of data. The results are only reported for 6th grade and with the number of student observations fixed to 12 in each year. Since the teacher effect estimates are based on two years, each teacher is linked to 24 student observations. The stability is of the teacher effect estimate in year $t-1$ which uses data from year $t-2$ and $t-1$ and the teacher effect estimate in year $t+1$, which uses years t and $t+1$. For a teacher to be included in one of the quartile groupings, the teacher had to have classes in that quartile range for all four years. This dramatically reduced the sample of teachers available to compare. Even still, statistically significant differences in inter-year stabilities are detected, suggesting that the differences possibly persist even as more data are added.

Table 8 shows a statistically significant difference at the 5% level of .309 between the middle 50% and bottom 25% for DOLS in the average lagged score category. This suggests that the inter-year stability for those in the bottom 25% is .181, while the stability for the middle 50% is approximately .490, although the estimates are somewhat noisy. This difference wasn't statistically different from zero for DOLS using only 1 year of data. This could be because of the additional screening that took place in the two year stabilities. The fact that only teachers with classes in the bottom 25% four years in a row were included in the bottom 25% category could mean that these teachers are particularly likely to receive extremely low scoring students in the prior year. This could lead to a more stark contrast than before in the inter-year stabilities using one year of data, which could be producing the statistically significant result. In that case, we are still discovering that teachers with different groups of students have different inter-year stabilities, but we are less sure exactly how much including additional years may have improved the relative stability across groups. This difference is nearly the same magnitude for POLS and is also statistically significant (Table 12 in the appendix of supplemental tables).

The only other statistically significant difference is for EB Lag for the proportion Hispanic category. (Table 8) Those teachers with the proportion of Hispanic students in the top 25% are estimated to have a .233 smaller stability than those in the bottom 25%. This is significant at the 10% level. There was also a statistically significant difference using one year of data for this categorization.

The inter-year stabilities using two years of data are not perfect for determining how much adding additional data improves the differences in stability, but perhaps provides some suggestive evidence that adding multiple years of data may not fully solve the issue.

10 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.

Since it conceivable that teachers of students with low average lag scores and high proportions of free and reduced price lunch, Hispanic, and African-American students are also low in experience, and low experienced teachers also have lower inter-year stabilities, the analysis was repeated dropping all teachers with less than 5 years of experience. However, the quadratic of teacher experience was included in the student level model, so the teacher effects should have been net of teacher experience, and the patterns described above still held in general in this analysis as expected.

As an additional sensitivity check, we repeated the analysis without school dummies. The analysis without school dummies tended to increase the inter-year stabilities for all sub-groups and all estimators. We were still able to detect statistically significant differences in inter-year stabilities across sub-groups.

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.

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

11 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 strong theoretical reason for suspecting that a student's error term is heteroskedastic and statistical test 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 suggestive 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. Some of this may be driven by a comparison of teachers with particularly extreme classroom compositions however. More research is needed to answer this question fully.

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 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. Other differences, the large differences between teachers in the top and bottom 25% in proportion free and reduced priced lunch, Hispanic, and African-American, may suggest that the heteroskedasticity goes beyond measurement, and could be related to differences in inputs to achievement itself. 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 re-

searchers and policymakers. Empirical Bayes estimators are commonly computed assuming homoskedastic student level error. This assumption doesn't 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. Future work will be to evaluate empirical Bayes estimators that do not make a homoskedasticity assumption.

Also, as shown in Jacob and Lefgren (2005), using an empirical Bayes estimate as a right hand side variable in a regression can circumvent the issue of attenuation bias caused by measurement error in the teacher effect estimate. This is true since the empirical Bayes estimate has the property of being the linear projection of the true teacher effect on the unshrunken estimated teacher effect, meaning that the empirical Bayes estimate is uncorrelated with the measurement error by definition. However, this feature of the empirical Bayes estimator relies on using the correct shrinkage term. Assuming the the student level error is homoskedastic can lead to the wrong shrinkage term and may produce biased estimates as a result. Future work will investigate how much bias may be created by using an improper shrinkage term.

Additionally, it is quite common for standard errors and the corresponding confidence intervals to also be based on a homoskedasticity assumption, particularly for estimators that treat teacher effects as random[7] and use a GLS approach for efficiency that assumes homoskedasticity. Making standard errors robust to heteroskedasticity is essentially admitting that the assumptions underlying the estimator are wrong, but should be done anyway. This is particularly important since the teacher value-added estimates are being

[7]This is typical of HLM estimators or empirical Bayes estimators. Ballou, Sanders, and Wright (2004) assume homoskedasticity in computing standard errors, as does the value-added estimator employed by the NYC school district (reference: NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model 2010)

made publicly available in some school districts. It is important that the confidence intervals accurately reflect imprecision caused by all sources of variability, not just the number of student observations.

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 author that care should be used in evaluating teachers using value-added estimators and that value-added estimates should not be used as the sole basis of any high stakes policy involving teachers.

12 Work Cited

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

American Institute for Research (2011), "Florida Value-Added Model Technical Report "Downloaded on April 1, 2012 at http://www.fldoe.org/committees/sg.asp

Ballou, D., Sanders, W., & Wright, P. (2004), "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Educational and Behavioral Statistics* 29(1), 37-65

Chetty, R., Freidman, J., & Rockoff, J. (2011), "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," *NBER*, Working Paper 17699

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012), "Can Value-Added Measures of Teacher Performance Be Trusted? ," *Education Policy Center at Michigan State University*, Working Paper 18

Hanushek, E. A. (1979), "Conceptual and empirical issues in the estimation of educational production functions," *Journal of Human Resources* 14(3), 351-388

Jacob, B. & Lefgren, L. (2005), "Principals as Agents: Subjective Performance Measurement in Education," *NBER*, Working Paper 11463

Kane, T. & Staiger, D. (2002), "The Promise and Pitfalls of Using Imprecise School Accountability Measures " *Journal of Economic Perspectives* 16(4), 91-114

Kane, T. & Staiger, D. (2008), "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," *NBER*, Working Paper 14607

Koedel, C. & Betts, J. (2007) "Re-Examining the Role of Teacher Quality In the Educational Production Function " *Working Paper* Downloaded on April 1, 2012 at http://econ.missouri.edu/working-papers/2007/wp0708_koedel.pdf

Lord, F.M. (1980) "Applications of Item Response Theory to Practical Testing Problems, "Hillsdale, NJ: Lawrence Erlbaum Associates

McCaffrey, D., Lockwood, J. R., Louis, T., & Hamilton, L. (2004), "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics* 29(1), 67101

McCaffrey, D., Lockwood, J. R., Sass, T., Mihaly, K. (2009), "The Inter-Temporal Variability of Teacher Effect Estimates " *Working Paper.* Downloaded on April 2, 2012 at http://www.performanceincentives.org/data/files/news/PapersNews/McCaffrey_et_al_20091.pdf

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association* 78(381), 47-55

Reckase, M. (2009) "Multidimensional Item Response Theory, "New York: Springer

Todd, P., & Wolpin, K. (2003), "On the Specification and Estimation of the Production Function for Cognitive Achievement," *The Economic Journal* 113, F3-F33

Value-Added Research Center (2010), "NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model 2010, "University of Wisconsin-Madison: Wisconsin Center for Education Research.

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

13 Appendix of Figures and Tables

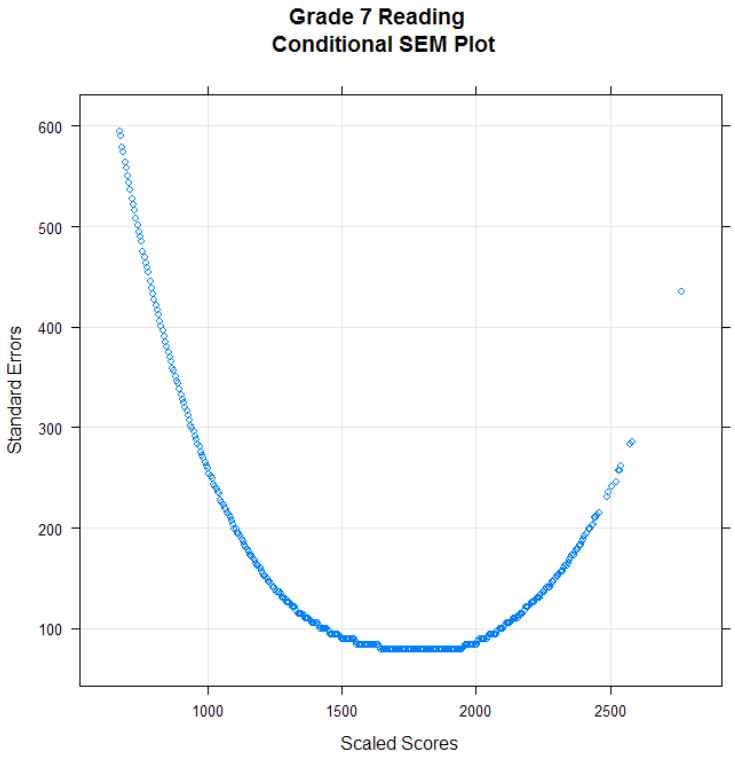

Figure 1: Estimated standard deviation of measurement error conditional on student achievement level in 7th grade reading test scores. Source: Florida Value-Added Technical Report, American Institute for Research, 2011

Table 1: Summary Statistics for Full Sample

Variable	Mean	Std. Dev.	Min.	Max.	N
Math Scale Score	1631.551	245.955	375	4007.625	
Math Gain Score	109.496	164.125	-996	999	
Reading Scale Score	1632.34	306.906	294	4004	
Reading Gain Score	115.875	216.202	-999	1000	
African-American	0.217	0.412	0	1	
Hispanic	0.219	0.414	0	1	
Free and Reduced Price Lunch	0.49	0.5	0	1	
Limited English Proficiency	0.176	0.38	0	1	
Teacher Years of Experience	9.834	9.332	0	80	
N		2196043			

Table 2: Summary Statistics for Restricted Sample

Variable	Mean	Std. Dev.	Min.	Max.
Math Scale Score	1639.186	246.566	512	2844
Math Gain Score	108.935	162.573	-969	998
Reading Scale Score	1639.762	307.054	294	2954
Reading Gain Score	116.062	216.367	-998	999
African-American	0.232	0.422	0	1
Hispanic	0.216	0.411	0	1
Free and Reduced Price Lunch	0.498	0.5	0	1
Limited English Proficiency	0.187	0.39	0	1
Teacher Years of Experience	9.944	9.352	0	80
# of Teachers	4096			
# of Schools	247			
N		210712		

Class Characteristics of Restricted Sample: 6th Grade

Variable	Mean	Std. Dev.	Min.	Max.
Avg. Lag Math Score	1656.925	120.807	569	2442
Prop. FRL	0.485	0.262	0	1
Prop. LEP	0.177	0.202	0	1
Prop. Hispanic	0.224	0.24	0	1
Prop. African- American	0.227	0.239	0	1
Class Size	37.144	16.755	1	139
# of Teachers	2697			
# of Schools	227			
N		185451		

Table 3: Heteroskedasticity Test: 6th Grade

Regressions of Squared Residuals on Sets of Covariates

VARIABLES	DOLS	POLS
Math Lagged Score	-142.1***	-845.3***
	(20.95)	(38.08)
Math Lagged Score Squared	-0.0297**	0.331***
	(0.0130)	(0.0234)
Math Lagged Score Cubed	2.51e-05***	-3.37e-05***
	(2.64e-06)	(4.73e-06)
African-American	7,041***	8,572***
	(206.1)	(236.4)
Hispanic	1,620***	2,197***
	(188.6)	(216.1)
FRL	6,290***	7,429***
	(148.6)	(167.3)
Observations	356,410	

Breusch-Pagan F-Test using Fitted Values

DOLS	F(1,356408)= 282.70	p-value < .0001
POLS	F(1,356408)= 15196.60	p-value < .0001

Summary Stats of Sq. Residuals	**Mean**	**Std. Dev.**
DOLS Squared Residuals	16874.361	43419.707
POLS Squared Residuals	19116.378	48795.105

Robust standard errors in parentheses

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

Table 4: DOLS 1 Year Inter-Year Stability with 4th Grade Observations

DOLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.132***	0.240***	0.280***	0.175***
	(0.0475)	(0.0346)	(0.0411)	(0.0550)
Middle 50% Difference	0.139**	0.0390	-0.0501	0.0580
	(0.0551)	(0.0462)	(0.0485)	(0.0602)
Top 25% Difference	0.110*	-0.108**	-0.180***	0.0227
	(0.0599)	(0.0506)	(0.0549)	(0.0657)
Observations	4,531	5,220	4,922	4,999
R^2	0.047	0.050	0.048	0.040
Joint Test	3.174	4.831	6.435	0.646
p-value	0.0420	0.00805	0.00163	0.524
Results with Student Observations Fixed at 12				
Bottom 25%	0.133***	0.252***	0.239***	0.278***
	(0.0381)	(0.0382)	(0.0379)	(0.0327)
Middle 50% Difference	0.127***	-0.0127	0.0166	-0.0643
	(0.0465)	(0.0453)	(0.0471)	(0.0437)
Top 25% Difference	0.122**	-0.107**	-0.0954**	-0.134***
	(0.0518)	(0.0493)	(0.0482)	(0.0458)
Observations	3,608	4,290	3,897	3,997
R^2	0.051	0.046	0.049	0.043
Joint Test	4.112	3.455	4.080	4.259
p-value	0.0165	0.0318	0.0171	0.0143

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference for teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference Difference for teachers above the 75th percenitle and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 5: DOLS 1 Year Inter-Year Stability with 6th Grade Observations

DOLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.134	0.158***	0.222**	0.175**
	(0.0831)	(0.0474)	(0.108)	(0.0745)
Middle 50% Difference	-0.0131	-0.0323	-0.159	-0.0626
	(0.104)	(0.0683)	(0.114)	(0.0851)
Top 25% Difference	0.0864	-0.0611	-0.0788	-0.0554
	(0.0927)	(0.0867)	(0.129)	(0.123)
Observations	2,355	2,400	2,558	2,430
R^2	0.020	0.015	0.026	0.016
Joint Test	1.081	0.274	1.289	0.274
p-value	0.340	0.760	0.276	0.760
Results with Student Observations Fixed at 12				
Bottom 25%	0.0976**	0.179***	0.238***	0.124***
	(0.0487)	(0.0474)	(0.0456)	(0.0426)
Middle 50% Difference	0.0365	-0.000178	-0.0427	0.0560
	(0.0581)	(0.0578)	(0.0579)	(0.0549)
Top 25% Difference	0.0462	-0.0626	-0.211***	0.00681
	(0.0626)	(0.0640)	(0.0607)	(0.0632)
Observations	2,144	2,433	2,225	2,212
R^2	0.021	0.024	0.035	0.027
Joint Test	0.292	0.754	7.317	0.645
p-value	0.747	0.471	0.000696	0.525

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference for teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference for teachers above the 75th percenitle and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 6: EB Lag 1 Year Inter-Year Stability with 4th Grade Observations

EB Lag	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.268***	0.362***	0.361***	0.368***
	(0.0346)	(0.0315)	(0.0489)	(0.0306)
Middle 50% Difference	0.177***	0.0527	0.00940	0.0226
	(0.0411)	(0.0384)	(0.0536)	(0.0402)
Top 25% Difference	0.0874*	-0.0543	-0.0360	-0.0274
	(0.0472)	(0.0445)	(0.0565)	(0.0496)
Observations	5,151	6,015	5,693	5,803
R^2	0.135	0.136	0.121	0.133
Joint Test	9.867	4.134	0.799	0.588
p-value	5.38e-05	0.0161	0.450	0.556
Results with Student Observations Fixed at 12				
Bottom 25%	0.204***	0.288***	0.319***	0.341***
	(0.0345)	(0.0355)	(0.0428)	(0.0305)
Middle 50% Difference	0.187***	0.0502	-0.00372	-0.0438
	(0.0412)	(0.0429)	(0.0510)	(0.0399)
Top 25% Difference	0.0828*	0.0201	-0.0119	-0.0784
	(0.0486)	(0.0482)	(0.0507)	(0.0491)
Observations	4,191	4,957	4,530	4,642
R^2	0.103	0.102	0.097	0.087
Joint Test	11.17	0.754	0.0363	1.348
p-value	1.49e-05	0.471	0.964	0.260

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 7: EB Lag 1 Year Inter-Year Stability with 6th Grade Observations

EB Lag	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.218***	0.321***	0.269***	0.258***
	(0.0415)	(0.0438)	(0.0531)	(0.0366)
Middle 50% Difference	0.144***	0.0294	0.0640	0.0890*
	(0.0556)	(0.0538)	(0.0618)	(0.0460)
Top 25% Difference	0.0937	-0.152***	-0.0520	-0.0889
	(0.0576)	(0.0580)	(0.0657)	(0.0565)
Observations	3,133	3,193	3,377	3,224
R^2	0.097	0.094	0.091	0.083
Joint Test	3.410	7.051	2.704	6.239
p-value	0.0333	0.000893	0.0673	0.00200
Results with Student Observations Fixed at 12				
Bottom 25%	0.163***	0.255***	0.255***	0.276***
	(0.0460)	(0.0436)	(0.0504)	(0.0443)
Middle 50% Difference	0.0295	0.00899	-0.00179	-0.0506
	(0.0542)	(0.0529)	(0.0603)	(0.0539)
Top 25% Difference	0.106*	-0.125**	-0.152**	-0.119**
	(0.0606)	(0.0579)	(0.0625)	(0.0595)
Observations	2,877	3,250	2,981	2,987
R^2	0.049	0.052	0.052	0.050
Joint Test	1.849	4.224	5.214	2.068
p-value	0.158	0.0148	0.00554	0.127

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 8: DOLS and EB Lag Stability, 6th Grade Observations, Student Observations fixed at 24, Pooled using Two Years

DOLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.181**	0.338***	0.265**	0.323***
	(0.0880)	(0.104)	(0.114)	(0.118)
Middle 50% Difference	0.309**	-0.105	-0.234	-0.0324
	(0.130)	(0.139)	(0.243)	(0.146)
Top 25% Difference	-0.0160	-0.144	0.0532	-0.137
	(0.115)	(0.145)	(0.154)	(0.164)
Observations	181	237	197	203
R^2	0.150	0.096	0.126	0.105
Joint Test	4.104	0.529	0.728	0.398
p-value	0.0190	0.590	0.485	0.672

EB Lag	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.355***	0.284**	0.479***	0.441***
	(0.0860)	(0.114)	(0.114)	(0.0922)
Middle 50% Difference	0.0487	0.0927	-0.173	-0.0746
	(0.129)	(0.134)	(0.171)	(0.112)
Top 25% Difference	-0.0794	0.000515	-0.233*	-0.0775
	(0.138)	(0.149)	(0.136)	(0.139)
Observations	244	325	274	288
R^2	0.140	0.131	0.185	0.166
Joint Test	0.394	0.415	1.472	0.249
p-value	0.675	0.661	0.232	0.780

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

14 Appendix of Supplemental Tables

Table 9: POLS 1 Year Inter-Year Stability with 4th Grade Observations

POLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.126**	0.239***	0.263***	0.193***
	(0.0496)	(0.0349)	(0.0442)	(0.0537)
Middle 50% Difference	0.124**	0.0367	-0.0316	0.0217
	(0.0571)	(0.0467)	(0.0517)	(0.0590)
Top 25% Difference	0.124*	-0.110**	-0.164***	0.00776
	(0.0638)	(0.0514)	(0.0563)	(0.0657)
Observations	4,531	5,220	4,922	4,999
R^2	0.040	0.046	0.043	0.037
Joint Test	2.558	4.674	6.069	0.0936
p-value	0.0776	0.00941	0.00235	0.911
Results with Student Observations Fixed at 12				
Bottom 25%	0.116***	0.243***	0.214***	0.277***
	(0.0392)	(0.0383)	(0.0383)	(0.0325)
Middle 50% Difference	0.138***	0.00108	0.0354	-0.0694
	(0.0467)	(0.0454)	(0.0472)	(0.0430)
Top 25% Difference	0.139***	-0.114**	-0.0694	-0.145***
	(0.0529)	(0.0500)	(0.0490)	(0.0460)
Observations	3,608	4,290	3,897	3,997
R^2	0.049	0.045	0.045	0.041
Joint Test	4.859	4.525	3.247	4.982
p-value	0.00785	0.0109	0.0391	0.00694

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 10: POLS 1 Year Inter-Year Stability with 6th Grade Observations

POLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.115	0.161***	0.209**	0.196***
	(0.0781)	(0.0496)	(0.102)	(0.0698)
Middle 50% Difference	-0.00699	-0.0270	-0.160	-0.0934
	(0.102)	(0.0737)	(0.108)	(0.0821)
Top 25% Difference	0.0999	-0.0630	-0.0693	-0.0943
	(0.0889)	(0.0857)	(0.128)	(0.116)
Observations	2,355	2,400	2,558	2,430
R^2	0.016	0.015	0.023	0.016
Joint Test	1.243	0.275	1.446	0.690
p-value	0.289	0.760	0.236	0.502
Results with Student Observations Fixed at 12				
Bottom 25%	0.0978*	0.180***	0.223***	0.127***
	(0.0502)	(0.0476)	(0.0438)	(0.0436)
Middle 50% Difference	0.0129	-0.0200	-0.0608	0.0267
	(0.0596)	(0.0575)	(0.0571)	(0.0557)
Top 25% Difference	0.0374	-0.0752	-0.202***	0.00354
	(0.0636)	(0.0647)	(0.0608)	(0.0640)
Observations	2,144	2,433	2,225	2,212
R^2	0.016	0.021	0.028	0.024
Joint Test	0.200	0.773	5.961	0.143
p-value	0.819	0.462	0.00266	0.867

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percenitle and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 11: EB Gain 1 Year Inter-Year Stability with 6th Grade Observations

EB Gain	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.217***	0.308***	0.250***	0.258***
	(0.0422)	(0.0436)	(0.0510)	(0.0392)
Middle 50% Difference	0.131**	0.0340	0.0735	0.0755
	(0.0561)	(0.0540)	(0.0597)	(0.0484)
Top 25% Difference	0.0891	-0.148**	-0.0342	-0.0919
	(0.0581)	(0.0580)	(0.0662)	(0.0575)
Observations	3,133	3,193	3,377	3,224
R^2	0.089	0.086	0.083	0.076
Joint Test	2.787	6.901	2.267	5.494
p-value	0.0619	0.00104	0.104	0.00419
Results with Student Observations Fixed at 12				
Bottom 25%	0.162***	0.240***	0.235***	0.277***
	(0.0467)	(0.0447)	(0.0483)	(0.0430)
Middle 50% Difference	0.0319	0.0232	0.000365	-0.0629
	(0.0553)	(0.0537)	(0.0588)	(0.0527)
Top 25% Difference	0.0988	-0.117**	-0.138**	-0.122**
	(0.0604)	(0.0589)	(0.0611)	(0.0593)
Observations	2,877	3,250	2,981	2,987
R^2	0.046	0.048	0.043	0.047
Joint Test	1.558	4.306	4.314	2.135
p-value	0.211	0.0136	0.0135	0.119

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 12: POLS and EB Gain Stability, 6th Grade Observations, Student Observations fixed at 24, Pooled using Two Years

POLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.194**	0.291**	0.207	0.314**
	(0.0847)	(0.127)	(0.135)	(0.140)
Middle 50% Difference	0.289**	-0.0776	-0.177	-0.0707
	(0.124)	(0.160)	(0.260)	(0.164)
Top 25% Difference	-0.0357	-0.0894	0.0986	-0.132
	(0.122)	(0.157)	(0.175)	(0.183)
Observations	181	237	197	203
R^2	0.150	0.094	0.119	0.093
Joint Test	3.980	0.174	0.650	0.261
p-value	0.0213	0.841	0.524	0.771

EB Gain	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.396***	0.344***	0.480***	0.399***
	(0.0794)	(0.114)	(0.119)	(0.0967)
Middle 50% Difference	0.0745	0.0286	-0.123	-0.0656
	(0.138)	(0.135)	(0.191)	(0.116)
Top 25% Difference	-0.0974	-0.0487	-0.211	-0.0200
	(0.137)	(0.150)	(0.142)	(0.138)
Observations	244	325	274	288
R^2	0.155	0.133	0.185	0.141
Joint Test	0.588	0.203	1.112	0.185
p-value	0.557	0.817	0.331	0.831

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 13: DOLS and POLS Stability, 6th Grade Observations, Student Observations fixed at 12, Alternate Sampling Seed

DOLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.107**	0.280***	0.107**	0.135***
	(0.0497)	(0.0497)	(0.0508)	(0.0504)
Middle 50% Difference	0.0807	-0.154***	0.0460	0.0622
	(0.0602)	(0.0585)	(0.0614)	(0.0587)
Top 25% Difference	0.121*	-0.159**	0.00431	-0.0455
	(0.0734)	(0.0726)	(0.0712)	(0.0686)
Observations	2,017	2,237	2,080	2,249
R^2	0.028	0.029	0.017	0.025
Joint Test	1.501	3.809	0.397	2.007
p-value	0.223	0.0224	0.672	0.135

POLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.129***	0.265***	0.0958*	0.150***
	(0.0476)	(0.0484)	(0.0509)	(0.0506)
Middle 50% Difference	0.0381	-0.145**	0.0617	0.0383
	(0.0581)	(0.0575)	(0.0611)	(0.0597)
Top 25% Difference	0.112	-0.130*	0.0326	-0.0525
	(0.0706)	(0.0698)	(0.0712)	(0.0683)
Observations	2,017	2,237	2,080	2,249
R^2	0.027	0.027	0.018	0.024
Joint Test	1.289	3.300	0.528	1.328
p-value	0.276	0.0372	0.590	0.266

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.

Table 14: DOLS and EB Lag Stability, 4th Grade Observations, Student Observations fixed at 24, Pooled using Two Years

DOLS	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.237*	0.510***	0.432***	0.404***
	(0.123)	(0.0990)	(0.106)	(0.0975)
Middle 50% Difference	0.329*	-0.244**	-0.270*	-0.210
	(0.183)	(0.122)	(0.141)	(0.150)
Top 25% Difference	0.0382	-0.119	-0.303**	0.0695
	(0.164)	(0.156)	(0.131)	(0.163)
Observations	260	409	386	390
R^2	0.151	0.124	0.075	0.102
Interactions Test	1.937	2.037	2.880	1.547
p-value	0.147	0.133	0.0582	0.215
EB Lag	Avg Lag Score	Prop FRL	Prop Hisp.	Prop Black
Bottom 25%	0.317***	0.511***	0.636***	0.375***
	(0.111)	(0.0970)	(0.163)	(0.0882)
Middle 50% Difference	0.332**	-0.0442	-0.463**	0.0284
	(0.138)	(0.135)	(0.190)	(0.122)
Top 25% Difference	0.0557	-0.161	-0.269	0.0266
	(0.166)	(0.162)	(0.171)	(0.169)
Observations	324	484	467	470
R^2	0.234	0.169	0.168	0.123
Interactions Test	3.572	0.498	3.197	0.0300
p-value	0.0296	0.608	0.0424	0.970

Standard errors clustered at teacher level in parentheses

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

Middle 50% Difference: Difference between teachers in the 25-75 percentile and those below 25th percentile.

Top 25% Difference: Difference between teachers above the 75th percentile and those below 25th percentile.

Joint Test: F-test statistic that Middle 50% Difference and Top 25% Difference coefficients zero.