Guarino, Cassandra; Reckase, Mark D.; Stacy, Brian; Wooldridge, Jeffrey M.

Working Paper

A Comparison of Growth Percentile and Value-Added Models of Teacher Performance

IZA Discussion Papers, No. 7973

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Guarino, Cassandra; Reckase, Mark D.; Stacy, Brian; Wooldridge, Jeffrey M. (2014) : A Comparison of Growth Percentile and Value-Added Models of Teacher Performance, IZA Discussion Papers, No. 7973, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/93312

DISCUSSION PAPER SERIES

IZA DP No. 7973

A Comparison of Growth Percentile and Value-Added Models of Teacher Performance

Cassandra Guarino
Mark Reckase
Brian Stacy
Jeffrey Wooldridge

February 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A Comparison of Growth Percentile and Value-Added Models of Teacher Performance

Cassandra Guarino
Indiana University and IZA

Mark Reckase
Michigan State University

Brian Stacy
Michigan State University

Jeffrey Wooldridge
Michigan State University and IZA

Discussion Paper No. 7973
February 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7973
February 2014

ABSTRACT

A Comparison of Growth Percentile and Value-Added Models of Teacher Performance*

School districts and state departments of education frequently must choose between a variety of methods to estimating teacher quality. This paper examines under what circumstances the decision between estimators of teacher quality is important. We examine estimates derived from growth percentile measures and estimates derived from commonly used value-added estimators. Using simulated data, we examine how well the estimators can rank teachers and avoid misclassification errors under a variety of assignment scenarios of teachers to students. We find that growth percentile measures perform worse than value-added measures that control for prior year student test scores and control for teacher fixed effects when assignment of students to teachers is nonrandom. In addition, using actual data from a large diverse anonymous state, we find evidence that growth percentile measures are less correlated with value-added measures with teacher fixed effects when there is evidence of nonrandom grouping of students in schools. This evidence suggests that the choice between estimators is most consequential under nonrandom assignment of teachers to students, and that value-added measures controlling for teacher fixed effects may be better suited to estimating teacher quality in this case.

JEL Classification: I20, J08, J24, J45

Keywords: teacher labor markets, teacher value-added, teacher quality

Corresponding author:

Cassandra Guarino
Indiana University
School of Education
201 N. Rose Avenue
Bloomington, IN 47405
USA
E-mail: guarino@indiana.edu

* 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.

1 Introduction

Currently researchers and policymakers can choose among a number of statistical approaches to measuring teacher effectiveness based on student test scores. Given a relative lack of easily accessible information on the pros and cons of different methodological choices, the choice of a method is often based on replicating what others in similar contexts or disciplines have done rather than carefully weighing the relative merits of each approach. Policymakers, for example, will often opt for a procedure that has been used in other states. An example would be the increasingly popular Colorado Growth Model based on the work of Betebenner (2012) and whose use is now spreading to other states, such as Indiana and Massachusetts. Researchers, on the other hand, have tended to rely on value-added models (VAMs) based on OLS or GLS regression techniques. The distinction between growth modeling procedures and OLS-based value-added models in the context of teacher performance evaluation—and the relative merits of each approach—has not been fully explored. This paper addresses this task.

Teacher performance measures can be used for different purposes. In some cases researchers or administrators wish to rank a set of teachers in terms of their effectiveness–those in a particular grade in a particular district, for example. Both growth percentile methods and VAMs can be used for this purpose. The primary technical distinction between VAMs and growth percentile models is that the former produce an estimate of the magnitude of a teacher's effectiveness relative to her peers and the latter yield information only on a teacher's rank in the distribu-

tion of student growth percentiles. In other words, the former method is potentially capable of revealing how much better or worse a teacher may be relative to peers whereas the latter reveals only the relative position a teacher holds among other teachers

When test scores are vertically scaled from one year to the next, VAMs may produce estimates of teacher effectiveness that can be interpreted as the average amount of achievement growth an individual teacher contributes to her students. Some argue that methods that do not use vertically scaled test scores are preferable. Barlevy and Neal (2011), for example, favor simple rankings based on growth percentile models, arguing that pay schemes based on performance measures that simply order teachers can induce teachers to exert an optimal level of effort. They point out that these models do not require a vertical scale, so the analyst may use test forms with no item overlap or equating, citing this as an advantage over VAMs. It is important to note, however, that VAMs can be employed regardless of whether test scores are vertically scaled. Therefore a central question is whether there are any advantages to using one approach versus another for the purpose of ranking teachers.

To explore this research question, we evaluate the merits of growth models versus VAMs with regard to the goal of ranking teachers, since both approaches can accomplish this task. Both types of approaches face a common set of challenges when applied to the task of determining teacher effectiveness rankings. Perhaps the most important of these is the issue of bias under conditions of nonrandom assignment of students to teachers. To compare how well the two approaches deal

with these challenges, we use them to rank teachers using simulated data in which the true underlying effects are known. The simulated data sets are created to represent varying degrees of challenge to the estimation process: some of our data generating processes randomly assign students to teachers, others do so in non-random ways. In addition to the simulation study, we compare growth percentile models to VAMs using administrative data from a large diverse southern state.

Previous studies comparing growth percentile models with VAMs in measuring educational performance have been limited to empirical investigations of actual data. Wright et al. (2010) compares the EVAAS methodology and student percentile growth models–both of which treat teacher effects as random–and finds substantial agreement. Goldhaber et al. (2013) compare a subset of value-added models that treat teacher effects as fixed with student growth percentile models and find varying degrees of divergence depending upon on the characteristics of the sample. Ehlert et al. (2013) investigate school-level value added and find substantial divergence between growth percentile models and different types of VAMs, although they endorse a noncausal approach, which treats school effects as random for reasons of expediency in policymaking.

A primary contribution of our study is to use simulations to understand and explain the fundamental differences among the estimators and to then target the investigation of the empirical data in ways that highlight the conditions under which they diverge and how these may affect policy applications regarding teacher value-added. We find that growth percentile models and VAMs rank teachers very similarly under random assignment of students to teachers. However, when stu-

dents are nonrandomly assigned to teachers, VAMs that treat teacher effects as fixed outperform both growth percentile models and VAMs that treat teacher effects as random, such as average residuals or empirical Bayes. Thus the primary distinction to be made among different approaches to modeling teacher performance is whether they merely describe classroom achievement growth or whether they make headway in isolating the teacher's role in producing that growth–a task that is particularly challenging in the context of nonrandom assignment.

We begin with a description of the different types of models, beginning with two different growth percentile approaches and following with four types of VAMs. We then apply the various estimators to the task of ranking teachers using simulated data and compare their ability to rank teachers accurately. This is followed by a discussion and conclusions.

1.1 Description of the Models

Both growth percentile and value-added approaches can take various forms. In this paper, we consider two types of growth modeling approaches: one based on quantile regression as in Betebenner (2012)and one based on nearest neighbor matching of students across classrooms, as implemented by Fryer et al. (2012).

Likewise, we consider more than one type of commonly-used value-added model. Three of these utilize OLS regression but differ in their specifications: one is based on a dynamic specification that treats teacher effects as fixed, one is based on a gain-score specification that also treats teacher effects as fixed, and one computes teacher effects by averaging residuals, thus treating teacher effects

as random. In addition to these three VAMs, we also consider an empirical Bayes approach, based on GLS, or more commonly referred to as HLM.

1.1.1 Colorado Growth Model Estimation Procedure

The growth model creates a metric of teacher effectiveness by calculating the median or mean conditional percentile rank of student achievement in a given year for students in a teacher's class. For a particular student with current year score A_{ig} and score history $\{A_{i,g-1}, A_{i,g-2}, ..., A_{i,1}\}$, one locates the percentile corresponding to the student's actual score, A_{ig}, in the distribution of scores conditional on having a test score history $\{A_{i,g-1}, A_{i,g-2}, ..., A_{i,1}\}$. In short, the analyst evaluates how high in the distribution the student achieved, given their past scores. Then teachers are evaluated by the median or average conditional percentile rank of their students.

Here, we briefly describe the estimation procedure used in the Colorado Growth Model. Details of this approach can be found in Betebenner (2011) . Quantile regressions are used to estimate features of the conditional distribution of student achievement. In particular, one estimates the conditional quantiles for all possible test score histories, which are then used for assigning percentile ranks to students. Using the notation in Betebenner (2011) , the τ-th conditional quantile is the value $Q_y(\tau|x)$ such that

$$Pr(y \leq Q_y(\tau|x)) = \tau$$

The conditional quantiles are then modeled for achievement scores as:

$$Q_{A_{ig}}(\tau|A_{i,g-1}, A_{i,g-2}, ..., A_{i,1}) = \sum_{j=1}^{g-1}\sum_{k=1}^{6}\phi_{ik}(A_{i,j})\beta_{ik}(\tau) \quad (1)$$

where ϕ_{ik} denote B-spline basis functions of prior test scores. Six knots are used at the lowest score, 20th percentile, 40th percentile, 60th percentile, 80th percentile, and the highest score[1]. As discussed in Betebenner (2011) , the B-spline functions are chosen to improve model fit by adding flexibility in the treatment of prior test scores as covariates, primarily in that they allow for nonlinearities in the relationship between current and prior scores. Several available prior year test scores can be used as regressors, if available, and estimation is done using quantile regression. In practice, student and family background variables are included in the regressions.

To be specific, 100 quantile regressions are estimated, one for each percentile. Regressions are run separately for each grade and year. Conditional test scores are estimated for each percentile by generating fitted values from the regressions as follows:

$$\hat{Q}_{A_{ig}}(\tau|A_{i,g-1}, A_{i,g-2}, ..., A_{i,1}) = \sum_{j=1}^{g-1}\sum_{k=1}^{6}\phi_{ik}(A_{i,j})\hat{\beta}_{ik}(\tau)$$

A student's conditional percentile rank is then computed by counting the number of conditional percentiles that result in fitted test scores that are smaller than

[1]These knots were chosen based on a phone conversation with Dr. Betebenner. We would like to thank him for his valuable time and generous help with the details of the model.

the student's current grade test score, A_{ig}. For example, a student has a conditional percentile rank of 20 if there are 20 percentiles estimated lower than or equal to their score, in which case:

$$\hat{Q}_{A_{ig}}(.20|A_{i,g-1}, A_{i,g-2}, ..., A_{i,1}) \leq A_{ig} < \hat{Q}_{A_{ig}}(.21|A_{i,g-1}, A_{i,g-2}, ..., A_{i,1})$$

Once conditional percentile ranks are computed for all students, teachers are assigned a score equal to the median conditional percentile rank of the students within their class. These scores cannot reveal how much better students performed in one teacher's class compared with another, but can be used to form rankings of teachers by their estimated effectiveness.

An attractive feature of growth percentile models is that the student growth percentiles,once computed, can be used to provide a variety of descriptive portraits. They can be averaged (or the median can be taken) by classroom, grade, school, district, etc. Such models were originally developed to provide a description of student growth and were not intended to form a basis for determining the causal impact of a teacher (Betebenner (2009)).

1.2 Growth Percentile Model using Nearest Neighbor Matching

The second growth percentile model we consider is the approach proposed in Barlevy and Neal (2011) as a basis for distributing merit pay to teachers and applied

by Fryer et al. (2012) in an experimental context. The method consists of matching students based on their test score histories. Each student is matched to nine other students in, say, the district, with similar prior year test scores. Fryer et al. (2012) match students based on up to three prior year test scores and match only to students in different schools. For students with fewer than three prior year scores, as many as are available are used.

The Mahalanobis distance measure is used for matching. For two students, i and h, the distance is computed by the following formula:

$$(\mathbf{A_h} - \mathbf{A_i})' \hat{\Sigma}_A^{-1} (\mathbf{A_h} - \mathbf{A_i}) \tag{2}$$

where $\mathbf{A_h}$ and $\mathbf{A_i}$ are the vectors of past achievement scores and $\hat{\Sigma}_A$ is the sample variance-covariance matrix of the past achievement scores.

Students are matched to other students with the smallest distance from their prior achievement scores. Once nine matches are found for each student, students are ranked within this group of ten according to how they perform on the achievement test in the current year. Teacher evaluations are then computed, based on the average percentile rank of students within their class.

1.3 VAMs

Value added models attempt to model the achievement process over time and are based on the broad notion that achievement at any grade can be modeled as a

function of both past and current child, family, and schooling inputs[2]. In its most general formulation, the model can be expressed as:

$$A_{ig} = f_g(E_{ig}, \ldots, E_{i0}, X_{ig}, \ldots, X_{i0}, c_i, u_{ig})$$

where A_{ig} is achievement of student i in grade g, E_{ig} is a vector of educational inputs including teacher, school, and classroom characteristics, and in some cases a set of teacher indicators, X_{ig} consists of a set of relevant time-varying student and family inputs, c_i is an unobservable student fixed effect (representing, for example, motivation, some notion of sustained ability, or some persistent behavioral or physical issue that affects achievement), and the u_{ig} is an idiosyncratic, time varying error term. In this very general formulation, the functional form is unspecified and can vary over time.

To estimate this function, several assumptions are generally made. The functional form is considered to be more or less linear and unchanging over time, learning "decay" (that is, the amount of forgetting that takes place over time) is generally assumed to be constant for all inputs over time, and the time-constant student effect is assumed to either be ignorable or, at least, constant in its impact over time[3]. The resultant value-added model is typically expressed as follows:

$$A_{ig} = \tau_g + \lambda A_{i,g-1} + E_{ig}\beta + X_{ig}\gamma + c_i + e_{ig} \tag{3}$$

[2]See Hanushek (1979) or Todd and Wolpin (2003)

[3]For a full explication of the assumptions applied in value-added models, see Todd and Wolpin (2003), Harris et al. (2011), and Guarino et al. (2012b)

where $A_{i,g-1}$ is the prior year achievement score of student i and only current schooling and family inputs are required for estimation.[4] When value-added models are used to estimate teacher effects, the E_{ig} vector generally consists of indicator variables for specific teachers.[5]

There are several ways of estimating equation (1) to compute teacher effects. We focus on four value-added estimators that form the basis for most of the common procedures currently in use. A nice feature of value-added estimators is that, with a vertical scale, an analyst can not only rank teachers but also judge, subject to sampling variation, how much better one teacher is compared with another in terms of the their contribution to student growth. Of course, when the assumptions are met that allow these estimators to recover each teacher's contribution to student growth, we can also successfully order teachers according to their effectiveness.

A full discussion of all of these estimators' econometric properties can be found in Guarino et al. (2012b).

1.3.1 Dynamic OLS (DOLS)

A simple estimator for equation (1) involves OLS regression to estimate λ, β, and γ. We refer to this estimator as DOLS because it contains the lagged test score (or in many applications, more than one lagged score) on the right hand

[4]It is also common to include multiple prior years of achievement, other subject scores, and sometimes polynomials of both as regressors.

[5]The vector may also consists of exposure variables (i.e. the fraction of the year that a student spend with a particular teacher)

side of the equation along with a full set of teacher indicator [6]variables. Teacher effect estimates are then constructed from the coefficients on the teacher indicator variables. This estimator ignores the presence of c_i, but the inclusion of teacher indicators in addition to prior year test scores specifically adjusts the teacher effect estimates for nonrandom assignment to students based on prior year scores, as explained in Guarino et al. (2012b).

1.3.2 OLS using Gain Scores (OLS-Gain)

A second value-added estimator assumes that the persistence rate of past inputs, λ, is equal to one. In this case, the model collapses to a gain score equation:

$$\Delta A_{ig} = \tau_g + E_{ig}\beta + X_{ig}\gamma + c_i + e_{ig} \tag{4}$$

This model is also estimated using OLS. As in the case of DOLS, a full set of teacher indicators are included to estimate teacher effects. If λ is not equal to one, then a portion of prior achievement–specifically $(\lambda - 1)A_{i,g-1}$ –is left in the error term and will cause omitted variable bias when assignment is based on prior test scores.

1.3.3 Average Residual (AR) and Empirical Bayes (EB)

Another estimator of equation (1) is the average residual estimator, which we will refer to as AR. This estimator also uses OLS regression to estimate λ and

[6]or exposure

coefficients on the other covariates, except that it is typical not to include teacher indicators in the regressions. Instead, teacher effect estimates are recovered by averaging the regression residuals within a classroom.

Often researchers and policy analysts choose to shrink the average residual measures towards the mean teacher effect, with the shrinkage term being related to the variance of the unshrunken estimator. This is often referred to as an empirical Bayes approach, although the true empirical Bayes relies on GLS rather than OLS[7]. The variance of the estimator for an individual teacher effect can differ from teacher to teacher because of differences in class size as well as other sources of heteroskedasticity. In our simulation, we only evaluate the unshrunken average residual measure, since we do not vary class size and there are no sources of heteroskedasticity. In this special case, the unshrunken average residuals are perfectly correlated with the shrunken estimates, since the shrinkage term is identical for every teacher. In our application of application of value-added models to actual administrative data, we include both the AR (unshrunken) and the empirical Bayes estimator based on GLS, which we abbreviate as EB.

As discussed in Guarino et al. (2012b) the decision not to include teacher indicators in the regression can be costly when the assignment of teachers to students is nonrandom because the correlation between the assignment mechanism (say, prior test scores) and the teacher effects is not partialled out of the effect estimates. Assuming a correct functional form and nonrandom assignment condi-

[7]See Guarino et al. (2012a) for a complete explanation and derivation of the empirical Bayes estimator in its application to teacher evaluation.

tional on the other covariates, then including teacher indicators can produce consistent estimates of teacher effects. However, that is not the case when averaging the residuals.

A type of omitted variable bias can result in the teacher effect estimates if we are unable to control for the assignment mechanism. Under random assignment of students to teachers, many omitted variable issues would be considerably mitigated. However, classroom assignments are not always random, and, indeed, it is not necessarily desirable to do so. Random assignment deprives parents of the ability to request teachers whom they believe to be best suited for their children, and it deprives principals of one of their most important functions: to maximize overall achievement by matching the individualized skills of teachers to those students most likely to benefit from them. Thus random assignment—while helpful from an evaluation standpoint—could result in sub-optimal learning conditions if particular teacher characteristics interact in a beneficial way with student characteristics in the learning process.

1.3.4 Colorado Growth Model under Nonrandom Assignment

Under random assignment of teachers to students, it is possible to attribute high achievement in the conditional distribution of test scores (conditional on prior scores) to strong instruction by the teacher. It is sensible to use the conditional distribution, since students start off with different levels of knowledge. However, under nonrandom assignment of teachers to student, it may no longer be possible to attribute high achievement in the conditional distribution to good teaching.

To illustrate the reason, consider a thought experiment in which the best students are assigned to the best teachers and the worst students are assigned to the worst teachers in a model school district with 4 teachers and 4 classrooms:

The four teachers have differing teacher abilities. Let teacher i have teaching ability β_i and

$$\beta_1 < \beta_2 < \beta_3 < \beta_4$$

Suppose that all students within a classroom are identical. Also, suppose that classroom 1 and classroom 2 have identical initial achievement, $A_{1,g-1} = A_{2,g-2}$ and classroom 3 and classroom 4 have identical initial achievement, $A_{3,g-1} = A_{4,g-2}$.

$$A_{1,g-1} = A_{2,g-2} < A_{3,g-1} = A_{4,g-2}$$

Also, assume for simplicity that teachers are the only input into achievement.

In the Colorado Growth Model approach, students are compared to other students with the same initial achievement levels. Since students in classrooms 1 and 2 are identical at the start of the year, students in classroom 1 and 2 will be compared to one another. Students in classrooms 3 and 4 will be compared to one another as well, since their initial achievement levels are the same. Also, since $\beta_1 < \beta_2$ then all students in class 1 score below students in class 2. In this case, the median conditional percentile of teacher 1's students will be below the median for teacher 2's students. Likewise the median conditional percentile of teacher 3's

students will be below teacher 4's.

Using the Colorado growth model approach, teachers 1 and 3 actually will have the same median conditional percentile and so teachers 1 and 3 will have the same ranking, even though $\beta_1 < \beta_3$. Teacher 3 is also rated below teacher 2, even though $\beta_2 < \beta_3$. Finally, teachers 2 and 4 will have the same rankings, even though $\beta_2 < \beta_4$. In this simple illustration, nonrandom assignment of teachers to students can lead to the wrong conclusions in some cases.

1.3.5 Nearest Neighbor Matching under Nonrandom Assignment

Under nonrandom assignment of teachers to students, the nearest neighbor matching method has issues similar to the Colorado growth model approach.

Assume again that the best students are assigned to the best teachers in each school. For simplicity, also assume that there are 10 identical schools with 2 teachers and 2 classrooms within each school.

Assume teacher 1 is a worse teacher than teacher 2 in every school, so:

$$\beta_1 < \beta_2$$

Also, that all students in classroom 1 have lower prior achievement than those in classroom 2, so:

$$A_{1,g-1} < A_{2,g-1}$$

Using the nearest neighbor matching method, students in classroom 1 will be

matched to students in classroom 1 in the other schools. Students in classroom 2 will be matched to other students in classroom 2 in the other schools.

In this case, assuming that the teachers numbered 1 in all schools have the same effects and that teachers numbered 2 in all schools are also the same and for simplicity that teachers are the only input into education, teacher 1 and 2 in every school will be evaluated as being identical, even though $\beta_1 < \beta_2$. This is true since classroom 1 students are only compared with classroom 1 students in other schools who also get the low ability teacher. Similarly for classroom 2 students.

This simple example is meant to illustrate the main point that matching similar students to one another is not enough to guarantee an accurate ranking of teachers under nonrandom assignment in which the best (or worst) teachers are assigned to the best students. If principals are tracking students by ability and assigning the gifted class to the best teacher, then nearest neighbor matching may not work effectively. No partialling out between teachers and students takes place, as it does in the DOLS estimator for instance.

2 Simulation

Our data are constructed to represent one elementary grade that normally undergoes standardized testing in a hypothetical district. To mirror the basic structural conditions of an elementary school system for, say, grade 3, we create data sets that contain students nested within teachers nested within schools. Our simple baseline data generating process is as follows:

$$A_{i3} = \lambda A_{i2} + \beta_{i3} + c_i + u_{i3}$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of child i entering third grade, A_{i3} is the achievement score of child i at the end of third grade, λ is a time constant persistence parameter, β_{i3} is the teacher-specific contribution to growth (the true teacher value-added effect), c_i is a time-invariant child-specific effect, and u_{i3} is a random deviation for each student. We assume independence of u_{i3}. We assume that the time-invariant child-specific heterogeneity c_i is correlated at about 0.5 with the baseline test score A_{i2}. In the simulations reported in this paper, the random variables A_{i2}, β_{i3}, c_i, and u_{i3} are drawn from normal distributions. The standard deviation of the teacher effect is .25, while that of the student fixed effect is .5, and that of the random noise component is 1, each representing approximately 5, 19, and 76 percent of the total variance in gain scores, respectively[8].

Our data structure has the following characteristics that do not vary across simulation scenarios:

- 10 schools
- 1 grade (3rd grade), with a base score in 2nd grade
- 4 teachers per grade and school (thus 40 teachers overall)

[8]These relativeeffect sizes are based on prior research (e.g. Nye et al. (2004), McCaffrey et al. (2004), and Lockwood et al. (2007)). We changed the relative effect sizes as sensitivity checks and found no substantive differences.

- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools

To create different scenarios, we vary certain key features: the grouping of students into classes, the assignment of classes of students to teachers within schools, and the amount of decay in prior learning from one period to the next. Students are grouped either randomly or dynamically. In the case of dynamic grouping, students are ordered by their prior year achievement scores and grouped into classrooms. In this scenario, the students with the lowest prior year scores tend to be grouped in classes together, and students with the highest scores tend to be grouped together [9]. Also, there is random assignment and nonrandom assignment of teachers to the classrooms. There are two nonrandom assignment scenarios. The first is positive assignment, where the best teachers are assigned to the highest performing classrooms. The second is negative assignment, where the worst teachers are assigned to the highest performing classes. We vary the amount of persistence in past test scores, λ, in the data generating process. We consider a case with full persistence, $\lambda = 1$, and partial persistence, $\lambda = .5$. 100 simulation replications are performed for each grouping-assignment-persistence rate combination. Finally, we perform the estimation using achievement scores that are vertically scaled and also using scores that are standardized within grade, which artificially breaks the vertical scaling.

[9] There is a small amount of noise built into the assignment process

2.1 Results

We will begin with the results in which vertically scaled test scores are used to compute the measures. As mentioned before, when test scores are vertically scaled, VAMs have the advantage of being able to estimate how much the average student gained with a given teacher and how much they could have gained with another teacher. This information is not available with the Colorado growth model (CGM) approach or with nearest neighbor matching (NNM) approach. We will present Spearman rank correlations of the estimated teacher effects with the true teacher effects as a measure of performance, since all estimators of teacher performance are capable of ranking teachers. In addition, we will present a measure of misclassification. The measure we choose is the percent of teachers that have a true teacher effect above the 25th percentile that are rated in the bottom 25% using a teacher quality measure.

In the random grouping and random assignment scenario (RG-RA) all of the estimators perform fairly well. The results for λ set to .5 and for λ set to 1 are similar. All three VAMs outperform the growth percentile models, but these latter models still perform reasonably well, with rank correlations around .82 and .85. The misclassification rates are best for DOLS, AR, and OLS-Gain as well, with rates of 6%, 8%, and 8%. The CGM and NNM estimators have a slightly larger misclassification rate of 10% and 9% respectively in the lambda 1 case.

In the case of dynamic grouping coupled with random assignment of groups to teachers (DG-RA) the results are quite similar. The rank correlations for DOLS and AR drop slightly to .87, and the rank correlation for OLS-Gain drops to .83.

The CGM and NNM rank correlations are nearly identical and actually slightly higher for the CGM estimator. The misclassificaiton rates are fairly stable as well.

Once assignment of teachers to students is nonrandom the results change considerably. In the DG-PA scenario, in which students with the highest prior year achievement level tend to be assigned to teachers with the highest value added, the growth percentile estimators perform far worse than DOLS. The DOLS estimator has a rank correlation of .88, whereas the CGM and NNM estimators have rank correlations of .71 and .75 respectively in the $\lambda = 1$ and $\lambda = .5$ cases. The rank correlation also decreases for AR, which also fails to properly partial out the relationship between teacher and student quality, with a correlation of .78. OLS-Gain performs well in the DG-PA scenario, but poorly in the DG-NA scenario, when lambda is 1. The pattern reverses when lambda is .5. This is due to biases sometimes amplifying the teacher effects, making the teachers with the good teacher effects look even better than they actually are and those that are worse look worse. This has the effect of making the rank correlations strong, even though there is substantial bias. The misclassification rates show a pattern as the rank correlations. DOLS does the best in terms of misclassification, while the CGM and NNM estimators have misclassification rates that increase roughly 2-3 percentage points compared to the random assignment scenarios.

The results for the dynamic grouping with negative assignment (DG-NA) case look similar to those for the DG-PA scenario. DOLS outperforms the AR, CGM, and NNM estimators that do not partial out.

In the case of the standardized test scores, the patterns are largely the same.

In the RG-RA and DG-RA scenarios, DOLS, AR, CGM, and NNM estimators perform similarly with rank correlations in the .82-.87 range. The results in the DG-PA and DG-NA look nearly identical to those in the vertically scaled test scores scenario. DOLS still outperforms the estimators that fail to partial out.

One estimator that is particularly harmed is the OLS-Gain estimator. When scores are standardized, the OLS-Gain estimator, which consists of a regression of gain scores on teacher indicators, performs very poorly under nonrandom assignment.

2.2 Simulation with Random Noise Component Drawn from t Distribution with 3 d.f.

One claim that could be made about the Colorado growth model is that the teacher rankings may be more robust to outliers, since the quantile regression estimators used in the ranking method are themselves less affected by outliers. If the distribution is thicker tailed, the Colorado growth model may perform better than the estimators based on OLS. As a robustness check, therefore, we examine the performance of the estimators when the idiosyncratic error term u_{i3} is drawn from a t distribution with three degrees of freedom. The t distribution with three d.f. has much thicker tails than the normal distribution. Figure 1 in the appendix shows the pdf of the Normal(0,1) pdf and the t(3) pdf.

Results are reported in table 3. Only results using the vertically scaled test scores are reported. Under random grouping and random assignment (RG-RA)

the CGM and NNM estimators slightly outperform the value-added estimators. The CGM estimator has a rank correlation of .72, and the NNM estimator has a rank correlation of .73 in the $\lambda = 1$ case. The value-added estimators have a slightly lower but very similar rank correlation of .71.

Under the dynamic grouping and nonrandom assignment (DG-PA and DG-NA) scenarios, DOLS again outperforms the CGM and NNM estimators that do not properly partial out the relationship between the covariates and the teacher's value-added. The rank correlation for DOLS is .70 and .71 for the DG-PA and DG-NA scenarios in the $\lambda = 1$ case. The rank correlation for the CGM estimator is .57 and .59 for the DG-PA and DG-NA scenarios, and the rank correlation is .60 and .59 for the NNM estimator.

An important takeaway from this analysis is that there may be cases were using the CGM or NNM estimators is preferable. One case may be when the distribution is thick tailed and there is random grouping and assignment. However, as the simulations show, even in the thick tailed case, nonrandom grouping and assignment still poses a threat to the CGM and NNM estimators.

3 Empirical Analysis

In the next section, we examine the correlations between the estimators using real data. A main finding from the simulations was that value-added estimators such as DOLS and the CGM estimator provide similar rankings under random assignment, but somewhat different rankings under nonrandom assignment.

Using real data, we find patterns suggesting a similar relationship between DOLS and the Colorado estimator when we compare correlations and classification rates for teachers in schools with little evidence of nonrandom grouping with those using teachers in schools with evidence of nonrandom grouping.

3.1 Data

The data come from an administrative data set in large and diverse anonymous school district. It consists of 215,411 usable student year observations from years 2002-2007 and grades 5 and 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. 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 to 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. Students in charter schools are not included in this analysis, since charter schools 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 Tables 4 and 5.

3.2 Analysis

The nearest neighbor matching estimator was dropped from the analysis using real data due to computational limitations and because the issues this estimator faces are mirrored in the CGM estimator. Matching each student to 9 other students cannot be done quickly when the data set is large[10], as far as we are aware. Also we use an empirical Bayes estimator of value-added instead of a more simple average residual estimator, since class size does vary substantially in the real data set. In the following only estimates from the DOLS, EB-Lag, OLS-Gain, and Colorado growth model estimators are reported.

The simulation results indicated that in situations where students were dynamically grouped based on prior year test scores and were nonrandomly assigned to teachers the DOLS estimator maintained a strong correlation with the true teacher effect, while the Colorado Growth estimator performed less well. In order to examine whether the Colorado growth model may perform less well in actual data, we performed the test of nonrandom grouping that was performed in Dieterle et al. (2012). The test involved a multinomial logit regression of each student's classroom assignment on the student's prior year test score for each school-grade-year combination in the data. Finding that students' prior year test scores significantly predict their classroom assignment is taken as evidence that nonrandom grouping based on prior test scores occurs in that particular school-grade-year. Since nonrandom grouping is a precondition for nonrandom grouping and assignment, we

[10]It will take approximately 76 days to calculate teacher quality measures using the NNM estimator in just one large in the state we examine using a high performance computer.

focus on teachers in schools that reject the test of random grouping and compare them with teachers in schools that fail to reject. Results are reported in tables 6, 7, and 8.

All value-added models include the student's free-and-reduced price lunch status, English learner status, gender, and indicators for whether the student is black or hispanic. Value-added estimates and the estimates for the Colorado growth model are computed using one year of data[11]. DOLS and the EB-Lag estimator include two prior years mathematics scores as controls. In the EB-Lag and EB-Gain estimator, the student's class average prior year test score is included as a control for peer effects. The Colorado growth model estimates only include two prior year mathematics scores as controls in the quantile regressions, since this is how the estimator is described in Betebenner (2011)[12].

The correlation between DOLS and the CGM using one year of data[13] is .713 in the real data. This correlation is appreciably smaller than the correlation between DOLS and the EB-Lag (.96), OLS-Gain (.87), and EB-Gain (.94) estimators.

When the sample of teachers is broken into those teachers in school-grade-years where we find evidence of nonrandom grouping and those in school-grade-

[11]We have also examined the correlations when we pool across years. All correlations across estimators are higher than when only one year of data is used. We speculate that this is driven by greater precision using pooled data.

[12]As a sensitivity check we also estimate the CGM rankings by also including the other student demographics. In another sensitivity check, we also estimate the value-added models using only previous test scores as controls. This somewhat alters the correlations, but the main patterns described below still hold

[13]Two prior years of test scores are included. We mean that each teacher quality measure is estimated cohort by cohort.

years where we do not, we see a pattern that accords with the pattern seen in the simulation. The correlation between DOLS and the CGM estimator, found in table 7, is .695 in school-grade-years with nonrandom grouping and .746 in school-grade-years where we can't reject the hypothesis of random grouping, in table 8. The correlations between DOLS and OLS-Gain, as well as the EB estimators, changes slightly when comparing the nonrandom and random grouping school-grade-years, rising from .842 to .917 for OLS-Gain and from .953 to .973 for EB-Lag. This again is similar to what took place in the simulations.

As another check we examine a measure of disagreement between the estimators in terms of who is classified in the bottom 25% of teachers. We calculate the fraction of teachers rated in the bottom 25% using one estimator that are not rated in the bottom 25% using the other estimators. Results are reported in tables 9 to 11. Similar to the pattern indicated by the rank correlations, there is less disagreement between the estimators in the cases of schools with little evident of nonrandom grouping. The fraction of teachers rated in the bottom 25% using the DOLS estimator not rated in the bottom 25% using the CGM estimator is .36 in nonrandom grouping schools and .32 in random grouping schools.

4 Conclusions

In this paper, we compare commonly used value-added estimators to two alternative estimators that have been proposed: the Colorado growth model approach and the nearest neighbor matching estimator.

Simulation evidence indicates that the performance of these estimators depends on how students are grouped and assigned to teachers. In cases where students are nonrandomly grouped based on prior year test scores and nonrandomly assigned to teachers, the Colorado growth model and nearest neighbor estimators perform poorly compared with the DOLS estimator, which partials out the relationship between student's prior year achievement and the teacher assignment. The DOLS estimator is also robust to the case where vertically scaled test scores are not used.

The performance of the estimators also depends to some extent on the distribution of the error term in the achievement model. When a fatter tailed t distribution with 3 d.f. is used for the error term, DOLS and the other value-added estimators perform worse than the Colorado growth model and nearest neighbor matching approacher, but only slightly so. In this case, DOLS still outperforms the Colorado growth model and nearest neighbor matching estimators when there is nonrandom grouping and assignment.

Additionally, we compare the estimators using actual data. The patterns of divergence between DOLS and the growth model approach when there is nonrandom grouping and assignment uncovered in the simulations are also detected when we divide the sample into teacher in schools with evidence of nonrandom grouping versus teachers in schools in which grouping is fairly random.

This paper provides evidence that nonrandom grouping and assignment can negatively affect the popular Colorado growth modeling approach, as well as other growth percentile models such as the nearest neighbor matching approach.

Care should be used by practitioners and researchers in evaluating teachers using these approaches when nonrandom grouping and assignment occurs in the school system. More generally, estimators that do not partial out teacher effects–not only growth models, but also value-added models that are relatively descriptive in nature–are less equipped to disentangle true teacher contributions to student achievement from other source of achievement than those that partial out these effects.

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.

Barlevy, Gadi and Derek Neal, "Pay for percentile," Technical Report, National Bureau of Economic Research 2011.

Betebenner, D, "Norm-and Criterion-Referenced Student Growth," *Educational Measurement: Issues and Practice*, 2009, *28* (4), 42–51.

Betebenner, Damian W, "A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment," 2011.

_ , "Growth, standards, and accountability," *GJ Cizek, Setting Performance Standards: Foundations, Methods & Innovations*, 2012, pp. 439–450.

Dieterle, Steven G, Cassandra Guarino, Mark D Reckase, and Jeffrey M Wooldridge, "How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added," *Michigan State Education Policy Center*, 2012.

Ehlert, Mark, Cory Koedel, Eric Parsons, and Michael J. Podgursky, "The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence From School- and Teacher-Level Models in Missouri," *Statistics and Public Policy*, 2013, *1* (1), 19–27.

Fryer, Roland G, Steven D Levitt, John List, and Sally Sadoff, "Enhancing the efficacy of teacher incentives through loss aversion: A field experiment," Technical Report, National Bureau of Economic Research 2012.

Goldhaber, Dan, Joe Walch, and Brian Gabele, "Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments," *Statistics and Public Policy*, 2013, *1* (1), 28–39.

Guarino, Cassandra M, Michelle Maxfield, Mark D Reckase, Paul Thompson, and Jeffrey M Wooldridge, "An Evaluation of Empirical Bayes Estimation of Value-Added Teacher Performance Measures," *Michigan State Education Policy Center*, 2012.

Guarino, Cassandra Marie, Mark D Reckase, and Jeffrey M Wooldridge, "Can Value-Added Measures of Teacher Performance Be Trusted?," *Michigan State University Education Policy Center*, 2012, (Working Paper 18).

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.

Koenker, Roger and Kevin F. Hallock, "Quantile Regression," *The Journal of Economic Perspectives*, 2001.

Lockwood, JR, Daniel F McCaffrey, Laura S Hamilton, Brian Stecher, Vi-Nhuan Le, and José Felipe Martinez, "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures," *Journal of Educational Measurement*, 2007, *44* (1), 47–67.

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.

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

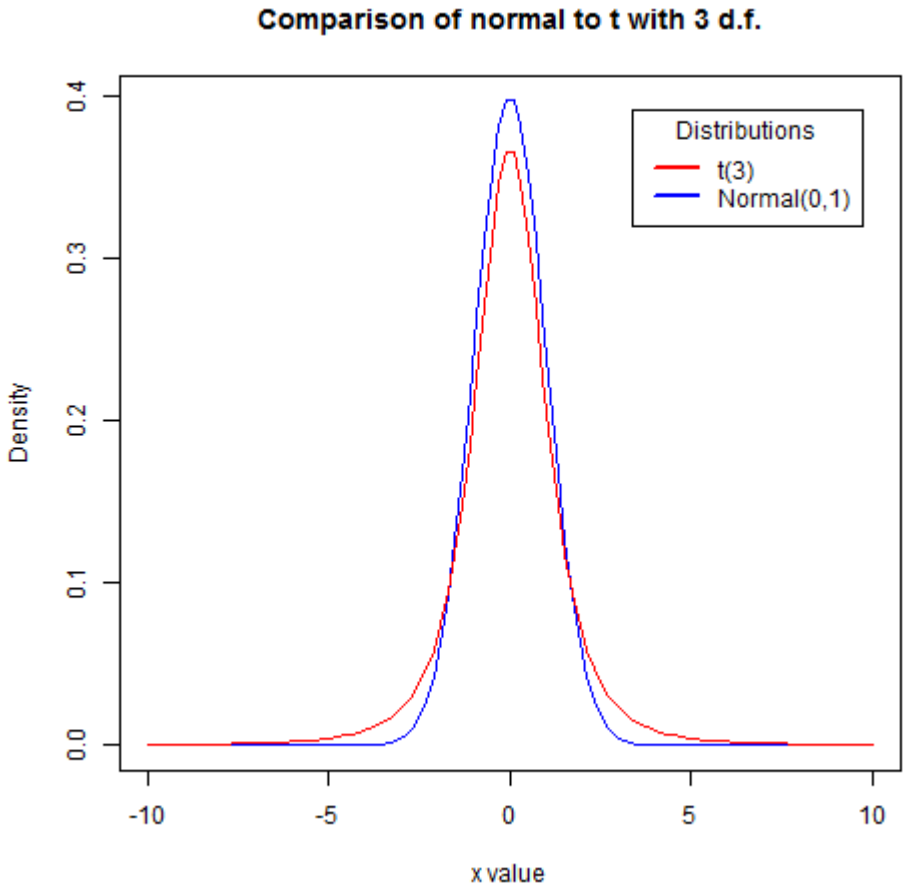
Nye, Barbara, Spyros Konstantopoulos, and Larry V Hedges, "How large are teacher effects?," *Educational evaluation and policy analysis*, 2004, *26* (3), 237–257.

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

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.

Wright, S Paul, John T White, William L Sanders, and June C Rivers, "SAS® EVAAS® statistical models," *SAS White Paper*, 2010.



5 Appendix of Tables and Figures

Table 1: Results from 100 replications. Normal(0,1) Errors. Row 1: Average rank correlation Row 2: Percentage of teachers above bottom 25% in true effect misclassified in bottom 25%

Estimator	**DOLS**	**AR/EB-Lag**	**OLS-Gain**	**CGM**	**NNM**	**Corr DOLS/CGM**
Assign Mech	$\lambda = 1$ Vertically Scaled Test Scores					
RG-RA	0.88	0.88	0.87	0.82	0.85	.93
	6%	8%	8%	10%	9%	
DG-RA	0.87	0.87	0.83	0.83	0.85	.93
	7%	8%	9%	9%	9%	
DG-PA	0.88	0.78	0.91	0.71	0.75	.87
	7%	11%	7%	12%	12%	
DG-NA	0.87	0.77	0.52	0.71	0.72	.87
	8%	10%	16%	12%	12%	
	$\lambda = 1$ Standardized Test Scores					
RG-RA	0.88	0.88	0.86	0.82	0.85	.93
	8%	8%	8%	9%	9%	
DG-RA	0.87	0.87	0.76	0.83	0.85	.93
	7%	8%	10%	9%	9%	
DG-PA	0.88	0.78	-0.11	0.71	0.75	.87
	8%	11%	27%	12%	12%	
DG-NA	0.87	0.77	0.91	0.71	0.72	.87
	7%	10%	7%	12%	12%	

Table 2: Results from 100 replications. Normal(0,1) errors. Row 1: Average rank correlation Row 2: Percentage of teachers above bottom 25% in true effect misclassified in bottom 25%

Estimator	**DOLS**	**AR/EB-Lag**	**OLS-Gain**	**CGM**	**NNM**	**Corr DOLS/CGM**
Assign Mech	$\lambda = .5$ Vertically Scaled Test Scores					
RG-RA	0.88	0.88	0.87	0.82	0.85	.93
	8%	8%	8%	10%	9%	
DG-RA	0.87	0.87	0.83	0.83	0.85	.93
	7%	8%	9%	9%	9%	
DG-PA	0.88	0.78	0.53	0.71	0.75	.87
	8%	11%	17%	12%	12%	
DG-NA	0.87	0.77	0.91	0.71	0.72	.87
	7%	10%	7%	12%	12%	
	$\lambda = .5$ Standardized Test Scores					
RG-RA	0.88	0.88	0.85	0.82	0.85	.93
	8%	8%	9%	9%	9%	
DG-RA	0.87	0.87	0.69	0.83	0.85	.93
	7%	7%	12%	9%	9%	
DG-PA	0.88	0.78	-0.33	0.71	0.75	.87
	8%	11%	30%	12%	12%	
DG-NA	0.88	0.77	0.89	0.72	0.73	.87
	7%	10%	8%	12%	12%	

Table 3: Results from 100 replications. Errors have t distribution with 3 d.f. Row 1: Average rank correlation Row 2: Percentage of teachers above bottom 25% in true effect misclassified in bottom 25%

Estimator	**DOLS**	**AR/EB-Lag**	**OLS-Gain**	**CGM**	**NNM**	**Corr DOLS/CGM**
Assign Mech		$\lambda = 1$ Vertically Scaled Test Scores				
RG-RA	0.71	0.71	0.71	0.72	0.73	.85
	10%	10%	10%	12%	11%	
DG-RA	0.72	0.72	0.67	0.71	0.72	.84
	11%	11%	13%	12%	11%	
DG-PA	0.70	0.58	0.84	0.57	0.60	.80
	11%	13%	08%	14%	14%	
DG-NA	0.71	0.59	0.16	0.59	0.59	.80
	11%	14%	23%	15%	15%	
	$\lambda = .5$ Vertically Scaled Test Scores					
RG-RA	0.71	0.71	0.71	0.71	0.73	.85
	9%	9%	10%	12%	11%	
DG-RA	0.72	0.72	0.71	0.71	0.72	.84
	11%	11%	11%	12%	11%	
DG-PA	0.70	0.58	0.57	0.57	0.59	.80
	11%	13%	14%	14%	14%	
DG-NA	0.71	0.59	0.78	0.59	0.59	.80
	11%	14%	9%	15%	15%	

Table 4: Summary statistics

Grade 5

Student Level Characteristics

Variable	Mean	Std. Dev.	Min.	Max.
Math Scale Score	1630.033	239.368	569	2456
Reading Scale Score	1552.276	321.701	474	2713
Math Scale Standardized Score	-0.081	1.009	-5.149	3.705
Reading Scale Standardized Score	-0.149	0.986	-4.020	3.605
Black	0.281	0.45	0	1
Hispanic	0.597	0.491	0	1
Free and Reduced Price Lunch	0.703	0.457	0	1
Limited English Proficiency	0.507	0.5	0	1
N		110970		

Teach Level Characteristics

Avg. Lag Math Score	1456.094	152.644	806.769	1986.808
Prop. FRL	0.718	0.249	0	1
Prop. LEP	0.508	0.259	0	1
Prop. Hispanic	0.584	0.322	0	1
Prop. Black	0.3	0.341	0	1
Class Size	24.019	7.929	12	145
Teacher Experience	9.374	10.101	0	47
N		4620		

Table 5: Summary statistics

Grade 6

Student Level Characteristics

Variable	Mean	Std. Dev.	Min.	Max.
Math Scale Score	1641.693	247.982	770	2492
Reading Scale Score	1618.179	311.402	539	2758
Math Scale Standardized Score	-0.14	0.971	-3.707	3.354
Reading Scale Standardized Score	-0.192	0.969	-4.049	3.526
Black	0.288	0.453	0	1
Hispanic	0.6	0.49	0	1
Free and Reduced Price Lunch	0.705	0.456	0	1
Limited English Proficiency	0.511	0.5	0	1
N		104441		

Teach Level Characteristics

Variable	Mean	Std. Dev.	Min.	Max.
Avg. Lag Math Score	1608.182	143.225	903.733	2053.576
Prop. FRL	0.727	0.218	0	1
Prop. LEP	0.515	0.238	0	1
Prop. Hispanic	0.589	0.31	0	1
Prop. Black	0.307	0.33	0	1
Class Size	65.113	42.807	12	216
Teacher Experience	7.668	8.978	0	40
N		1604		

Table 6: Correlations across Estimators

Variables	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
DOLS	1.00				
EB-Lag	0.96	1.00			
OLS-Gain	0.87	0.91	1.00		
EB-Gain	0.94	0.98	0.93	1.00	
CGM	0.71	0.68	0.60	0.69	1.00
Teacher/Year Obs	5666	5666	5666	5666	5666

Table 7: Correlations across Estimators - Nonrandom Grouping Schools

Variables	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
DOLS	1.00				
EB-Lag	0.95	1.00			
OLS-Gain	0.84	0.89	1.00		
EB-Gain	0.94	0.98	0.91	1.00	
CGM	0.70	0.65	0.56	0.67	1.00
Teacher/Year Obs	3672	3672	3672	3672	3672

Table 8: Correlations across Estimators - Random Grouping Schools

Variables	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
DOLS	1.00				
EB-Lag	0.97	1.00			
OLS-Gain	0.92	0.94	1.00		
EB-Gain	0.96	0.99	0.96	1.00	
CGM	0.75	0.72	0.68	0.73	1.00
Teacher/Year Obs	1876	1876	1876	1876	1876

Table 9: Fraction of Teachers Rated in Bottom 25% in the Initial Estimator Who are Not Rated in Bottom 25% in Another

		Not Rated Bottom 25%				
	Initial Estimator	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
	DOLS	0				
Rated	EB-Lag	.14	0			
Bottom	OLS-Gain	.26	.22	0		
25%	EB-Gain	.17	.10	.21	0	
	CGM	.34	.38	.42	.38	0
	Teacher/Year Obs	5666	5666	5666	5666	5666

Table 10: Fraction of Teachers Rated in Bottom 25% in the Initial Estimator Who are Not Rated in Bottom 25% in Another - Nonrandom Grouping Schools

		Not Rated Bottom 25%				
	Initial Estimator	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
	DOLS	0				
Rated	EB-Lag	.15	0			
Bottom	OLS-Gain	.27	.23	0		
25%	EB-Gain	.17	.09	.21	0	
	CGM	.36	.40	.44	.39	0
	Teacher/Year Obs	3672	3672	3672	3672	3672

Table 11: Fraction of Teachers Rated in Bottom 25% in the Initial Estimator Who are Not Rated in Bottom 25% in Another - Random Grouping Schools

		Not Rated Bottom 25%				
	Initial Estimator	DOLS	EB-Lag	OLS-Gain	EB-Gain	CGM
	DOLS	0				
Rated	EB-Lag	.12	0			
Bottom	OLS-Gain	.23	.20	0		
25%	EB-Gain	.17	.11	.17	0	
	CGM	.32	.35	.38	.35	0
	Teacher/Year Obs	1876	1876	1876	1876	1876