The Education Policy Center
AT MICHIGAN STATE UNIVERSITY

WORKING PAPER #39

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

Cassandra M. Guarino
Mark D. Reckase
Brian W. Stacy
Jeffrey M. Wooldridge

Michigan State University

February 6, 2014
Revised September 9, 2014

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

Author Information

Cassandra M. Guarino, Indiana University

Mark D. Reckase
Brian W. Stacy
Jeffrey W. Wooldridge
Michigan State University

This work was supported by grant numbers R305D100028 and R305B0030011 from the Institute for Education Sciences in the U.S. Department of Education.

Abstract

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 measure and estimates derived from commonly used valued-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.

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

Cassandra Guarino

Mark Reckase

Brian Stacy

Jeffrey Wooldridge

September 8, 2014

Abstract

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 student 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 include 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.

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 is the increasingly popular student growth percentile (SGP) model, which has been used extensively in Colorado and has now spread to other states such as Indiana and Massachusetts.[1] 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, have not been fully explored. This paper contributes to this investigation.

Teacher performance measures can be used for different purposes. Typically, researchers or administrators use them to rank a set of teachers in terms of their effectiveness–those in a particular grade or district, for example. Both SGPs and VAMs can be used for this purpose. One distinction between VAMs and SGPs, however, is that the former can produce an estimate of the magnitude of a teacher's effectiveness in terms of achievement and the latter yield information only on a

[1] Due to its long term use in Colorado, this method is sometimes referred to as the "Colorado growth model."

teacher's effect on his or her students' relative position in the growth distribution.[2] When test scores are vertically scaled from one year to the next or in a standardized form, VAM estimates can be interpreted as the average amount of achievement growth an individual teacher contributes to his or her students.[3] This distinction between VAMs and SGPs disappears, however, when percentile scores are used in place of vertically scaled or standardized test scores in value-added regressions.

Since both SGPs and VAMs are primarily used in practice to rank teachers on the basis of their measured effectiveness, we investigate the relative merits of SGPs versus VAMs with regard to the goal of ranking teachers by their effectiveness, since both approaches can accomplish this task and are typically used in this manner. 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 this challenge, 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

[2]This distinction permits policymakers and others to claim that growth models are not value-added models.

[3]Vertically scaled test scores allow for a comparison of student knowledge across years. In theory, with a vertical scale, a student score of 500 in third grade and a score of 550 in fourth grade would indicate that the student made a 50 point learning gain. However, it is sometimes the case that so-called vertical scales produce very similar scores for individual students from year to year, leading one to question whether they truly capture growth over time. The issue of whether the vertical scales can successfully be produced by test developers is controversial. For instance, Ballou (2009) and Barlevy & Neal (2011) critique the use of a vertical scale in teacher performance evaluation, and Briggs & Weeks (2009) show that school-level value-added estimates can be sensitive to different scaling methods.

varying degrees of challenge to the estimation process: some of our data generating processes randomly assign students to teachers, others do so in nonrandom 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 SGPs with VAMs in measuring educational performance have focused on empirical investigations of actual data. Wright, White, Sanders, & Rivers (2010) compares the EVAAS methodology with student percentile growth models – both of which make the assumption that teacher effects are uncorrelated with the regressors – and finds substantial agreement. Goldhaber, Walch, & Gabele (2013) compare a subset of value-added models that treat teacher effects as fixed, meaning the teacher effects can be arbitrarily correlated with the regressors, with student growth percentile models and find varying degrees of divergence depending upon on the characteristics of the sample. Ehlert, Koedel, Parsons, & Podgursky (2013) investigate school-level value added and find substantial divergence between growth percentile models and certain types of VAMs.

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 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 when students are randomly assigned to teachers. However, when students are nonrandomly assigned to teachers, VAMs that control for teacher assignment

outperform both growth percentile models and other VAMs that assume the regressors and teacher effects are uncorrlelated, such as those that average residuals or employ empirical Bayes. Thus a key distinction to be made among different models to estimate and rank teacher effectiveness is whether or not they control for teacher assignment.

We begin with a description of the different types of models, beginning with two SGP approaches and following with three 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. Following this, we compare teacher rank correlations across the different estimators in real data to investigate the consequences of using one method versus another. This is followed by a discussion and conclusions.

2 Description of the Models

Both growth percentile and value-added approaches can take various forms. In this paper, we consider two SGPs commonly used in practice, both based on the work of Betebenner (2012).[4] In one case, teachers are rated on the *median* stu-

[4]There are other percentile (or rank) based methods that are similar to the SGP methods, such as the approach proposed in Barlevy & Neal (2011) as a basis for distributing merit pay to teachers and applied by Fryer, Levitt, List, & Sadoff (2012) in an experimental context, although its use in accountability policies is rare. 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, and then teachers are evaluated on how their students compare to the nine other students they are matched with. We have examined the matching estimator proposed by Fryer, Levitt, List, & Sadoff (2012) using our simulations and found that it performed similarly to the SGP methods evaluated.

dent growth percentile of their students, and, in the other, teachers are rated on the *mean* student growth percentile. We also consider more than one type of commonly-used value-added model. One is based on a dynamic specification that treats teacher effects as fixed by partialling them out from other covariates. Another computes teacher effects by averaging residuals, thus not partialling out teacher assignment from the covariates. The third is an Empirical Bayes' (EB) approach, which uses generalized least squares (GLS) to estimate the parameters on the covariates and then uses a shrinkage estimator on the GLS residuals to obtain the teacher effects. It, too, does not partial out teacher assignment from the other covariates. The EB approach is a special case of a hierarchical linear model (HLM).

2.1 SGP Estimation Procedure

The SGP 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 mean conditional percentile rank of their students.

Here, we briefly describe the estimation procedure used in the SGP 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

$$Pr(y \leq Q_y(\tau|x)|x) = \tau. \tag{1}$$

The conditional quantiles are then modeled for achievement scores as:

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

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[5]. 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 not included in the regressions.[6]

[5]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.

[6]There is no conceptual reason why these other student background variables cannot be included. Future work examining how the omission of these variables affects estimates may be useful, although it is beyond the scope of this study.

To be specific, 100 quantile regressions are estimated, one for each percentile.[7] 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:

$$\hat{Q}_{A_{ig}}(\tau|A_{i,g-1},A_{i,g-2},...,A_{i,1}) = \sum_{j=1}^{g-1}\sum_{k=1}^{6}\phi_{ik}(A_{i,j})\hat{\beta}_{ik}(\tau). \quad (3)$$

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 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,[8] in which case:

$$\hat{Q}_{A_{ig}}(.20|A_{i,g-1},A_{i,g-2},...,A_{i,1}) \leq A_{ig} < \hat{Q}_{A_{ig}}(.21|A_{i,g-1},A_{i,g-2},...,A_{i,1}). \quad (4)$$

Once conditional percentile ranks are computed for all students, teachers are assigned a score equal to the median or mean conditional percentile rank of the students within their class. These scores cannot reveal how much better students

[7] In practice, sometimes more than 100 quantile regressions are estimated. It is more correct to say that 100 quantile regressions are run for each unique combination of prior year scores. As described in Betebenner (2011), if students are missing prior year scores, then all available scores up to, say, three are used. This means that multiple sets of 100 quantile regressions are computed for the different combinations of available prior year scores.

[8] In this illustration and throughout the paper, we allow for the estimation of 100 conditional quantiles for simplicity. In the R SGP software, it is possible to estimate several more intermediate percentiles, such as .005, .015, etc.

performed in one teacher's class compared with another, but can be used to form rankings of teachers by their estimated effectiveness.[9]

An attractive feature of growth percentile models is that, once computed, the student growth percentiles can be used to provide a variety of descriptive portraits. Such models were originally developed to provide a description of student growth and were not intended to form the basis for determining the impact of individual teachers (Betebenner (2009)). However, it is important to note that these measures have played a role in school accountability policies for several years, particularly in states such as Colorado.

2.2 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.[10] In its most general formulation, the model can be expressed as:

$$A_{ig} = f_g(E_{ig}, \ldots, E_{i0}, X_{ig}, \ldots, X_{i0}, c_i, u_{ig}), \tag{5}$$

[9] VAM models attempt to show how much a student's achievement increases after being exposed to a teacher. SGPs position students on a percentile distribution corresponding to their growth. Therefore, an underlying achievement distribution with a large spread can produce the same teacher ratings for the SGP model as an underlying distribution with a tight spread, so we do not know how much growth is associated with a particular teacher. VAMs using percentile scores instead of actual or standardized achievement scores would also be subject to this limitation, however.

[10] See Hanushek (1979) or Todd & Wolpin (2003)

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[11]. The resultant value-added model is typically expressed as follows:

$$A_{ig} = \lambda A_{i,g-1} + E_{ig}\beta + X_{ig}\gamma + c_i + e_{ig}, \tag{6}$$

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.[12] When value-added models are used to estimate teacher effects, the E_{ig} vector generally consists of

[11]For a full explication of the assumptions applied in value-added models and the statistical properties of different value-added estimators, see Todd & Wolpin (2003), Harris, Sass, & Semykina (2011), and Guarino, Reckase, & Wooldridge (2015)

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

indicator variables for specific teachers.[13]

There are several ways of estimating equation (6) to compute teacher effects. We focus on three value-added estimators that form the basis for most of the common procedures currently in use. A potentially useful 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 more one teacher contributes to student achievement than another. Of course, the estimates can also be used simply to order teachers according to their effectiveness.

2.2.1 Dynamic OLS (DOLS)

A simple estimator for equation (6) involves OLS regression to estimate λ, β, and γ. We refer to this estimator as "dynamic" OLS, because it contains the lagged test score (or in many applications, more than one lagged score) on the right hand side of the equation. The DOLS estimator, using our terminology, also contains a full set of teacher indicator variables.[14] 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, Reckase, & Wooldridge (2015). An additional feature of DOLS is that it allows for the di-

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

[14]Instead of binary indicator variables, one can also include a student's level of exposure to a teacher.

rect estimation of standard errors pertaining to each teacher effect estimate, thus enabling researchers to determine whether a teacher is statistically significantly different from, say, an average teacher.[15]

2.2.2 Average Residual (AR) and Empirical Bayes' (EB-Lag)

Another approach to estimating equation (6) is to use OLS regression to estimate λ and coefficients on the other covariates without including teacher indicators in the regressions. The student-level residuals from this regression are then averaged for each teacher to provide a measure of teacher effectiveness: hence we refer to this method as the average residual estimator (AR). Since AR does not partial out teacher assignment from the covariates, an assumption is made that teacher assignment is not correlated with the regressors – meaning essentially that students are randomly assigned to teachers – if the goal is to isolate the contribution of teachers from other important factors.[16]

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 Em-

[15]The proper computation of these standard errors is still an under researched topic and is beyond the scope of this paper. See Bibler, Guarino, Reckase, Vosters, & Wooldridge (2014) for an investigation of this issue. Moreover, it may be possible to compute a type of standard error using bootstrapping techniques for SGP and other models as well, but little work has been done in this area, and it beyond the scope of this paper.

[16]Ehlert, Koedel, Parsons, & Podgursky (2013) argue that methods such as AR, which control for student covariates but do not adjust for teacher assignment, may induce teachers to exert optimal effort and be preferable to methods that more accurately identify individual teachers' causal impact on learning. However, we maintain that identifying causal effects is an important policy goal and treat it as the object of interest in this paper.

pirical Bayes' approach, although the true Empirical Bayes' relies on GLS rather than OLS.[17] 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. Estimates for teachers with smaller class sizes will be shrunk more than those with larger class sizes. In our simulation, we only estimate 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 value-added models to actual administrative data, we examine the Empirical Bayes' estimator based on GLS, which we abbreviate as EB-Lag.[18]

2.3 Adjusting for Teacher Assignment

As discussed in Guarino, Reckase, & Wooldridge (2015) the decision not to include teacher indicators in the VAM 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 teacher effects is not partialled out of

[17] See Guarino, Maxfield, Reckase, Thompson, & Wooldridge (Accepted for Publication) for a complete derivation and explanation of the Empirical Bayes' estimator in its application to teacher evaluation. As desribed there, for mechanical reasons the EB estimator is often much closer to DOLS than is the AR estimator under nonrandom assignment.

[18] The results with the AR estimator are available upon request. The results for the shrunken AR estimator are very similar to the Empirical Bayes' estimator based on GLS. In our actual data, the correlation between the two estimators is .998, and the differences seen in the correlations and misclassification rates for teachers in schools with evidence of nonrandom grouping and teachers in schools with little evidence of nonrandom grouping discussed below for the Empirical Bayes' estimator are very similar for the shrunken AR estimator.

the effect estimates, such as in the case of the estimators that average the residuals. A type of omitted variable bias can affect 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.

Also, under nonrandom assignment of teachers to students, it may no longer be possible to attribute high performance in rankings produced by the SGP estimator to good teaching. To illustrate the reason, consider a case 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, $A_{1,g-1} = A_{2,g-1}$ and classroom 3 and classroom 4 have identical initial achievement, $A_{3,g-1} = A_{4,g-1}$.

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

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

In the SGP approach, students are compared with 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 with one another. Students in classrooms 3 and 4 will be compared with 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 at the end of the year. In this case, the median or mean 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 SGP 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 will also be 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. While this problem could be potentially addressed by including teacher indicators in the quantile regressions, in practice, including these variables can make the estimation procedure very computationally intensive, and quick techniques such as demeaning data do not have theoretical justification in quantile regression. This makes the problem of nonrandom assignment of teachers to students difficult to address in SGP approaches.

3 Simulation

3.1 Data Generating Process

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:

$$A_{i3} = \lambda A_{i2} + \beta_{i3} + c_i + u_{i3}, \tag{7}$$

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

achievement gains over the course of the year, respectively.[19]

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)
- 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, with some noise included, 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

[19]These relative effect sizes are based on prior research (e.g. Nye, Konstantopoulos, & Hedges (2004), McCaffrey, Lockwood, Koretz, Louis, & Hamilton (2004), and Lockwood, McCaffrey, Hamilton, Stecher, Le, & Martinez (2007)). We changed the relative effect sizes as sensitivity checks and found no substantive differences.

highest scores tend to be grouped together.[20]

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, $\lambda = .5$.

Simulations are performed using Stata. One hundred simulation replications are performed for each grouping-assignment-persistence rate combination.

3.2 Simulation Results from Main Analysis

Table 1 displays Spearman rank correlations of the estimated teacher effects with the true teacher effects for each estimator under each grouping and assignment scenario. In addition, we present a measure of misclassification. The measure we choose is the percentage of teachers who have a true teacher effect above the 25th

[20]The amount of noise built into the assignment process yields patterns of variance that are consistent with what is found in Aaronson, Barrow, & Sander (2007) in their investigation using real data from Chicago public schools. The authors compare the average standard deviation of prior year achievement within classrooms in their data with a simulated average standard deviation in the case in which students are randomly assigned to classrooms. The authors find that the average standard deviation under random assignment is roughly 1.2, while the average standard deviation in their data is roughly 1. In our simulations, the average standard deviation of prior year achievement within classrooms under random assignment is 1, while the average standard deviation under nonrandom assignment is .75. Of course, the degree of sorting can greatly vary from school to school, as found in Dieterle, Guarino, Reckase, & Wooldridge (2015, published online July 2014). The degree of sorting introduced into this simulation may actually understate the amount of sorting in a nontrivial number of schools.

percentile but who are rated in the bottom 25% using the estimated 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 1 and for λ set to .5 are similar. Both VAMs outperform the SGP models, but these latter models still perform reasonably well, with rank correlations of around .82 and .87. Misclassification rates are around 8% for all estimators, with the exception of DOLS at 6% in one scenario and the SGP-Median estimator, with a slightly larger misclassification rate of 10% in the case of the true vertically scaled scores and 9% in the case of the standardized scores.

In the case of dynamic grouping coupled with random assignment of groups to teachers (DG-RA) the results are quite similar to those for the RG-RA scenario. The rank correlations for DOLS and AR drop only slightly to .87, stay the same for the SGP-Mean, and increase one percentage point for the SGP-Median. The misclassificaiton rates are fairly stable as well.

Once assignment of teachers to students is nonrandom the patterns 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 maintains a rank correlation of .88, whereas the rank correlations of the SGP-Median and SGP-Mean estimators fall to .71 and .76 respectively in the $\lambda = 1$ and $\lambda = .5$ cases. The rank correlation also decreases markedly for AR, which, like the SGPs, fails to properly partial out the relationship between teacher and student

quality. The misclassification rates show a similar pattern. DOLS does the best in terms of misclassification, while the SGP-Median and SGP-Mean estimators have misclassification rates that increase roughly 2-3 percentage points compared with 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, SGP-Median, and SGP-Mean estimators because it partials out the relationship between teacher and student quality.

3.3 Simulation Results for Sensitivity Analyses

Our simulations are intentionally simplified to highlight the behavior of the estimators we consider under conditions of random and nonrandom assignment. For example, for simplicity, we have thus far utilized a data-generating process in which the relationship between current and prior scores is linear. To relax this restriction and test the sensitivity of our results to this assumption, we also generated test score data in which a squared prior year test score term is included in the data-generating process. In this case, the SGP estimators can outperform DOLS when the generated coefficient on the squared term is set to an implausibly large number such as 1 (results are not reported but are available upon request). However, the difference is driven simply by the functional form in which prior test scores enter the model. Once the DOLS specification is rendered more flexible by including a polynomial or B-spline function of prior test scores, it regains its status as the best estimator across random and nonrandom assignment scenarios

in the simulation.

Moreover, it is unlikely that nonlinear relationships between current and prior scores are a significant feature of real data. In our actual data, we find that the coefficient on the squared term in a regression of math achievement on math prior achievement, prior achievement squared, other student demographics, and teacher indicators is estimated to be around .05 rather than a large number like 1. Moreover, when we compare estimated teacher effects from the basic DOLS estimator with those from a DOLS estimator that includes the square and cube of prior year test scores in the real data, we find that the estimates are very highly correlated (around .99), signaling that nonlinearities do not exist in actuality in a way that impacts teacher quality measures.

Also in an effort to keep the simulation simple, we did not include other demographic characteristics of students, such as, say, ELL status, in the test-score data-generating process, nor did we sort students on the basis of these characteristics in our simulation. However, it would logically follow that a DOLS type estimator that also controls for such observables will outperform the other estimators under nonrandom assignment based on these characteristics, since neither the SGP nor the AR models control for teacher assignment and the SGP estimators, in particular, typically omit student demographics as well.

One claim that could be made about the SGP approaches 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 SGP model may perform better than the estimators based

on OLS. As a further 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 2. Only results using the vertically scaled test scores are reported. Under random grouping and random assignment (RG-RA) the SGP-Median and particularly the SGP-Mean estimators outperform the value-added estimators. The SGP-Median estimator has a rank correlation of .72, and the SGP-Mean estimator has a rank correlation of .79 in the $\lambda = 1$ case. The value-added estimators have a slightly lower rank correlation of .71.

Under the dynamic grouping and nonrandom assignment (DG-PA and DG-NA) scenarios, however, DOLS again outperforms the SGP estimators, which do not properly partial out the relationship between the covariates and the teacher's value-added. The rank correlation for DOLS remains relatively stable at .70 and .71 for the DG-PA and DG-NA scenarios in the $\lambda = 1$ case. The rank correlation for the SGP-Median estimator drops to .57 and .59 for the DG-PA and DG-NA scenarios, and the rank correlation drops to .66 and .66 for the SGP-Mean estimator.

An important takeaway from this analysis is that there may be cases in which using the SGP 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 SGP estimators.

4 Empirical Analysis

We now examine the correlations between the estimators using real data. A main finding from the simulations was that the DOLS estimator and the SGP estimator provide similar rankings under random assignment, but somewhat different rankings under nonrandom assignment. Using the real data, we find patterns suggesting a similar relationship between DOLS and the SGP estimators 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.

4.1 Data

We use administrative data from a 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 provided. 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 focuses on effectiveness estimates 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 3 and 4.

4.2 Analysis of Administrative Data

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 SGP estimator performed less well. In order to examine whether the SGP model may perform less well in actual data, we performed the test of nonrandom grouping that was developed in Dieterle, Guarino, Reckase, & Wooldridge (2015, published online July 2014). The test involves running a student-level multinomial logit regression of classroom assignment on prior year test scores and other observables 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 focus on teachers in schools that reject the test of random grouping and compare them with teachers

in schools that fail to reject.[21]

We report Spearman rank correlations across estimators as well as two misclassifcation measures, which are the fraction of teachers rated in the bottom (or top) 25% in one estimator not rated in the bottom (or top) 25% in the other estimators in tables 5 through 13. Although our main focus is the comparison between the DOLS and SGP estimators, we also include statistics for the Empirical Bayes' estimator in our comparison.[22]

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 SGP estimators are computed using one year of data[23]. DOLS and the EB-Lag estimator include two prior years mathematics scores as controls. In the EB-Lag estimator, the student's class average prior year test score is included as a control for peer effects. The SGP 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)[24].

[21]The data contain 314 unique schools with 1,683 unique school-year-grade observations. Out of the 1683 school-year-grade observations, there is evidence of nonrandom grouping in 1,032 (61.32%).

[22]In this analysis, 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.

[23]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.

[24]As a sensitivity check we also estimate the SGP 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 still hold

The correlation between DOLS and the SGP-Median using one year of data[25] is .808 in the real data. The correlation between DOLS and the SGP-Mean estimator is .833. These correlations are appreciably smaller than the correlation between DOLS and the EB-Lag (.953) estimator.

Based on the simulations, we expect that the correlations will increase and the misclassification rates will be lower in schools where there is little or no evidence of nonrandom grouping, and this is what we find. When the sample of teachers is broken into those teachers in school-grade-years in which we find evidence of nonrandom grouping and those in school-grade-years in which we do not, we see a pattern that accords with the pattern seen in the simulation. In the simulation, the correlation between the DOLS estimator and the SGP-Median estimator went from around .87 under nonrandom assignment to .93 under random assignment as is visible in the final column of Tables 1. The correlation between DOLS and the SGP-Median estimator, found in table 7, is .802 in school-grade-years with nonrandom grouping and .818 in school-grade-years where we can't reject the hypothesis of random grouping, found in table 8. The correlations between DOLS and the EB-Lag estimator changes from .945 to .967 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% and top 25% of teachers. We calculate the fraction of teachers rated in the bottom 25% (or top 25%) using one

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

estimator that are not rated in the bottom 25% (or top 25%) using the other estimators. Results are reported in tables 9 to 14. Similar to the pattern indicated by the rank correlations, there is less disagreement between the estimators in the cases of schools with little evidence of nonrandom grouping. The fraction of teachers rated in the bottom 25% using the DOLS estimator not rated in the bottom 25% using the SGP-Median estimator is .3 in nonrandom grouping schools and .277 in random grouping schools.

5 Conclusions

In this paper, we compare commonly used value-added estimators to two SGP models: one based on the median student growth percentile for students assigned to the teacher and the other based on the mean.

Simulation evidence indicates that the relative 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 SGP 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.

A key finding is that DOLS, because it controls for teacher assignment in estimating the parameters on the other covariates, including lagged test scores, outperforms not only SGP estimators but also value-added models that do not that

do not control for teacher assignment when assignment is nonrandom.

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 in our simulations, DOLS and the other value-added estimators perform worse than the SGP estimators in the random assignment scenario, but only slightly so. DOLS still outperforms the SGP estimators in the nonrandom grouping and assignment scenarios.

Additionally, we compare the estimators using actual data. In accordance with the predictions of the simulation analysis, we detect stronger patterns of divergence between the DOLS and SGP estimates in for teachers in school contexts that exhibit evidence of nonrandom grouping than for teachers in school contexts in which grouping is fairly random.

This paper provides evidence that nonrandom grouping and assignment can negatively affect the popular SGP modeling approaches. Care should be taken by practitioners and researchers in evaluating teachers using these approaches when nonrandom grouping and assignment occurs in the school system. More generally, estimators that partial out teacher effects are better equipped to disentangle teacher contributions to student achievement from other factors affecting achievement than estimators that do not adjust for teacher assignment, whether they be SGPs or VAMs.

References

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

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, *4*(4), 351–383.

Barlevy, G., & Neal, D. (2011). Pay for percentile. *National Bureau of Economic Research*.

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42–51.

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

Betebenner, D. W. (2012). Growth, standards, and accountability. *GJ Cizek, Setting Performance Standards: Foundations, Methods & Innovations*, (pp. 439–450).

Bibler, A. J., Guarino, C. M., Reckase, M. D., Vosters, K. N., & Wooldridge, J. M. (2014). Precision for policy: Calculating standard errors in value-added models. *Unpublished Draft*.

Briggs, D. C., & Weeks, J. P. (2009). The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4*(4), 384–414.

Dieterle, S. G., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015, published online July 2014). How do principals assign students to teachers? finding evidence in administrative data and the implications for value-added. *Journal of Policy Analysis and Management*.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2013). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in missouri. *Statistics and Public Policy*, *1*(1), 19–27.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2012). Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. *National Bureau of Economic Research*.

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

Guarino, C. M., Maxfield, M., Reckase, M. D., Thompson, P., & Wooldridge, J. M. (Accepted for Publication). An evaluation of empirical bayes estimation of value-added teacher performance measures. *Journal of Educational and Behavioural Statistics*.

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

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

Harris, D., Sass, T., & Semykina, A. (2011). Value-added models and the measurement of teacher productivity.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? evidence on subjective performance evaluation in education. *Journal of Labor Economics*, *26*(1), 101–136.

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

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Tech. rep., National Bureau of Economic Research.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *The Journal of Economic Perspectives*.

Lockwood, J., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to

different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

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

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

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational evaluation and policy analysis*, *26*(3), 237–257.

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

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

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

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). Sas® evaas® statistical models. *SAS White Paper*.

Tables and Figures

Table 1: Rank Correlations and Misclassification Measures Across Estimators with Simulated Data Generated with Normal(0,1) Errors. Results from 100 replications. Row 1: Average rank correlation Row 2: Percentage of teachers above bottom 25% in true effect misclassified in bottom 25%

Estimator	**DOLS**	**AR**	**SGP-Median**	**SGP-Mean**	**Corr DOLS/ SGP-Median**
Assign Mech			$\lambda = 1$		
RG-RA	0.88	0.88	0.82	.87	.93
	6%	8%	10%	8%	
DG-RA	0.87	0.87	0.83	.87	.93
	7%	8%	9%	8%	
DG-PA	0.88	0.78	0.71	.76	.87
	7%	11%	12%	11%	
DG-NA	0.87	0.77	0.71	.76	.87
	8%	10%	12%	10%	
			$\lambda = .5$		
RG-RA	0.88	0.88	0.82	.87	.93
	8%	8%	10%	8%	
DG-RA	0.87	0.87	0.83	.87	.93
	7%	8%	9%	8%	
DG-PA	0.88	0.78	0.71	.76	.87
	8%	11%	12%	11%	
DG-NA	0.87	0.77	0.71	.76	.87
	7%	10%	12%	10%	

Table 2: Rank Correlations and Misclassification Measures Across Estimators with Simulated Data Generated with t(3) Errors. Results from 100 replications. Row 1: Average rank correlation Row 2: Percentage of teachers above bottom 25% in true effect misclassified in bottom 25%

Estimator	**DOLS**	**AR**	**SGP-Median**	**SGP-Mean**	**Corr DOLS/ SGP-Median**
Assign Mech			$\lambda = 1$		
RG-RA	0.71	0.71	0.72	0.79	.85
	10%	10%	12%	10%	
DG-RA	0.72	0.72	0.71	0.78	.84
	11%	11%	12%	11%	
DG-PA	0.70	0.58	0.57	0.66	.80
	11%	13%	14%	13%	
DG-NA	0.71	0.59	0.59	0.66	.80
	11%	14%	15%	14%	
			$\lambda = .5$		
RG-RA	0.71	0.71	0.71	0.79	.85
	9%	9%	12%	10%	
DG-RA	0.72	0.72	0.71	0.78	.84
	11%	11%	12%	11%	
DG-PA	0.70	0.58	0.57	0.66	.80
	11%	13%	14%	13%	
DG-NA	0.71	0.59	0.59	0.66	.80
	11%	14%	15%	14%	

Table 3: Summary Statistics for Administrative Data in 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 4: Summary Statistics for Administrative Data in 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 5: Spearman Rank Correlations across Estimators using Administrative Data

Variables	DOLS	EB-Lag	SGP-Median	SGP-Mean
DOLS	1.000			
EB-Lag	0.953	1.000		
SGP-Median	0.808	0.769	1.000	
SGP-Mean	0.833	0.790	0.975	1.000
Teacher/Year Obs	5661	5661	5661	5661

Table 6: Spearman Rank Correlations across Estimators using Administrative Data - Nonrandom Grouping Schools

Variables	DOLS	EB-Lag	SGP-Median	SGP-Mean
DOLS	1.000			
EB-Lag	0.945	1.000		
SGP-Median	0.802	0.759	1.000	
SGP-Mean	0.828	0.781	0.974	1.000
Teacher/Year Obs	3674	3674	3674	3674

Table 7: Spearman Rank Correlations across Estimators using Administrative Data - Random Grouping Schools

Variables	DOLS	EB-Lag	SGP-Median	SGP-Mean
DOLS	1.000			
EB-Lag	0.967	1.000		
SGP-Median	0.818	0.785	1.000	
SGP-Mean	0.841	0.806	0.977	1.000
Teacher/Year Obs	1991	1991	1991	1991

Table 8: Fraction of Teachers Rated in Bottom 25% in the Initial Estimator Who are Not Rated in Bottom 25% in Another Estimator

		Not Rated Bottom 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.146	0		
Bottom	SGP-Median	.291	.328	0	
25%	SGP-Mean	.273	.311	.102	0
	Teacher/Year Obs	5661	5661	5661	5661

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

		Not Rated Bottom 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.156	0		
Bottom	SGP-Median	.3	.337	0	
25%	SGP-Mean	.276	.316	.1	0
	Teacher/Year Obs	3674	3674	3674	3674

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

		Not Rated Bottom 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.127	0		
Bottom	SGP-Median	.277	.309	0	
25%	SGP-Mean	.265	.297	.108	0
	Teacher/Year Obs	1991	1991	1991	1991

Table 11: Fraction of Teachers Rated in Top 25% in the Initial Estimator Who are Not Rated in Top 25% in Another Estimator

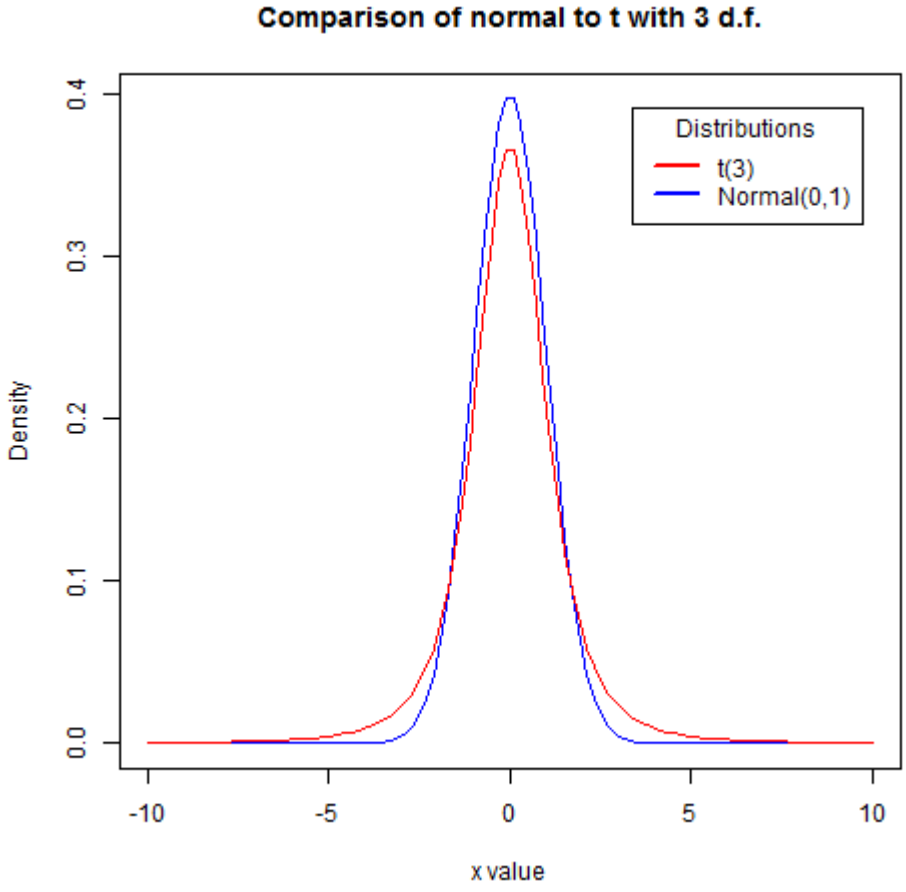
		Not Rated Top 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.148			
Top	SGP-Median	.288	.349	0	
25%	SGP-Mean	.271	.344	.086	0
	Teacher/Year Obs	5661	5661	5661	5661

Table 12: Fraction of Teachers Rated in Top 25% in the Initial Estimator Who are Not Rated in Top 25% in Another Estimator - Nonrandom Grouping Schools

		Not Rated Top 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.167	0		
Top	SGP-Median	.292	.362	0	
25%	SGP-Mean	.275	.356	.093	0
	Teacher/Year Obs	3674	3674	3674	3674

Table 13: Fraction of Teachers Rated in Top 25% in the Initial Estimator Who are Not Rated in Top 25% in Another Estimator - Random Grouping Schools

		Not Rated Top 25%			
	Initial Estimator	DOLS	EB-Lag	SGP-Median	SGP-Mean
	DOLS	0			
Rated	EB-Lag	.109			
Top	SGP-Median	.286	.336	0	
25%	SGP-Mean	.27	.32	.095	0
	Teacher/Year Obs	1991	1991	1991	1991


Comparison of normal to t with 3 d.f.
Distributions
t(3)
Normal(0,1)
Density
0.0
0.1
0.2
0.3
0.4
-10
-5
0
5
10
x value