An Evaluation of Empirical Bayes’ Estimation of Value-Added Teacher Performance Measures

Cassandra M. Guarino, Indiana University
Michelle Maxfield, Michigan State University
Mark D. Reckase, Michigan State University
Paul Thompson, Michigan State University
Jeffrey M. Wooldridge, Michigan State University

December 12, 2012
Revised: February 28, 2014

The content of this paper does not necessarily reflect the views of The Education Policy Center or Michigan State University
An Evaluation of Empirical Bayes’ Estimation of Value-Added Teacher Performance Measures

Author Information
Cassandra M. Guarino, Indiana University
Michelle Maxfield, Michigan State University
Mark D. Reckase, Michigan State University
Paul Thompson, Michigan State University
Jeffrey M. Wooldridge, Michigan State University

The work here was supported by IES Statistical Research and Methodology grant #R305D10028 and in part by a Pre-Doctoral Training Grant from the IES, U.S. Department of Education (Award # R305B090011) to Michigan State University. The opinions expressed here are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Abstract
Empirical Bayes’ (EB) estimation is a widely used procedure to calculate teacher value-added. It is primarily viewed as a way to make imprecise estimates more reliable. In this paper we review the theory of EB estimation and use simulated data to study its ability to properly rank teachers. We compare the performance of EB estimators with that of other widely used value-added estimators under different teacher assignment scenarios. We find that, although EB estimators generally perform well under random assignment of teachers to classrooms, their performance generally suffers under non-random teacher assignment. Under nonrandom assignment, estimators that explicitly (if imperfectly) control for the teacher assignment mechanism perform the best out of all the estimators we examine. We also find that shrinking the estimates, as in EB estimation, does not itself substantially boost performance.
An Evaluation of Empirical Bayes’ Estimation of Value-Added Teacher Performance Measures

February 28, 2014

Abstract: Empirical Bayes’ (EB) estimation is a widely used procedure to calculate teacher value-added. It is primarily viewed as a way to make imprecise estimates more reliable. In this paper we review the theory of EB estimation and use simulated data to study its ability to properly rank teachers. We compare the performance of EB estimators with that of other widely used value-added estimators under different teacher assignment scenarios. We find that, although EB estimators generally perform well under random assignment of teachers to classrooms, their performance generally suffers under nonrandom teacher assignment. Under nonrandom assignment, estimators that explicitly (if imperfectly) control for the teacher assignment mechanism perform the best out of all the estimators we examine. We also find that shrinking the estimates, as in EB estimation, does not itself substantially boost performance.
1 Introduction

Empirical Bayes’ (EB) estimation of teacher effects has gained recent popularity in the value-added research community (see, for example, McCaffrey et al. 2004; Kane and Staiger 2008; Chetty, Friedman, and Rockoff 2011; Corcoran, Jennings, and Beveridge 2011; and Jacob and Lefgren 2005, 2008). Researchers motivate the use of EB estimation as a way to decrease classification error of teachers, especially when limited data is available to compute value-added estimates. When there are only a small number of students per teacher, teacher value-added estimates can be very noisy. EB estimates of teacher value-added reduce the variability of the estimates by shrinking them toward the average estimated teacher effect in the sample and, therefore, are often referred to as “shrinkage estimators.” As the degree of shrinkage depends on class size, estimates for teachers with smaller class sizes are more affected, potentially helping with the misclassification of these teachers. In addition, EB estimation may be less computationally demanding than methods that view the teacher effects as fixed parameters to estimate.

Despite the potential shrinkage benefits of EB estimation, the estimated teacher effects can suffer from severe bias under nonrandom teacher assignment. By treating the teacher effects as random, EB estimation assumes that teacher assignment is uncorrelated with factors that predict student achievement – including observed factors such as past test scores. While the bias (technically, the inconsistency) disappears as the number of students per teacher increases – because the EB estimates converges to the so-called fixed effects estimates – the bias still can be important for the kinds of data used to estimate teacher VAMs. This is because the EB estimators of the coefficients on the covariates in the model are inconsistent for fixed class sizes as the number of classrooms grows. By contrast, estimators that include the teacher assignment indicators along with the covariates in a multiple regression analysis are consistent (as the number of classrooms grows) for the coefficients on the covariates. This
generally leads to less bias in the estimated teacher VAMs under nonrandom assignment without many students per teacher.

In this paper we address the following research questions: (1) How does the performance of EB estimators compare with that of estimators that treat the teacher effect as fixed under random teacher assignment and various nonrandom assignment scenarios? (2) Are there cases where it is beneficial to use an EB-type approach to shrink estimates of teacher fixed effects? (3) How do recently proposed simplified versions of EB perform?

2 Empirical Bayes’ Estimation

There are several ways to derive empirical Bayes’ estimators of teacher value added. We adopt a so-called “mixed estimation” (ME) approach, as in Ballou, Sanders, and Wright (2004), because it is fairly straightforward and does not require delving into Bayesian estimation methods. Our focus is on estimating teacher effects grade by grade. Therefore, we assume either that we have a single cross section or multiple cohorts of students for each teacher. We do not include cohort effects; multiple cohorts are allowed by pooling students across cohorts for each teacher.

Let \( y_i \) denote a measure of achievement for student \( i \), randomly drawn from the population. This measure could be a test score or a gain score. Suppose there are \( G \) teachers and the teacher effects are \( b_g, \ g = 1, ..., G \). In the mixed effects setting, these are treated as random variables as opposed to fixed population parameters. Viewing the \( b_g \) as random variables independent of other observable factors affecting test scores has consequences for the properties of EB estimators.

Typically VAMs are estimated controlling for other factors, which we denote by a row vector \( x_i \). These factors include student demographics and, in some cases, prior test scores. We assume the coefficients on these covariates are fixed parameters. We can write a mixed
effects linear model as
\[ y_i = x_i \gamma + z_i b + u_i, \]  
(1)

where \( z_i \) is a \( 1 \times G \) row vector of teacher assignment dummies, \( b \) is the \( G \times 1 \) vector of teacher effects, and \( u_i \) contains the unobserved student-specific effects. Because a student is assigned to one and only one teacher, \( z_{i1} + z_{i2} + \cdots + z_{iG} = 1 \). Equation (1) is an example of a “mixed model” because it includes the usual fixed population parameters \( \gamma \) and the random coefficients \( b \). Even if there are no covariates, \( x_i \) typically includes an intercept. If \( x_i \gamma \) is only a constant, so \( x_i \gamma = \gamma \), then \( \gamma \) is the average teacher effect and we can then take \( E(b) = 0 \). This means that \( b_g \) is the effect of teacher \( g \) net of the overall mean teacher effect.

Equation (1) is written for a particular student \( i \) so that teacher assignment is determined by the vector \( z_i \). A standard assumption is that, conditional on \( b \), (1) represents a linear conditional mean:
\[ E(y_i|x_i, z_i, b) = x_i \gamma + z_i b \]  
(2)

which follows from
\[ E(u_i|x_i, z_i, b) = 0. \]  
(3)

An important implication of (3) is that \( u_i \) is uncorrelated with \( z_i \), so that teacher assignment is not systematically related to unobserved student characteristics once we have controlled for the observed factors in \( x_i \).

If we assume a sample of \( N \) students assigned to one of \( G \) teachers we can write (1) in matrix notation as
\[ y = X\gamma + Zb + u, \]  
(4)

where \( y \) and \( u \) are \( N \times 1 \), \( X \) is \( N \times K \), and \( Z \) is \( N \times G \). In order to obtain the best linear unbiased estimator (BLUE) of \( \gamma \) and the best linear unbiased predictor (BLUP) of \( b \), we
assume that the covariates and teacher assignments satisfy a strict exogeneity assumption:

\[ E(u_i|X, Z, b) = 0, \ i = 1, ..., N. \] (5)

An implication of assumption (5) is that inputs and teacher assignment of other students does not affect the outcome of student \( i \).

Given assumption (5) we can write the conditional expectation of \( y \) as

\[ E(y|X, Z, b) = X\gamma + Zb \] (6)

In the EB literature a standard assumption is

\[ b \text{ is independent of } (X, Z), \] (7)

in which case

\[ E(y|X, Z) = X\gamma + ZE(b|X, Z) = X\gamma = E(y|X) \] (8)

because \( E(b|X, Z) = E(b) = 0 \). Assumption (7) has the implication that teacher assignment for student \( i \) does not depend on the quality of the teacher (as measured by the \( b_g \)).

From an econometric perspective, equation (8) means that \( \gamma \) can be estimated in an unbiased way by OLS regression of

\[ y_i \text{ on } x_i, \ i = 1, ..., N. \] (9)

Consequently, we can estimate the effects of the covariates \( x_i \) by omitting the teacher assignment dummies. Practically, this means we are assuming teacher assignment is uncorrelated with the covariates \( x_i \).

Under (5) and (7), the OLS estimator of \( \gamma \) is unbiased and consistent, but it is inefficient
if we impose the standard classical linear model assumptions on \( u \). In particular, if

\[
\text{Var}(u|X, Z, b) = \text{Var}(u) = \sigma^2 I_N
\]  

(10)

then

\[
\text{Var}(y|X, Z) = E[(Zb + u)(Zb + u)'|X, Z] = Z\text{Var}(b)Z' + \text{Var}(u) = \sigma_b^2 ZZ' + \sigma_u^2 I_N,
\]

where we also add the standard assumption that the elements of \( b \) are uncorrelated

\[
\text{Var}(b) = \sigma_b^2 I_G, \tag{11}
\]

and \( \sigma_b^2 \) is the variance of the teacher effects, \( b_g \).

Under the assumption that \( \sigma_b^2 \) and \( \sigma_u^2 \) are known – actually, it suffices to know their ratio – the BLUE of \( \gamma \) under the preceding assumptions is the generalized least squares (GLS) estimator,

\[
\gamma^* = [X'(\sigma_b^2 ZZ' + \sigma_u^2 I_N)^{-1}X]^{-1}X'(\sigma_b^2 ZZ' + \sigma_u^2 I_N)^{-1}y. \tag{12}
\]

The \( N \times N \) matrix \( ZZ' \) is a block diagonal matrix with \( G \) blocks, where block \( g \) is an \( N_g \times N_g \) matrix of ones and \( N_g \) is the number of students taught by teacher \( g \). The GLS estimator \( \gamma^* \) is the well-known “random effects” (RE) estimator popular from panel data and cluster sample analysis. Note that the “random effects” in this case are teacher effects, not student-specific effects.

Before we discuss \( \gamma^* \) further, as well as estimation of \( b \), it is helpful to write down the mixed effects model in perhaps a more common form. After students have been designated to classrooms, we can write \( y_{gi} \) as the outcome for student \( i \) in class \( g \), and similarly for \( x_{gi} \)
and \( u_{gi} \). Then, for classroom \( g \), we have
\[
y_{gi} = x_{gi}\gamma + b_g + u_{gi} \equiv x_{gi}\gamma + r_{gi}, \ i = 1, \ldots, N_g, \quad (13)
\]
where \( r_{gi} \equiv b_g + u_{gi} \) is the composite error term. Equation (13) makes it easy to see that the BLUE of \( \gamma \) is the random effects estimator. It also highlights the assumption that \( b_g \) is independent of the covariates \( x_{gi} \). Further, the assumption \( E(u_{gi}|X_g, b_g) = 0 \) implies that covariates from student \( h \) do not affect the outcome of student \( i \). We can also see that OLS pooled across \( i \) and \( g \) is unbiased for \( \gamma \) because we are assuming \( E(b_g|X_g) = 0 \).

As shown in, say, BSW, the BLUP of \( b \) under assumptions (5), (7), and (10) is
\[
b^* = (Z'Z + \rho I_G)^{-1}Z'(y - X\gamma^*) \equiv (Z'Z + \rho I_G)^{-1}Z'r^*, \quad (14)
\]
where \( \rho = \sigma_u^2/\sigma_b^2 \) and \( r^* = y - X\gamma^* \) is the vector of residuals. Straightforward matrix algebra shows each \( b_g^* \) can be expressed as
\[
b_g^* = (N_g + \rho)^{-1} \sum_{i=1}^{N_g} r_{gi}^* = \left( \frac{N_g}{N_g + \rho} \right) \bar{r}_g^* = \left( \frac{\sigma_b^2}{\sigma_b^2 + (\sigma_u^2/N_g)} \right) \left( \bar{y}_g - \bar{x}_g\gamma^* \right), \quad (15)
\]
where
\[
\bar{r}_g^* = N_g^{-1} \sum_{i=1}^{N_g} r_{gi}^* = \bar{y}_g - \bar{x}_g\gamma^* \quad (16)
\]
is the average of the residuals \( r_{gi}^* = y_{gi} - x_{gi}\gamma^* \) within classroom \( g \).

To operationalize \( \gamma^* \) and \( b_g^* \) we must replace \( \sigma_b^2 \) and \( \sigma_u^2 \) with estimates. There are different ways to obtain estimates depending on whether one uses OLS residuals after an initial estimation or a joint estimation method. With the composite error defined as \( r_{gi} = b_g + u_{gi} \) we can write \( \sigma_r^2 = \sigma_b^2 + \sigma_u^2 \). An estimator of \( \sigma_r^2 \) can be obtained from the usual sum of
squared residuals from the OLS regression

\[ y_{gi} \text{ on } x_{gi}, \ i = 1, ..., N_g, \ g = 1, ..., G. \] (17)

Call the residuals \( \tilde{r}_{gi} \). Then a consistent estimator is

\[ \tilde{\sigma}^2_r = \frac{1}{(N - K)} \sum_{g=1}^{G} \sum_{i=1}^{N_g} \tilde{r}_{gi}^2, \] (18)

which is just the usual degrees-of-freedom (df) adjusted error variance estimator from OLS.

To estimate \( \sigma^2_u \), write

\[ r_{gi} - \bar{r}_g = u_{gi} - \bar{u}_g, \]

where \( \bar{r}_g \) is the within-teacher average, and similarly for \( \bar{u}_g \). A standard result on demeaning a set of uncorrelated random variables with the same variance gives \( Var(u_{gi} - \bar{u}_g) = \sigma^2_u (1 - N_g^{-1}) \) and so, for each \( g \), \( E \left[ \sum_{i=1}^{N_g} (r_{gi} - \bar{r}_g)^2 \right] = \sigma^2_u (N_g - 1) \). When we sum across teachers it follows that

\[ \frac{1}{(N - G)} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (r_{gi} - \bar{r}_g)^2 \] (19)

has expected value \( \sigma^2_u \). To turn (19) into an estimator we can replace the \( r_{gi} \) with the OLS residuals, as before, \( \tilde{r}_{gi} \), from the regression in (17). The estimator based on the OLS residuals is

\[ \tilde{\sigma}^2_u = \frac{1}{(N - G)} \sum_{g=1}^{G} \sum_{i=1}^{N_g} (\tilde{r}_{gi} - \bar{r}_g)^2. \] (20)

With fixed class sizes and \( G \) getting large, the estimator that uses \( N \) in place of \( N - G \) is not consistent. Therefore, we prefer the estimator in equation (20), as it should have less bias in applications where \( G/N \) is not small. With many students per teacher the difference should be minor. We could also use \( N - G - K \) as a further df adjustment, but subtracting off \( K \) does not affect the consistency.

7
Given $\tilde{\sigma}_r^2$ and $\tilde{\sigma}_u^2$ we can estimate $\sigma_b^2$ as

$$\sigma_b^2 = \tilde{\sigma}_r^2 - \tilde{\sigma}_u^2. \quad (21)$$

In any particular data set – especially if the data have been generated to violate the standard assumptions listed above – there is no guarantee that expression (21) is nonnegative. A simple solution to this problem (and one used in is software packages, such as Stata) is to set $\tilde{\sigma}_b^2 = 0$ whenever $\tilde{\sigma}_r^2 < \tilde{\sigma}_u^2$. In order to ensure this happens infrequently with multiple cohorts, we compute $\tilde{\sigma}_u^2$ by replacing $\tilde{r}_g$ with the average obtained for the particular cohort. This ensures that, for a given cohort, the terms $\sum_{i=1}^{N_g} (\tilde{r}_{gi} - \tilde{r}_g)^2$ are as small as possible. In theory, if there are no cohort effects we could use an overall cohort mean for $\tilde{r}_g$. But using cohort-specific means reduces the problem of negative $\tilde{\sigma}_b^2$ when the model is misspecified.

An appealing alternative is to estimate $\sigma_b^2$ and $\sigma_u^2$ jointly along with $\gamma$, using software that ensures nonnegativity of the variance estimates. The most common approach to doing so is to assume joint normality of the teacher effects, $b_g$, and the student effects, $u_{gi}$, across all $g$ and $i$ – along with the previous assumptions. One important point is that the resulting estimators are consistent even without the normality assumption; so, technically, we can think of them as “quasi-” maximum likelihood estimators. The maximum likelihood estimator of $\sigma_u^2$ has the same form as in equation (20), except the residuals are based on the MLE of $\gamma$ rather than the OLS estimator. A similar comment holds for the MLE of $\sigma_b^2$ (if we do not constrain it to be nonnegative). See, for example, Hsiao (2003, Section 3.3.3).

Unlike the GLS estimator of $\gamma$, the feasible GLS (FGLS) estimator is no longer unbiased (even under assumptions (5) and (7)), and so we must rely on asymptotic theory. In the current context, the estimator is consistent and asymptotically normal provided $G \to \infty$ with $N_g$ fixed. In practice, this means that the number of teachers, $G$, should be substantially larger than the number of students per teacher, $N_g$. Typically this is the case in VAM
studies, which are applied to large school districts or entire states and therefore include many teachers. Often the number of students per teacher is fewer than 100 with several hundred or even several thousand teachers.

When \( \gamma^* \) is replaced with the FGLS estimator and the variances \( \sigma_b^2 \) and \( \sigma_u^2 \) are replaced with estimators, the EB estimator of \( b \) is no longer a BLUP. Nevertheless, we use the same formula as in (15) for operationalizing the BLUPs. Conveniently, certain statistical packages – such as Stata with its “xtmixed” command – allow one to recover the operationalized BLUPs after maximum likelihood estimation. When we use the (quasi-) MLEs to obtain the \( b^*_g \) we obtain what are typically called the empirical Bayes’ estimates.

One way to understand the shrinkage nature of \( b^*_g \) is to compare it with the estimator obtained by treating the teacher effects as fixed parameters. Let \( \hat{\gamma} \) and \( \hat{\beta} \) be the OLS estimators from the regression

\[
y_i \text{ on } x_i, \ z_i, \ i = 1, \ldots, N. \tag{22}
\]

Then \( \hat{\gamma} \) is the so-called “fixed effects” (FE) estimator obtained by a regression of \( y_i \) on the controls in \( x_i \) and the teacher assignment dummies in \( z_i \). In the context of the model

\[
y = X\gamma + Z\beta + u \tag{23}
\]

\[
E(u|X,Z) = 0, \ Var(u|X,Z) = \sigma_u^2 I_N,
\]

\( \hat{\gamma} \) is the BLUE of \( \gamma \) and \( \hat{\beta} \) is the BLUE of \( \beta \). As is well-known, \( \hat{\gamma} \) can be obtained by an OLS regression where \( y_{gi} \) and \( x_{gi} \) have been deviated from within-teacher averages (see, for example, Wooldridge 2010, Chapter 10). Further, the estimated teacher fixed effects can be obtained as

\[
\hat{\beta}_g = \bar{y}_g - \bar{x}_g \hat{\gamma}. \tag{24}
\]
Equation (24) makes computation of the teacher VAMs fairly efficient if one does not want to run the long regression in (22).

By comparing equations (15) and (24) we see that the EB estimator $b_g^*$ differs from the fixed effects estimator $\hat{\beta}_g$ in two ways. First, and most importantly, the RE estimator $\gamma^*$ is used in computing $b_g^*$ while $\hat{\beta}_g$ uses the FE estimator $\hat{\gamma}$. Second, $b_g^*$ shrinks the average of the residuals toward zero by the factor

$$\frac{\sigma_b^2}{\sigma_b^2 + (\sigma_u^2/N_g)} = \frac{1}{1 + (\rho/N_g)}$$  \hspace{1cm} (25)$$

where

$$\rho = \frac{\sigma_u^2}{\sigma_b^2}. \hspace{1cm} (26)$$

Equation (25) illustrates the well-known result that the smaller is the number of students for teacher $g$, $N_g$, the more the average residual is shrunk toward zero.

A well-known algebraic result – for example, Wooldridge (2010, Chapter 10) – that holds for any given number of teachers $G$ is that

$$\gamma^* \rightarrow \hat{\gamma} \text{ as } \rho \rightarrow 0 \text{ or } N_g \rightarrow \infty. \hspace{1cm} (27)$$

Equation (27) can be verified by noting that the RE estimator of $\gamma$ can be obtained from the pooled OLS regression

$$y_{gi} - \theta_g \bar{y}_g \text{ on } x_{gi} - \theta_g \bar{x}_g \hspace{1cm} (28)$$

where

$$\theta_g = 1 - \left( \frac{\sigma_u^2}{\sigma_b^2 + N_g \sigma_b^2} \right)^{1/2} = 1 - \left( \frac{1}{1 + (N_g/\rho)} \right)^{1/2}. \hspace{1cm} (29)$$

It is easily seen that $\theta_g \rightarrow 1$ as $\rho \rightarrow 0$ or $N_g \rightarrow \infty$. In other words, with many students per teacher or large teacher effects relative to student effects, the RE and FE estimates can be
very close. But they are never identical. Not coincidentally, the shrinkage factor in equation (25) tends to unity if \( \rho \to 0 \) or \( N_g \to \infty \). The bottom line is that with a “large” number of students per teacher the shrinkage estimates of the teacher effects can be close to the fixed effects estimates. The RE and FE estimates also tend to be similar when \( \sigma_u^2 \) (the student effect) is “small” relative to \( \sigma_b^2 \) (the teacher effect), but this scenario seems unlikely.

An important point that appears to go unnoticed in applying the shrinkage approach is that in situations where \( \gamma^* \) and \( \hat{\gamma} \) substantively differ, \( \gamma^* \) suffers from systematic bias because it assumes teacher assignment is uncorrelated with \( x_i \). Because \( \gamma^* \) is used in constructing the \( b_g^* \) in equation (15), the bias in \( \gamma^* \) generally results in biased teacher effects, and the teacher effects would be biased even if (15) did not employ a shrinkage factor. The shrinkage likely exacerbates the problem: the estimates are being shrunk toward values that are systematically biased for the teacher effects.\(^3\)

The expression in equation (15) motivates a common two-step alternative to the EB approach proper. In the first step of the procedure, one obtains \( \hat{\gamma} \) using the OLS regression in equation (17), and obtains the residuals, \( \tilde{r}_{gi} \). In the second step, one averages the residuals \( \tilde{r}_{gi} \) within each teacher to obtain the teacher effect for teacher \( g \). We call this approach the “average residual” (AR) method. After obtaining the averages of the residuals one can, in a third step, shrink the averages using the empirical Bayes’ shrinkage factors in equation (15). Typically the estimates in equations (18) and (20), based on the OLS residuals, are used in obtaining the shrinkage factors. We call the resulting estimator the “shrunken average residual” (SAR) method.

With or without shrinking, the AR approach suffers from systematic bias if teacher assignment, \( z_i \), is correlated with the covariates, \( x_i \). In effect, the AR approach partials \( x_i \) out of \( y_i \) but does not partial \( x_i \) out of \( z_i \), the latter of which is crucial if \( x_i \) and \( z_i \) are correlated. The so-called “fixed effects” regression in (22) partials \( x_i \) out of \( z_i \), which makes it a more reliable estimator under nonrandom teacher assignment – perhaps much
more reliable with strong forms of nonrandom assignment.

It is also important to know that the SAR approach is inferior to the EB approach under nonrandom assignment. The logic is simple. First, the algebraic relationship between RE and FE means that $\gamma^*$ tends to be closer to the FE estimator, $\hat{\gamma}$, than the OLS estimator, $\tilde{\gamma}$. Consequently, under nonrandom teacher assignment, the estimated teacher effects using the RE estimator of $\gamma$ will have less bias than the estimates that begin with OLS estimation of $\gamma$. Second, if teacher assignment is uncorrelated with the covariates, the OLS estimator of $\gamma$ is inefficient relative to the RE estimator under the standard random effects assumptions (because the RE estimator is FGLS). Thus, the only possible justification for SAR is computational simplicity. But the saving is likely to be minor unless the number of controls in $x_i$ is very large. For the kinds of data sets widely available, the computational saving from using SAR rather than EB is likely to be minor.

Before we leave this section we must emphasize that fixed effects estimation of the teacher VAMs allows any correlation between $z_i$ and $x_i$, and thus expect it to outperform EB estimation and strongly outperform SAR under nonrandom assignment. The bias due to nonrandom allocation of students to teachers is also discussed in Rothstein (2009, 2010).

3 Summary of Estimation Methods

In this paper we examine five different value-added estimators used to recover the teacher effects and apply them to both real and simulated data. Some of the estimators use EB or shrinkage techniques, while others do not. They can all be cast as a special case of the estimators described in the previous section. For clarity, we briefly describe each one, with additional reference to each of these specifications provided in Table 1.
The estimators can be obtained from a dynamic equation of the form

$$A_{it} = \lambda A_{i,t-1} + X_{it}\delta + Z_{it}\beta + v_{it},$$

(30)

in which $A_{it}$ is achievement (measured by a test score) for student $i$ in grade $t$, $X_{it}$ is a vector of student characteristics, and $Z_{it}$ is the vector of teacher assignment dummies. This is similar to equation (1) but with the lagged test score written separately from $X_{it}$ for clarity. Also, $X_{it}$ is omitted from the estimation of the teacher effects in the simulation analysis below, as student characteristics are not included in the data generating process.

The EB estimator we analyzed in Section 2 was for the case of a single cross-section of students. Thus, we use only one grade – fifth grade – for the analysis.

We first analyze EB LAG, a dynamic MLE version of the EB estimator that treats the teacher effects as random. This technique obtains the estimates of the teacher effects using the normal maximum likelihood in the first stage, regressing $A_{it}$ on its lag, $A_{i,t-1}$, and $X_{it}$. In the second stage, the shrinkage factor is applied to these teacher effects. A second estimator we consider is the average residual (AR) method described in Section 2. This technique mainly differs from EB LAG in that it uses OLS in the first stage. The residuals of this OLS regression are obtained, and then we average these residuals by classroom to calculate the estimated teacher effects. We expect the EB LAG estimator to outperform the AR estimator in most scenarios, given that MLE is being used in the first-stage instead of OLS.

We compare the estimators that treat the teacher effect as random with an estimator that explicitly controls for the teacher effect through the inclusion of teacher assignment dummy variables. This third estimator applies OLS to (30) by pooling across students and classrooms. We refer to this estimator as “dynamic OLS,” or DOLS. DOLS treats the teacher effects as fixed parameters to estimate. The inclusion of the lagged test score accounts for the possibility that teacher assignment is related to the past test score. Guarino, Reckase, and
Wooldridge (forthcoming) discuss the assumptions under which DOLS consistently estimates $\beta$ when (30) is obtained from a structural cumulative effects model, and the assumptions are quite restrictive. Nevertheless, their simulations show the DOLS estimator often estimates $\beta$ well even when the assumptions underlying the consistency of DOLS fails.

Given that EB estimation is often motivated as a way to increase precision and decrease misclassification, we also analyze whether shrinking AR and DOLS estimates enhances performance. Thus, the fourth estimator we analyze, SAR (for shrunken average residual), takes the AR estimates and shrinks them by the shrinkage factor described in Section 2 using the variance estimates from equations (18), (20), and (21). Shrinking the AR estimates does not result in a true EB estimator since AR uses OLS in the first stage, but it is commonly used as a simpler way of operationalizing the EB approach. As discussed in Section 2, with a sufficiently large number of students per teacher, the EB LAG estimator converges to the DOLS estimator but SAR does not. Thus, as the number of students per teacher grows, we would expect EB LAG to perform more similarly to DOLS than SAR. Finally, we consider a shrunken DOLS (SDOLS) estimator, which takes the DOLS estimated teacher fixed effects and shrinks them by the shrinkage factor derived in Section 2. Although SDOLS is rarely done in practice and is not a true EB estimator, we include it as an exploratory exercise in order to better determine the effects of shrinking itself when the number of students per teacher differs. When the class sizes are all the same, the SDOLS and DOLS estimates differ only by a constant positive multiple and shrinking the DOLS estimates will have no effect in terms of ranking teachers.

4 Comparing VAM Methods Using Simulated Data

The question of which VAM estimators perform the best can only be addressed in simulations in which the true teacher effects are known. Therefore, to evaluate the performance
of EB estimators relative to other common value-added estimators, we apply these methods to simulated data. This approach allows us to examine how well various estimators recover the true teacher effect under a variety of assignment scenarios. Using data generated from the processes described in Section 4.1, we apply the set of value-added estimators discussed in Section 3 and compare the resulting estimates with the true underlying teacher effects.

4.1 Simulation Design

Much of our main analysis focuses on a base case that restricts the data generating process to a relatively narrow set of idealized conditions. These ideal conditions do not allow for measurement error or peer effects and assume that teacher effects are constant over time. The data are constructed to represent grades three through five (the tested grades) in a hypothetical school. For simplicity and comparison with the theoretical predictions, we assume that the learning process has been going on for a few years but only calculate estimates of teacher effects for fifth grade teachers – a single cross section. We create data sets that contain students nested within teachers nested within schools, with students followed longitudinally over time in order to reflect the institutional structure of an elementary school. Our simple baseline data generating process is as follows:

\[
\begin{align*}
A_{i3} &= \lambda A_{i2} + \beta_{i3} + c_i + e_{i3} \\
A_{i4} &= \lambda A_{i3} + \beta_{i4} + c_i + e_{i4} \\
A_{i5} &= \lambda A_{i4} + \beta_{i5} + c_i + e_{i5}
\end{align*}
\]

in which \(A_{i2}\) is a baseline score reflecting the subject-specific knowledge of child \(i\) entering third grade, \(A_{it}\) is the grade-\(t\) test score (\(t = 3, 4, 5\)), \(\lambda\) is a time constant decay parameter, \(\beta_{it}\) is the teacher-specific contribution to growth (the true teacher value-added effect), \(c_i\) is a time-invariant student-specific effect (may be thought of as “ability” or “motivation”),
and $e_{it}$ is a random deviation for each student. In all of the simulations reported in this paper, the random variables $A_{i2}$, $c_i$, and $e_{it}$ are drawn from normal distributions with means of zero. The true teacher effect, $\beta_{it}$, is drawn from a normal distribution with a mean of 0.5. The standard deviation of the teacher effect is .25, the standard deviation of the student fixed effect is .5, and the standard deviation of the random noise component is 1. These give relative shares of 5, 19, and 76 percent of the total variance in gain scores (when $\lambda = 1$), respectively. Given that the student and noise components are larger than the teacher effects, we call these “small” teacher effects. The baseline score is drawn from a distribution with a standard deviation of 1. We also allow for correlation between the time-invariant child-specific heterogeneity, $c_i$, and the baseline test score, $A_{i2}$, which we set to 0.5. This correlation reflects that students with better unobserved “ability” likely have higher test scores as well.

Our data are simulated using 36 teachers and 720 students per cohort. In order to create a situation in which there is a substantial variation in class size – to showcase the potential disparities between EB/shrinkage and other estimators – we vary the number of students per classroom. Teachers receive classes of varying sizes, but receive the same number of students in each cohort. Of the 36 teachers we simulate, nine teachers have classes of 10 students, nine teachers have a class size of 20, and nine teachers have class sizes of 30. We simulate the data using both one and four cohorts of students to provide further variance in the amount of data from which the teacher effects are calculated. In the case of four cohorts, data are pooled across the cohorts so that value-added estimates are based off of sample sizes of 40, 80, and 120, instead of 10, 20, and 30 as in the one cohort case. Therefore, we would expect that estimates in the four cohort case to be less noisy than those from the one cohort case, possibly mitigating the potential gains from EB estimation.

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 level of
persistence in prior learning from one year to the next. We generate data using each of the
nine different mechanisms for the assignment of students outlined in Table A.1. Students are
grouped into classrooms either randomly, based on their prior year achievement level (dy-
namic grouping or DG), or based on their unobserved heterogeneity (heterogeneity grouping
or HG). In the random case, students are assigned a random number and then grouped into
classrooms of various sizes based on that random number. In the grouping cases, students are
ranked by either the prior test score or the student fixed effect and grouped into classrooms
of various sizes based on that ranking. Teachers are assigned to these classrooms either
randomly (denoted RA) or nonrandomly. Teachers assigned nonrandomly can be assigned
positively (denoted PA), meaning the best teachers are assigned to classrooms with the best
students, or negatively (denoted NA), meaning the best teachers are assigned to classrooms
with the worst students.

These grouping and assignment procedures are not purely deterministic, as we allow for
a random component with standard deviation of one in the assignment mechanism. We use the estimators discussed in Section 3, but with only a constant, teacher dummies (if applicable), and, for the dynamic specifications, the lagged test score included as covariates. We use 100 Monte Carlo replications per scenario in evaluating each estimator.

4.2 Evaluation Measures

For each estimator across each iteration, we save the individual estimated teacher ef-
fects and also retain the true teacher effects, which are fixed across the iterations for each
teacher. To study how well the methods uncover the true teacher effects, we adopt five
simple summary measures using the teacher level data. The first is a measure of how well
the estimates preserve the rankings of the true teacher effects. We compute the Spearman
rank correlation, $\hat{\rho}$, between the estimated teacher effects and the true effects and report
the average $\hat{\rho}$ across the 100 iterations. Second, we compute a measure of misclassification.
These misclassification rates are obtained as the percentage of above average teachers (in the true quality distribution) who are misclassified as below average in the distribution of estimated teacher effects for the given estimator.

In addition to examining rank correlations and misclassification rates, it is also helpful to have a measure that quantifies some notion of the magnitude of the bias in the estimates. Given that some teacher effects are biased upwards and others downwards, it is difficult to capture the overall bias in the estimates in a simple way. We create a statistic, $\hat{\theta}$, that captures how closely the magnitude of the deviation of the estimates from their mean tracks the size of the deviation of the true effects from the true mean. To calculate this measure, we regress the deviation of the estimated teacher effects from their overall estimated means on the analogous deviation of the true effects generated from the simulation for each estimator. We can represent this simple regression as

$$\hat{\beta}_j - \bar{\beta} = \hat{\theta}(\beta_j - \bar{\beta}) + \text{residual}_j,$$

in which $\hat{\beta}_j$ is the estimated teacher effect and $\beta_j$ is the true effect of teacher $j$. From this simple regression, we report the average coefficient, $\bar{\hat{\theta}}$, across the 100 replications of the simulation for each estimator. This regression tells us whether the estimated teacher effects are correctly distributed around the average teacher. If $\hat{\theta} = 1$, then a movement of $\beta_j$ away from its mean is tracked by the same movement of $\hat{\beta}_j$ from its mean.

When $\hat{\theta} \approx 1$, the magnitudes of the estimated teacher effects can be compared across teachers. If $\hat{\theta} > 1$, the estimated teacher effects amplify the true teacher effects. In other words, teachers above average will be estimated to be even more above average and vice versa for below average teachers. An estimation method that produces $\hat{\theta}$ substantially above one generally does a good job of ranking teachers, but the magnitudes of the differences in estimated teacher effects cannot be trusted. The magnitudes also cannot be trusted if $\hat{\theta} < 1$;
in this case, ranking the teachers becomes more difficult because the estimated effects are compressed relative to the true teacher effects.

In addition to ranking teachers correctly, the magnitude of the estimated teacher effects is also important in policy applications. It is helpful to examine the extent to which shrinking the estimates, as in the EB methods, increases bias in these noisy estimates. Thus, we report the average value of $\hat{\theta}$ across the simulations because it provides evidence of which methods, under which scenarios, produce estimated teacher effects whose magnitudes have meaning. This measure also provides insight into why some methods rank teachers relatively well even when the estimated effects are systematically biased.

The precision of these methods is also a key consideration when evaluating the overall performance. As described in Section 2, EB methods are not unbiased when thinking about the teacher effects as fixed parameters we are trying to estimate. However, if the identifying assumptions hold, these methods should provide more precise estimates. This is one motivation for using such methods, as estimates should be more stable over time, leading to a smaller variance in the teacher effects. As the teacher effect is fixed for each teacher across the 100 iterations, we have 100 estimates of each teacher effect. As a summary measure for the precision of the estimators, we calculate the standard deviation of the 100 teacher effect estimates for each teacher and then take a simple average across all teachers.

To further analyze the variance-bias tradeoff for each of these estimators, we also include average mean squared error (MSE). This measure averages the following across all $j$ teachers and across simulation runs:

$$\text{MSE} = (\beta_j - \hat{\beta}_j)^2$$ (34)

This provides a simple statistic to determine whether the bias induced by shrinking is justifiable due to gains in precision.
5 Simulation Results

Tables 1 and 2 report the five evaluation measures described in Section 4.2 for each particular estimator-assignment scenario combination. For ease in interpreting the tables, a quick guide to descriptions of each of these estimators, grouping-assignment mechanisms, and evaluation measures can be found in Appendix tables A.1 through A.3. As these shrinkage and EB estimators are often motivated as a way to reduce noise, one might expect these approaches to be most beneficial with very limited student data per teacher. Thus, we estimate teacher effects using both four cohorts and one cohort of data. The tables show results for the case $\lambda = .5$. Though not reported, we also conducted a full set of simulations for $\lambda = 0.75$ and $\lambda = 1$, and the main conclusions are unchanged. The full set of simulation results is available upon request from the authors.

5.1 Fixed Teacher Effects versus Random Teacher Effects

We first compare the performance of the DOLS estimator, which treats teacher effects as fixed parameters to estimate, to the AR and EB LAG estimators that treat teacher effects as random in Table 1. Under nonrandom assignment of teachers, we expect DOLS, which explicitly controls for teacher assignment through the inclusion of teacher assignment indicators, to perform better than those estimators treating the teacher effects as random. When teacher assignment is based on the lagged test score, DOLS directly controls for the assignment mechanism by including both the lagged score and teacher assignment indicators and should perform particularly well in this case. The simulation results presented here largely support this hypothesis.
5.1.1 Random Assignment

We begin with the pure random assignment (RA) case, where EB-type estimation methods are theoretically justified. The results of the random assignment case are given in the top panel of Table 1 and suggest that the difference between fixed and random effects estimators is not that substantial under this scenario. As the theory suggests, EB LAG performs well in the four cohort case, with rank correlations between the estimated and the true teacher effects near 0.76, slightly better than the 0.75 rank correlation for DOLS. The AR estimator, which uses OLS in the first stage instead of MLE, outperforms both DOLS and EB LAG in terms of the rank correlation even though it is not theoretically preferred. In addition to very similar rank correlations, the misclassification rates are very similar across the three estimators, with between 19 to 21 percent of above average teachers misclassified as below average. The similarities between the three estimators in terms of rank correlation and misclassification rates remains when using only one cohort. Reducing the amount of data used to estimate the teacher effects lowers the performance of all estimators, decreasing the rank correlations and increasing the misclassification rates. With one cohort, rank correlations between the estimated and true teacher effects are about 0.61 to 0.63, and between 27 and 28 percent of above average teachers are misclassified as below average.

In addition to rank correlations and misclassification rates, we also examine the bias and precision of the estimators. While DOLS and AR appear to be unbiased with average $\hat{\theta}$ values close to 1, EB LAG substantially underestimates the magnitudes of the true teacher effects with an average $\hat{\theta}$ value of 0.82 using four cohorts and 0.53 using one cohort. This bias is likely the result of the shrinkage technique that is applied. However, this shrinkage causes EB LAG to be slightly more precise than AR and DOLS. While DOLS and AR both have similar average standard deviations of the estimated teacher effects near 0.28 and 0.38 in the four and one cohort cases, respectively, EB LAG has lower average standard deviations of 0.23 and 0.20, respectively. Given the precision gain in EB LAG, the MSE measure suggests
that EB LAG may be preferred to DOLS under random assignment. The MSE for AR suggests that even under random assignment, DOLS and EB LAG would be preferred.

We now move to the cases where the students are *nonrandomly grouped* together, but teachers are still *randomly assigned* to classrooms, the DG-RA and HG-RA panels in Table 1. Under these two scenarios, we see a fairly similar pattern as in the RA scenario, although the overall performance of all estimators is slightly diminished.

### 5.1.2 Dynamic Grouping and Nonrandom Assignment

The performance of the various estimators diverges noticeably under *nonrandom* teacher assignment. We allow for nonrandom grouping based on either the prior year test score or student-level heterogeneity, but now allow for nonrandom assignment of students to teachers. Classes with high test scores or high unobserved ability can be assigned to either the best (positive assignment) or worst (negative assignment) teachers. A key finding of this analysis is the disparity in performance of estimators that treat teacher effects as random (e.g. AR and EB LAG) compared with the DOLS estimator. These results suggest that estimators explicitly controlling for the teacher assignment should be preferred to those that treat the teacher effects as random.

DOLS substantially outperforms AR and EB LAG under the DG-PA scenario. When using four cohorts, DOLS has a rank correlation of 0.76 under DG-PA, while AR and EB LAG have rank correlations of 0.60 and 0.70, respectively. AR and EB LAG also have large misclassification rates, with 28 to 34 percent of above average teachers being misclassified as below average compared with only 23 percent for DOLS. In addition to misclassifying and poorly ranking teachers, the AR and EB LAG methods also underestimate the magnitudes of the true teacher effects. While DOLS has an average \( \hat{\theta} \) value of 0.99, the AR and EB LAG estimators have average \( \hat{\theta} \) values of 0.75 and 0.70, respectively. While some of the bias of the EB LAG estimates can be attributed to shrinkage, the larger issue is the bias caused by the
failure of the AR and EB LAG approaches to net out the correlation between the lagged test score and the teacher assignment (i.e. the assignment mechanism in these DG scenarios), a correlation that DOLS explicitly allows for with the inclusion of teacher dummies in the regression. Just as in the random assignment case, DOLS and EB LAG have similar MSE measures, while the MSE for AR is substantially larger. In the four cohort case, DOLS, EB LAG and AR have MSE values of 0.15, 0.14, and 0.29, respectively.

These differences are even more noticeable under the DG-NA scenario. Again examining the four cohort case, DOLS has a rank correlation of 0.74, while AR and EB LAG have rank correlations of 0.54 and 0.68, respectively. Only 22 percent of above average teachers are misclassified by DOLS, while 31 and 25 percent are misclassified by AR and EB LAG, respectively. AR and EB LAG are even more biased under DG-NA, with $\hat{\theta}$ values of 0.46 and 0.51, respectively. This severe underestimation again offsets the gain in precision of the AR and EB LAG estimators, leading to MSE measures that are larger than for DOLS.

These simulation results also verify an important result of the theoretical discussion: the performance of EB LAG approaches the performance of DOLS as the number of students per teacher grows. We see less of a disparity in the performance of DOLS and EB LAG when computing VAMs using four cohorts compared to one, but the relative performance of AR does not improve with more students per teacher. For example, under DG-PA with one cohort of students, AR and EB LAG have similar rank correlations of 0.45 and 0.48, respectively, compared to 0.58 for DOLS. With four cohorts of students, the rank correlation for EB LAG is much closer to that for DOLS (0.70 and 0.76, respectively) than is the rank correlation for AR (0.60). This theoretical result is also applicable to the SAR estimator we examine below, which is used as a simpler way to operationalize the EB approach. In summary, EB LAG, which uses random effects estimation in the first stage, is preferred to those using OLS (AR and SAR) under nonrandom teacher assignment, as these estimates approach the preferred DOLS estimates that treat teacher effects as fixed.
5.1.3 Heterogeneity Grouping and Nonrandom Assignment

As a final scenario we examine the case of nonrandom teacher assignment to students grouped on the basis of student-level heterogeneity. The results for these HG scenarios are especially unstable: all estimators do an excellent job ranking teachers under positive teacher assignment, and all estimators do a very poor job under negative teacher assignment. In the HG-PA case with four cohorts of students, the magnitudes of the estimated VAMs are amplified as seen by the large average values for $\hat{\theta}$ between 1.45 and 1.61. This improves the ability of the various estimators to rank teachers as evidenced by the high rank correlations of about 0.89 for all estimators. The EB LAG estimator performs the best in this scenario, as it performs as well as the other estimators in terms of ranking and misclassification of teachers but has the smallest MSE measure. Under HG-NA with four cohorts, the performance of all estimators falls substantially, largely caused by severely underestimated teacher effects ($\hat{\theta}$ values between 0.18 and 0.33). These compressed teacher effect estimates make it difficult to rank teachers in this scenario, resulting in low rank correlations for all estimators between 0.29 and 0.32. Just as in the HG-PA scenario, the performance of the three estimators under HG-NA is very similar across the evaluation measures we examine.

Why is the performance of DOLS, AR, and EB LAG so similar under HG-PA and HG-NA, while differing so greatly under DG-PA and DG-NA? Despite correlation between the baseline test score and the student fixed effect, the lagged test score appears to be a weak proxy for the assignment mechanism in the HG scenarios. Since none of the three estimators do well at allowing for the correlation between the assignment mechanism and the teacher assignment in these cases, the distinction between estimators that include teacher fixed effects and those that treat teacher effects as random is less stark. As found in Guarino, Reckase, and Wooldridge (forthcoming), a gain score estimator with student fixed effects included is the most robust in these HG scenarios, as it does allow for the correlation between the assignment mechanism (i.e. student fixed effect) and the teacher assignment (i.e teacher
dummy variables). Their results lend further support for our conclusions here that allowing for this correlation is extremely important in the performance of these value added estimators when there is nonrandom assignment.

5.2 Shrinkage versus Non-Shrinkage Estimation

Use of EB and other shrinkage estimators is often motivated as a way to reduce the noise in the estimation of teacher effects, particularly for teachers with a small number of students. Greater stability in the estimated effects is thought to reduce misclassification of teachers. We observed in section 3.1 that EB LAG was generally outperformed by the fixed effects estimator, DOLS. However, under nonrandom teacher assignment, we are unable to tell how much of the bias in the EB LAG estimator is due to treating the teacher effects as random and how much is due to the shrinkage procedure. To examine the effects of shrinkage itself, we compare the performance of unshrunken estimators, DOLS and AR, with their shrunken versions, SDOLS and SAR, in Table 2. Although SDOLS is not a commonly used or theoretically justified estimator, we explore it here to identify whether shrinking teacher fixed effect estimates could be useful in practice.

Our simulation results show that there is no substantial improvement in the performance of the DOLS or AR estimators after applying the shrinkage factor to the estimates. Using four cohorts of students, the performance measures for DOLS and AR compared to their shrunken counterparts are nearly identical to two decimal places across all grouping and assignment scenarios. Even with very limited data per teacher in the one cohort case, when we would expect shrinkage to have a greater effect on the estimates, we find very little change in the performance of the estimators after the shrinkage factor is applied.

In the one cohort case, shrinking either the DOLS or AR estimates slightly decreases (in the second decimal place) both the average $\hat{\theta}$ values and average standard deviation of the estimated teacher effects. This increased bias in the estimates is expected when
applying the shrinkage factor and, depending on the scenario and estimator we examine, the effect of this precision-bias tradeoff may increase or decrease the MSE measure when comparing the shrunken and unshrunken estimates. Shrinking the DOLS estimates generally reduces both the misclassification of teachers and the MSE measure slightly. Shrinking the AR estimates doesn’t affect misclassification in most cases, and it actually increases misclassification slightly in the DG-NA scenario. Effects on the MSE measure are mixed for shrinking the AR estimates, but generally reduces the MSE measure slightly.

The effect of shrinkage itself does not appear to be practically important or ameliorate the performance of the biased AR estimator found in the DG-PA and DG-NA scenarios. Given that shrinking the AR estimates does little to mitigate the performance drop of AR under DG-PA and DG-NA, our evidence suggests that shrinking teacher fixed effects estimates is preferred over shrinking teacher random effects if such techniques are desired.

6 Comparing VAM Methods Using Real Data

We also apply these estimation methods to actual student-level test score data and examine the rank correlations between the estimated teacher effects of the various estimators for each school district. In addition to rank correlations, we also examine whether teachers are being classified in the extremes uniformly across all of the estimators we examine. Although the real data does not allow comparison between the estimated effects and the true teacher effects, we are able to make comparisons between the estimated effects of different estimators. This comparison provides a measure of the sensitivity of the estimated teacher effects to specifications that shrink the estimates and/or treat the teacher effects as random or fixed. The results of this analysis provide some perspective on the impact of shrinking and Empirical Bayes’ methods in a real-world setting.
6.1 Data

We apply the five methods described in Section 3 to data from an anonymous southern U.S. state. The data span 2001 through 2007 and grades four through six, but test scores are collected for each student from grades three through six. The data set includes 1,488,253 total students from which we have at least one current year score and one lagged score. Only 482,031 students have test scores for all grades. The data set also contains 43,868 unique teachers that we observe for a varying number of cohorts of students. We observe 39 percent of teachers for only one year, but we do see 20 percent of teachers for four or more years. These teachers, on average, teach about 26 students per year, with only a small percentage (less than two percent) teaching more than 30 students per year. The high percentage of teachers that we observe for only one year could motivate researchers to employ shrinkage and EB estimators as a way to reduce precision problems due to minimal data.

We estimate teacher effects district-by-district using equation (30) with controls for various student characteristics and include dummies for the year. Student characteristics include race, gender, disability status, free/reduced price lunch eligibility, limited English proficiency status, and the number of student absences from school. As discussed above, the teacher effects are estimated using data on multiple cohorts (between one and seven) of students. For simplicity and comparison with the simulation results, we estimate the value-added measures for those teachers with fifth grade students in the 67 districts, but again teachers receive multiple cohorts of students. Overall, we estimate 20,749 teacher effects using test score data from the annual assessment exam administered by the state.

6.2 Results

Figure 1 presents box plots that depict the distributions of the within-district rank correlations between the various lagged score estimators, DOLS, SDOLS, AR, SAR, and EB
LAG. As in the discussion of the simulation results, we first compare the DOLS estimator, which treats the teacher effects as fixed, with the estimators that treat the teacher effects as random. Comparing DOLS and AR, we find that the median rank correlation is around 0.99, but there are nine districts with rank correlations below 0.90 and 2 districts with correlations below 0.50. We see a slightly lower median rank correlation between DOLS and EB LAG, at around 0.97, with five districts with rank correlations below 0.90 and 3 below 0.50. These results are not inconsistent with our simulation results: the performance of DOLS, AR, and EB LAG is very similar under cases of random assignment of teachers to classrooms, but the performance of AR and EB LAG is substantially different from DOLS under non-random assignment based on prior test scores. Thus, it could be the case that the outlier districts we see in the left tails of the top two box plots may be composed of schools that engage more heavily in nonrandom assignment of teachers to classrooms.

Comparing the two estimators that treat teacher effects as random, AR and EB LAG, we find that while the median rank correlation is 0.96, nine districts have rank correlations of between 0.82 and 0.92. These results suggest that the estimates are somewhat sensitive to how the teacher effects are calculated in the first stage. This was also the case in the simulated results, where the performance of the AR estimator suffered more than the performance of the EB LAG estimator in cases of non-random assignment based on the prior test score.

For a thorough comparison with the simulation results, we also compare the shrunken and unshrunken estimates of DOLS and AR using the real data. We find median rank correlations of around 0.97 for both the DOLS and SDOLS comparison and the AR and SAR comparison, suggesting that shrinkage has a small impact on the estimates. Shrinkage may have a slightly larger impact on the DOLS estimates, as two districts have rank correlations of 0.50 and 0.72. Our simulation results suggested that shrinking the estimates had very little impact on estimator performance, but the SDOLS estimator showed the greatest boost in performance from shrinking, especially in the case of one cohort of students.
In addition to rank correlation comparisons, we also examine the extent to which teachers are classified in the tails of the distribution by the different estimators. If shrinkage is having some effect, we would expect to see some teachers classified in the extremes to be pushed toward the middle of the distribution after applying the shrinkage factor. Table 3 lists the fraction of teachers ranked in the same quintile, either the top or bottom, by different pairs of estimators. Comparing across estimators that assume fixed teacher effects to those that assume random teacher effects, we do not see much movement across quintiles. For example, comparing DOLS to EB LAG, we find that about 91 percent of the teachers that are classified in the top quintile using DOLS are also in this quintile using EB LAG. This suggests that teacher assignment may not be largely based on prior student achievement or that the prior test score is a poor proxy for the true assignment mechanism. If the prior test score or other covariates insufficiently proxy for the underlying assignment mechanism, then the choice to include teacher assignment variables will matter little in how teachers are ranked.

Comparing the rankings of unshrunk and corresponding shrunken estimators, we see that about 90 percent of teachers are ranked in the same quintile by both the unshrunk estimators (DOLS and AR) and their shrunken counterparts (SDOLS and SAR). This suggests that shrinking the estimates results in some reclassification of teachers in the tails to quintiles in the middle of the distribution. Using real data, however, we are unable to tell whether this reclassification is appropriate. Our simulated analysis suggested that shrinking the estimates had little impact if any on misclassification rates.

7 Conclusion

Using simulation experiments where the true teacher effects are known, we have explored the properties of two commonly used Empirical Bayes’ estimators as well as the effects of shrinking estimates of teacher effects in general. Overall, EB methods do not appear to have
much advantage, if any, over simple methods such as DOLS that treat the teacher effects as fixed, even in the case of random teacher assignment where EB estimation is theoretically justified. Under random assignment, all estimators perform well in terms of their ability of ranking teachers, properly classifying teachers, and providing unbiased estimates. EB methods have a very slight gain in precision compared to the other methods in this case.

We generally find that EB estimation is not appropriate under nonrandom teacher assignment. The hallmark of EB estimation of teacher effects is to treat the teacher effects as random variables that are independent (or at least uncorrelated) with any other covariates. This assumption is tantamount to assuming that teacher assignment does not depend on other covariates such as past test scores (this is also true for the AR methods). When teacher assignment is not random, estimators that either explicitly control for the assignment mechanism or proxy for it in some way typically provide more reliable estimates of the teacher effects. Among the estimators and assignment scenarios we study, DOLS and SDOLS are the only estimators that control for the assignment mechanism (again, either explicitly or by proxy) through the inclusion of both the lagged test score and teacher assignment dummies. As expected, DOLS and SDOLS outperform the other estimators in the nonrandom teacher assignment scenarios. In the analysis of the real data, we found that the rank correlations between, say, DOLS and EB LAG or DOLS and SAR are quite low for some districts, suggesting that the decision among these estimators is important. Thus, if there is a possibility of nonrandom assignment, DOLS should be the preferred estimator.

As predicted by theory and seen in the simulation results, the random effects estimator, EB LAG, converges to the fixed effects estimator, DOLS, as the number of students per teacher gets large. Therefore, it could be that EB LAG is performing well in large samples simply because the estimates are approaching the DOLS estimates. However, the average residual methods, AR and SAR, do not have this property. Thus, despite the recent popularity, we strongly caution using SAR as a simpler way to operationalize the EB LAG estimator.
If EB-type methods are being used, it is important to estimate the coefficients in the first stage using random effects estimation (as in our EB LAG estimator) rather than OLS.

Lastly, we find that shrinking the estimates of the teacher effects does not seem to improve the performance of the estimators, even in the case where estimates are based on one cohort of students. The performance measures are extremely close in our simulations for those estimators that differ only due to the shrinkage factor – DOLS and SDOLS or AR and SAR. The rank correlations for these two pairs of estimators are also very close to one in almost all districts. Also, we find in the simulations that shrinking the AR estimates, which is a popular way to operationalize the EB approach, doesn’t reduce misclassification of teachers. Thus, our evidence suggests that the rationale for using shrinkage estimators to reduce the misclassification of teachers due to noisy estimates of teacher effects should not be given much weight. Accounting for nonrandom teacher assignment when choosing among estimators is more imperative.

Given the robust nature of the DOLS estimator to a wide variety of grouping and assignment scenarios, it should be widely preferred to AR and EB methods when there is uncertainty about the true underlying assignment mechanism. If the assignment mechanism is known to be random, applying these AR and EB estimators can be appropriate, especially when the amount of data per teacher is minimal. However, given that the assignment mechanism is not likely known, blindly applying these AR and EB methods can be extremely problematic, especially if teachers are truly assigned nonrandomly to classrooms. Therefore, we stress caution in applying theses AR and EB methods and urge researchers to be mindful of the underlying assignment mechanism when choosing between the various value-added methods.
Notes

1. In simulations, Hansen (2007) shows that the asymptotic properties work well when $G$ and $N_g$ are roughly around 40.

2. Lockwood and McCaffrey (2007) have highlighted equation (27) in the context of student-level panel data, essentially appealing to the first edition of Wooldridge (2010). In the panel data setting (27) is arguably less relevant, as one rarely has more than a handful of time periods per student. For additional discussion of the relationship between random and fixed effects estimators, see Raudenbush (2009). In addition, Reardon and Raudenbush (2009) lay out the various assumptions underlying value-added estimation.

3. Without covariates, the difference between the EB and fixed effects estimates of the $b_g$ is much less important: they differ only due to the shrinkage factor. In practice, the fixed effects estimates, $\hat{\beta}_g$, are obtained without removing an overall teacher average, which means $\hat{\beta}_g = \bar{y}_g$. To obtain a comparable expression for $b_g^*$ we must account for the GLS estimator of the mean teacher effect, which would be obtained as the intercept (the only parameter) in the RE estimation. Call this estimator $\mu_{b_g}^*$, which in the case of no covariates is $\gamma^*$. Then the teacher effects are

$$b_g^* = \mu_{b_g}^* + \eta_g(\bar{y}_g - \mu_{b_g}^*) = \eta_g \bar{y}_g + (1 - \eta_g)\mu_{b_g}^* = \bar{y}_g - (1 - \eta_g)(\bar{y}_g - \mu_{b_g}^*),$$

where $\eta_g$ is the shrinkage factor in equation (25). Compared with the FE estimate of $b_g$, $b_g^*$ is shrunk toward the overall mean $\mu_{b_g}^*$. When the teacher effects are treated as parameters to estimate, the $b_g^*$ are biased because of the shrinkage factor, even in the case in which they are BLUP.

4. As described in Rabe-Hesketh and Skrondal (2012), this two step procedure can be performed in one-step using the “xtmixed” command in Stata with teacher random effects.
The predicted random effects of this regression are identical to shrinking the MLE estimates by the shrinkage factor. This procedure is generally justified even if the unobservables do not have normal distributions, in which case we are applying quasi-MLE.

5. See, for example, Kane and Staiger, 2008). AR and SAR will be fairly similar with large class sizes and will be consistent under the same assumptions. The finite-sample performance of these estimators will differ only due to the shrinkage factor. It is important to keep in mind that, unlike DOLS and SDOLS, the AR and SAR estimators do not allow for general correlation between the teacher assignment and past test scores (or other covariates).

6. Despite only estimating value-added for grade 5 teachers, we keep the three grade structure when generating the student test scores to more realistically capture how fifth grade test scores are determined. The fifth grade achievement is based on more than just the current teacher and prior test score of the student; it is a function of all prior teacher, unobservable student, and random influences. Thus, to ignore that process and generate fifth grade test scores based on a “baseline” fourth grade test score seems inappropriate given this context.

7. For lag scores greater than one year prior, lambda is set equal to zero in the simulations. Some models, however, use multiple prior test scores (e.g. EVAAS, VARC) and we estimate DOLS, AR, and EB LAG with multiple lagged test scores as a sensitivity analysis. Although adding multiple lags improves the performance of AR and EB LAG in the random assignment case, the performance of these estimators still suffers greatly compared to DOLS in the DG-PA and DG-NA scenarios. Thus, the results of this sensitivity analysis suggest that adding additional lags to equation (30) does little to change our overall conclusions.

8. Because we assume independence of $e_{it}$ over time we are maintaining the so-called “common factor restriction” in the underlying cumulative effects model. This restriction implies that past shocks to student learning decay at the same rate as all inputs. See Guarino, Reckase, and Wooldridge (2012) for a more detailed discussion of this assumption.
9. The mean and standard deviation of the true teacher effects of the 36 teachers we estimate are 0.501 and 0.244, respectively.

10. We also conduct a sensitivity analysis using “large” teacher effects, where the true teacher effects are drawn from a distribution with a standard deviation of 1. When teacher effects are large, the performance of all estimators is increased and EB LAG performs similarly to DOLS in the DG-PA and DG-NA cases. The AR method, however, continues to suffer in performance under DG-PA and DG-NA compared to DOLS.

11. As a sensitivity analysis, we also run simulations with this standard deviation = 0.1, meaning the grouping of students into classrooms is more nonrandom. In this case, the performance of AR and EB LAG suffers even more greatly in terms of lower rank correlations and higher misclassification rates than what is observed in the results in Table 1 and 2.

12. Although the results are omitted from the paper we also examine the fraction of above average teachers that are misclassified in the bottom quintile and the fraction of below average teachers that are misclassified in the top quintile. As expected, the conclusions are very similar to those drawn from the misclassification rates reported in the tables. Teachers are more likely to be misclassified in the extremes by AR and EB LAG under DG-PA and DG-NA than under random assignment, while DOLS misclassifies teachers at similar rates across all scenarios. Shrinking the estimates also does not appear to have much impact on these misclassification rates.
References


Journal of the American Statistical Association 78(381), 47-55


Table 1: Simulation Results: Comparing Fixed and Random Teacher Effects Estimators

<table>
<thead>
<tr>
<th>G-A Mechanism</th>
<th>Evaluation Type</th>
<th>Four Cohorts</th>
<th>One Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.75</td>
<td>AR 0.78</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td>RA</td>
<td>Avg. Theta</td>
<td>1.01</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.75</td>
<td>AR 0.77</td>
</tr>
<tr>
<td>DG-RA</td>
<td>Misclassification</td>
<td>0.22</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.76</td>
<td>AR 0.60</td>
</tr>
<tr>
<td>DG-PA</td>
<td>Misclassification</td>
<td>0.23</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>0.99</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.74</td>
<td>AR 0.54</td>
</tr>
<tr>
<td>DG-NA</td>
<td>Misclassification</td>
<td>0.22</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.01</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.14</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.67</td>
<td>AR 0.67</td>
</tr>
<tr>
<td>HG-RA</td>
<td>Misclassification</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.02</td>
<td>1.01</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.17</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.89</td>
<td>AR 0.88</td>
</tr>
<tr>
<td>HG-PA</td>
<td>Misclassification</td>
<td>0.13</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.61</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.17</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Rank Correlation</td>
<td>DOLS 0.32</td>
<td>AR 0.29</td>
</tr>
<tr>
<td>HG-NA</td>
<td>Misclassification</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>0.33</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.20</td>
<td>0.32</td>
</tr>
</tbody>
</table>

**Note:** Rows of each scenario represent the following:
First - Rank corr. of estimated effects and true effects
Second - Fraction of above avg. teachers misclassified below avg.; Third - Avg. value of $\hat{\theta}$
Fourth - Average standard deviation of estimated teacher effects across 100 reps
Fifth - MSE measure
<table>
<thead>
<tr>
<th>G-A Mechanism</th>
<th>Evaluation Type</th>
<th>DOLS</th>
<th>SDOLS</th>
<th>AR</th>
<th>SAR</th>
<th>DOLS</th>
<th>SDOLS</th>
<th>AR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RA</td>
<td>Rank Correlation</td>
<td>0.75</td>
<td>0.75</td>
<td>0.78</td>
<td>0.78</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.21</td>
<td>0.21</td>
<td>0.19</td>
<td>0.19</td>
<td>0.28</td>
<td>0.26</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.01</td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.03</td>
<td>0.99</td>
<td>1.02</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.28</td>
<td>0.27</td>
<td>0.27</td>
<td>0.38</td>
<td>0.36</td>
<td>0.37</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.15</td>
<td>0.28</td>
<td>0.28</td>
<td>0.20</td>
<td>0.18</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>DG-RA</td>
<td>Rank Correlation</td>
<td>0.75</td>
<td>0.75</td>
<td>0.77</td>
<td>0.77</td>
<td>0.59</td>
<td>0.58</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.22</td>
<td>0.22</td>
<td>0.20</td>
<td>0.20</td>
<td>0.29</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
<td>0.97</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.37</td>
<td>0.36</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.14</td>
<td>0.29</td>
<td>0.29</td>
<td>0.20</td>
<td>0.17</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>DG-PA</td>
<td>Rank Correlation</td>
<td>0.76</td>
<td>0.76</td>
<td>0.60</td>
<td>0.60</td>
<td>0.58</td>
<td>0.59</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.23</td>
<td>0.23</td>
<td>0.34</td>
<td>0.34</td>
<td>0.30</td>
<td>0.30</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>0.99</td>
<td>0.99</td>
<td>0.75</td>
<td>0.75</td>
<td>0.96</td>
<td>0.95</td>
<td>0.73</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.28</td>
<td>0.23</td>
<td>0.23</td>
<td>0.37</td>
<td>0.36</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.15</td>
<td>0.15</td>
<td>0.33</td>
<td>0.33</td>
<td>0.21</td>
<td>0.18</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>DG-NA</td>
<td>Rank Correlation</td>
<td>0.74</td>
<td>0.74</td>
<td>0.54</td>
<td>0.54</td>
<td>0.60</td>
<td>0.61</td>
<td>0.36</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.22</td>
<td>0.22</td>
<td>0.31</td>
<td>0.31</td>
<td>0.28</td>
<td>0.28</td>
<td>0.37</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.01</td>
<td>1.01</td>
<td>0.46</td>
<td>0.46</td>
<td>1.03</td>
<td>1.01</td>
<td>0.48</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.28</td>
<td>0.28</td>
<td>0.18</td>
<td>0.18</td>
<td>0.38</td>
<td>0.36</td>
<td>0.30</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.14</td>
<td>0.14</td>
<td>0.31</td>
<td>0.31</td>
<td>0.20</td>
<td>0.16</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>HG-RA</td>
<td>Rank Correlation</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.67</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.02</td>
<td>1.02</td>
<td>1.01</td>
<td>1.01</td>
<td>0.99</td>
<td>0.96</td>
<td>0.98</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
<td>0.40</td>
<td>0.39</td>
<td>0.39</td>
<td>0.38</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.17</td>
<td>0.17</td>
<td>0.33</td>
<td>0.33</td>
<td>0.22</td>
<td>0.20</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>HG-PA</td>
<td>Rank Correlation</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
<td>0.78</td>
<td>0.78</td>
<td>0.77</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.13</td>
<td>0.13</td>
<td>0.14</td>
<td>0.14</td>
<td>0.20</td>
<td>0.20</td>
<td>0.21</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>1.61</td>
<td>1.61</td>
<td>1.57</td>
<td>1.57</td>
<td>1.60</td>
<td>1.57</td>
<td>1.55</td>
<td>1.52</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.41</td>
<td>0.41</td>
<td>0.40</td>
<td>0.40</td>
<td>0.48</td>
<td>0.47</td>
<td>0.46</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.17</td>
<td>0.16</td>
<td>0.36</td>
<td>0.35</td>
<td>0.22</td>
<td>0.20</td>
<td>0.41</td>
<td>0.39</td>
</tr>
<tr>
<td>HG-NA</td>
<td>Rank Correlation</td>
<td>0.32</td>
<td>0.32</td>
<td>0.29</td>
<td>0.29</td>
<td>0.25</td>
<td>0.27</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Misclassification</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
<td>0.40</td>
<td>0.39</td>
<td>0.42</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Avg. Theta</td>
<td>0.33</td>
<td>0.33</td>
<td>0.27</td>
<td>0.27</td>
<td>0.34</td>
<td>0.37</td>
<td>0.28</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>Avg. Std. Dev.</td>
<td>0.18</td>
<td>0.18</td>
<td>0.17</td>
<td>0.17</td>
<td>0.29</td>
<td>0.27</td>
<td>0.28</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.20</td>
<td>0.20</td>
<td>0.32</td>
<td>0.32</td>
<td>0.25</td>
<td>0.20</td>
<td>0.36</td>
<td>0.35</td>
</tr>
</tbody>
</table>

**Note:** Rows of each scenario represent the following: First - Rank corr. of estimated effects and true effects; Second - Fraction of above average teachers misclassified as below average; Third - Average value of $\hat{\theta}$; Fourth - Average standard deviation of estimated teacher effects across 100 reps; Fifth - MSE measure.
Figure 1: Spearman Rank Correlations Across Different VAM Estimators

Table 3: Fraction of Teachers Ranked in Same Quintile by Estimator Pairs

<table>
<thead>
<tr>
<th></th>
<th>DOLS</th>
<th>SDOLS</th>
<th>AR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top Quintile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDOLS</td>
<td>0.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.94</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>0.89</td>
<td>0.94</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>EB LAG</td>
<td>0.87</td>
<td>0.95</td>
<td>0.86</td>
<td>0.93</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>DOLS</th>
<th>SDOLS</th>
<th>AR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom Quintile</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDOLS</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AR</td>
<td>0.96</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAR</td>
<td>0.88</td>
<td>0.95</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>EB LAG</td>
<td>0.87</td>
<td>0.98</td>
<td>0.86</td>
<td>0.96</td>
</tr>
</tbody>
</table>
A Appendix

Table A.1: Definitions of Grouping-Assignment Mechanisms

<table>
<thead>
<tr>
<th>Name of G-A Mechanism</th>
<th>Acronym</th>
<th>Grouping students in classrooms</th>
<th>Assigning students to teachers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Assignment</td>
<td>RA</td>
<td>Random</td>
<td>Random</td>
</tr>
<tr>
<td>Dynamic Grouping - Random Assignment</td>
<td>DG-RA</td>
<td>Dynamic (based on prior test scores)</td>
<td>Random</td>
</tr>
<tr>
<td>Dynamic Grouping - Positive Assignment</td>
<td>DG-PA</td>
<td>Dynamic (based on prior test scores)</td>
<td>Positive corr. between teacher effects and prior student scores</td>
</tr>
<tr>
<td>Dynamic Grouping - Negative Assignment</td>
<td>DG-NA</td>
<td>Dynamic (based on prior test scores)</td>
<td>Negative corr. between teacher effects and prior student scores</td>
</tr>
<tr>
<td>Heterogeneity Grouping - Random Assignment</td>
<td>HG-RA</td>
<td>Static (based on student heterogeneity)</td>
<td>Random</td>
</tr>
<tr>
<td>Heterogeneity Grouping - Positive Assignment</td>
<td>HG-PA</td>
<td>Static (based on student heterogeneity)</td>
<td>Positive corr. between teacher effects and student fixed effects</td>
</tr>
<tr>
<td>Heterogeneity Grouping - Negative Assignment</td>
<td>HG-NA</td>
<td>Static (based on student heterogeneity)</td>
<td>Negative corr. between teacher effects and student fixed effects</td>
</tr>
</tbody>
</table>

Table A.2: Description of Evaluation Measures of Value-Added Estimator Performance

<table>
<thead>
<tr>
<th>Evaluation Measure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank Correlation</td>
<td>Rank correlation between estimated teacher effect and true teacher effect</td>
</tr>
<tr>
<td>Misclassification</td>
<td>Fraction of above average teachers that are misclassified as below average</td>
</tr>
<tr>
<td>Average Theta</td>
<td>Average value of ( \hat{\theta} )</td>
</tr>
<tr>
<td>Avg. Std. Dev.</td>
<td>Average standard deviation of estimated teacher effects across the 100 simulation reps</td>
</tr>
<tr>
<td>MSE</td>
<td>Average value of ( \bar{MSE} = (\bar{\beta}_j - \hat{\beta}_j)^2 )</td>
</tr>
</tbody>
</table>
### Table A.3: Description of Value-Added Estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Acronym</th>
<th>Description</th>
<th>Teacher Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical Bayes'</td>
<td>EB LAG</td>
<td>Two-step approach: Estimate teacher effects using MLE on dynamic equation and then shrink estimates by shrinkage factor</td>
<td>Random</td>
</tr>
<tr>
<td>Average Residual</td>
<td>AR</td>
<td>Estimate dynamic equation by OLS and compute residuals for each student. Then compute the average of these residuals for each teacher to get estimated teacher effect</td>
<td>Random</td>
</tr>
<tr>
<td>Dynamic OLS</td>
<td>DOLS</td>
<td>Estimate teacher effects using ordinary least squares on dynamic equation</td>
<td>Fixed</td>
</tr>
<tr>
<td>Shrunken DOLS</td>
<td>SDOLS</td>
<td>Two-step approach: Estimate teacher effects using dynamic equation and then shrink estimates by shrinkage factor</td>
<td>Fixed</td>
</tr>
<tr>
<td>Shrunken Avg. Residual</td>
<td>SAR</td>
<td>Two-step approach: Compute average residual for each teacher using residuals from OLS on dynamic equation. Then shrink average residual for each teacher by shrinkage factor</td>
<td>Random</td>
</tr>
</tbody>
</table>