UNC CHARLOTTE ECONOMICS WORKING PAPER SERIES

UNDERSTANDING AND EVALUATING THE SAS® EVAAS® UNIVARIATE RESPONSE MODEL (URM) FOR MEASURING TEACHER EFFECTIVENESS

Kelly N. Vosters,
Cassandra M. Guarino,
Jeffrey M. Wooldridge

Working Paper No. 2018-001

THE UNIVERSITY OF NORTH CAROLINA AT CHARLOTTE
BELK COLLEGE OF BUSINESS
DEPARTMENT OF ECONOMICS
9201 University City Blvd
Charlotte, NC 28223-0001
February 2018

UNC Charlotte economics working papers represent research work-in-progress and are circulated for discussion and comment purposes. They have not been peer reviewed and the expressed views solely represent those of the authors. All rights reserved to the authors. Unless otherwise indicated below, short sections of the text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit is given to the source.

ABSTRACT

Despite little attention or exposure in the evaluation literature, the two SAS® EVAAS® models for estimating teacher effectiveness are used by several states and districts, in some cases for high stakes policies regarding teacher tenure, retention, or incentive pay. The EVAAS approach involves using one of two distinct models, the Multivariate Response Model (MRM) or the Univariate Response Model (URM). In this paper, we discuss and illustrate advantages or disadvantages of the EVAAS URM relative to the other widely used and studied value-added methods. We perform simulations to evaluate their ability to uncover true teacher effects under various teacher assignment scenarios. We also use administrative data to illustrate the extent of agreement between the URM and other common value-added approaches. Although the differences are small in our administrative data, we show with theory and simulations that standard linear regression using OLS often performs at least as well as—and some- times better than—the more complicated EVAAS URM.

Understanding and evaluating the SAS® EVAAS® Univariate Response Model (URM) for measuring teacher effectiveness

Kelly N. Vosters[a*], Cassandra M. Guarino[b], and Jeffrey M. Wooldridge[c]

[a] *University of North Carolina Charlotte, 9201 University City Blvd, Charlotte, NC 28223*
[*] *Corresponding author. Email: kvosters@uncc.edu*
[b] *University of California Riverside, 900 University Ave, Riverside, CA 92521*
[c] *Michigan State University, 110 Marshall-Adams Hall, East Lansing, MI 48824*

February 25, 2018

Abstract

Despite little attention or exposure in the evaluation literature, the two SAS® EVAAS® models for estimating teacher effectiveness are used by several states and districts, in some cases for high stakes policies regarding teacher tenure, retention, or incentive pay. The EVAAS approach involves using one of two distinct models, the Multivariate Response Model (MRM) or the Univariate Response Model (URM). In this paper, we discuss and illustrate advantages or disadvantages of the EVAAS URM relative to the other widely used and studied value-added methods. We perform simulations to evaluate their ability to uncover true teacher effects under various teacher assignment scenarios. We also use administrative data to illustrate the extent of agreement between the URM and other common value-added approaches. Although the differences are small in our administrative data, we show with theory and simulations that standard linear regression using OLS often performs at least as well as—and sometimes better than—the more complicated EVAAS URM.

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

JEL codes: I20, I21, I28, J08, J24, J45

1 Introduction

The federal Race to the Top policy, established in 2009 with funding from the American Recovery and Reinvestment Act (ARRA), led to a rise of teacher evaluation systems as a component of state and district education policies, and, in particular, mandated that student-achievement-based measures be included in these evaluation systems (see, e.g., Doherty & Jacobs, 2015; Steinberg & Donaldson, 2016). As a result, teacher performance measures have been considered for tenure decisions in 23 states and for dismissal decisions in 28 states (Doherty & Jacobs, 2015). Because high stakes decisions were being tied to teacher evaluation in these states, a large body of literature has examined the statistical merits of methods that use student achievement test scores to estimate teacher effectiveness (e.g., Backes, Cowan, Goldhaber, Koedel, Miller & Xu, 2018; Chetty, Friedman, & Rockoff, 2014; Goldhaber & Chaplin, 2015; Goldhaber, Walch, & Gabele, 2014; Guarino, Maxfield, Reckase, Thompson, & Wooldridge, 2015; Guarino, Reckase, & Wooldridge, 2015; Harris, 2009; Ishii & Rivkin, 2009; Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008; Koedel & Betts, 2009; Koedel & Betts, 2011; Mariano, McCaffrey, & Lockwood, 2010; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rothstein, 2010; Sass, Semykina, & Harris, 2014).

More recently, the Every Student Succeeds Act (ESSA) has backed away from emphasizing teacher evaluation, but since many states and districts continue to use student test scores in evaluating teachers and since the policy may resurge in the future, it is important that efforts to fill out the literature examining the strengths and weakness of different value-added models continue.

The SAS® EVAAS® model is one of the most commonly used approaches for evaluating teacher performance (Ballou & Springer, 2015). EVAAS is used statewide in North Carolina, Ohio, Pennsylvania, South Carolina, and Tennessee, providing reporting services to all districts, public schools, and charter schools (SAS Institute Inc., 2014, 2015a). EVAAS

has also been used by various districts in several other states, including Arkansas, California, Colorado, Connecticut, Delaware, Georgia, Indiana, Louisiana, Missouri, New Jersey, New York, Texas, Virginia, and Wyoming (SAS Institute Inc., 2011, 2015a).

The SAS EVAAS approach has a long history and was the first value-added approach to be used to evaluate teachers.[1] The current name, EVAAS, stands for Education Value-Added Assessment System, a variant on the earlier name Tennessee Value-Added Assessment System (TVAAS), as Tennessee was where it was developed and used since the early 1990's.[2] EVAAS methods include two options for estimating teacher effectiveness; the Univariate Response Model (URM) and the multivariate response model (MRM). The MRM, also referred to as the "layered" teacher model, involves joint modeling of scores from multiple tested subjects for multiple grades and cohorts in up to a five-year period. Jointly modeling the test scores aims to improve efficiency, and using the complete set of scores available for a student attempts to account for any other student characteristics that might affect achievement. This model is sometimes not feasible if data requirements cannot be met. Hence, the URM was developed for these situations. The URM focuses on a single subject and is thus less intensive computationally and more flexible with respect to data requirements. The method involves computing a single composite score for each student based on their lagged scores in the same subject as well as others, and then using this composite score as the only regressor in empirical Bayes' estimation of the teacher effects.

While the properties of the MRM have been investigated in a small number of research studies, the URM has received very little attention in the literature, and yet is currently used in at least three states (North Carolina, Ohio, and Tennessee).[3] This paper fills a

[1]The approach was first developed by William Sanders, a statistician studying agricultural phenomena related to cattle reproduction, who recognized the potential for applying his methods to teacher evaluation.

[2]The name is often modified in a similar fashion in states which adopt the EVAAS methods, such as "PVAAS" for Pennsylvania (e.g., www.portal.state.pa.us, accessed 1/12/2015).

[3]SAS documentation describes the URM and other methods used in each of these states: North Carolina, https://ncdpi.sas.com/support/EVAAS-NC-TechnicalDocumentation-2016.pdf; Ohio, http://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/Technical-Documentation-of-EVAAS-Analysis.pdf.aspx; Tennessee, http://tn.gov/assets/entities/education/attachments/tvaas_technical_documentation_2016.pdf.

gap by assessing the strengths and limitations of the URM relative to other methods. Our paper further contributes to the literature by providing researchers with a helpful mapping between the URM and value-added models. We draw key theoretical connections between the URM and other estimation approaches, focusing primarily on areas of overlap between URM, OLS, and empirical Bayes' estimation of typical value-added models, and we also show how and where the various estimation approaches differ. Through the theoretical discussion, simulations, and empirical work, we show that standard linear regression techniques perform very similarly—and in certain cases better—under plausible data scenarios. In addition, our detailed descriptions of the URM help make it more readily available for other researchers to implement and include in future evaluation studies.

We begin by reviewing the prior literature on EVAAS models in Section 2, and then provide a detailed technical description of common value-added models (VAMs) and the EVAAS URM approach in Section 3. In Section 4, we discuss our simulation design and present results from the simulation. Section 5 describes our empirical analysis and results using administrative data. We summarize and conclude in Section 6.

2 Prior Literature Evaluating EVAAS Methods

While the research literature on estimating teacher effectiveness has been growing rapidly in recent years, only a small number of papers made available by the SAS Institute and a handful of research studies external to SAS have implemented either of the EVAAS teacher models in simulations or using administrative data. The papers made available by the SAS Institute discuss theoretical advantages of the EVAAS methods, and some also evaluate the performance of these methods (e.g., Sanders, 2006; Wright, 2010; Wright, 2015; Wright, Sanders, & Rivers, 2006). These papers tend to focus on the scaling of the test scores, measurement error in the test scores, missing data, and shrinkage.

One of the only SAS papers to examine the URM is focused on the concern that the measurement error in the lagged test scores will cause bias in the estimates of teacher effects, leading to instability in the estimates as well. In a paper comparing a standardized gain model, a student growth percentile model, the URM, and the MRM, Wright (2010) found that the estimated teacher effects in the URM and MRM had smaller correlations with the percentage of students in a teacher's class who were eligible for free-and-reduced-price lunches than those in the other two models. The proposed method for mitigating measurement error bias was to include multiple lag scores (at least three), based on the argument that the measurement error tends to average out. Wright (2015) compared the mean squared error (MSE) for estimators that treat the teacher effects as fixed (such as standard linear regression) or random (as the URM does) when test scores are measured with error. When including three or more scores in a scenario where students are randomly assigned to teachers, the estimators performed fairly similarly. Under positive assignment though (higher achieving students assigned to better teachers), the random effects estimators did best while under negative assignment the fixed effects estimators did best. However, requiring additional lag test scores can also worsen missing data issues. Thus, although SAS papers assert that an advantage of the URM is that it mitigates missing data issues, the URM still requires observing at least four test scores for each student (specifically, the current score and at least three prior scores) (Wright et al., 2010).

While the SAS papers have provided some evidence on the statistical merits of the EVAAS methods, the details that allow researchers external to the SAS Institute to easily replicate the approach remain somewhat elusive. The brief nature of the documentation combined with proprietary programs and data have been thought to limit the implementation of EVAAS in external evaluation and replication studies (Amrein-Beardsley, 2008; Kupermintz, 2003). Of the small number of such studies, the majority focus on a specific assumption or characteristic of the EVAAS MRM.

Turning briefly to the small body of existing literature on EVAAS by authors outside the SAS Institute, we first mention those that study the MRM, and then discuss URM-related studies in more detail. Ballou and Springer (2015) point out weaknesses of EVAAS implementation in particular teacher evaluation systems, such as how teachers are classified into groups using a t-statistic constructed from their estimated effectiveness. A few studies have addressed signature features of the MRM, such as the omission of student covariates or joint modeling of subjects, typically focusing on a comparison to a generalized or modified version of the model (e.g., Ballou, Sanders, & Wright, 2004; Lockwood & McCaffrey, 2007; Lockwood, McCaffrey, Mariano, & Setodji, 2007; Mariano, McCaffrey, & Lockwood, 2010; McCaffrey et al., 2004).

For example, one of the most commonly raised concerns about EVAAS is the omission of student covariates. A few studies have identified scenarios where the MRM or MRM-type approaches produce teacher effect estimates that are biased and scenarios where the bias is reduced (Ballou, Sanders, & Wright, 2004; Lockwood & McCaffrey, 2007; McCaffrey et al., 2004). Although these studies suggest bias from omitting student covariates may not be large in certain scenarios, the applicability of this evidence is limited to the MRM, as none of these studies have included the URM. The URM might not exhibit the same type of bias reduction properties described in these studies because the URM does not exploit within-student correlation in test scores.

Another commonly raised concern with using administrative data is how to handle missing data. Researchers have also found some robustness of the MRM-type approaches with respect to the treatment of missing data (Lockwood et al., 2007; McCaffrey & Lockwood, 2011). Again, the applicability of these results are limited to the MRM. The joint modeling of test scores plays a key role in robustness properties of the MRM. Although the URM does, to some extent, control for student test scores from multiple years and subjects, the theoretical implications are very different from jointly modeling the scores. Thus, we now

turn to the limited existing evidence related directly to the URM.

The more limited non-SAS research evaluating the URM consists mainly of two reports, Rose, Henry, and Lauen (2012) and Henry and Rose (2014), that provide comparisons among a broad set of value-added approaches. They discuss assumptions needed for causal inference and implications of violations, and also provide simulation and statewide empirical evidence using three years of administrative data from North Carolina. The estimation approaches include several methods that treat the teacher effects as random, including the URM, three HLM approaches, two which take the within-teacher average of the residuals, and then three methods that treat the teacher effects as fixed. The specifications vary with respect to the number of (and subject of) lagged scores, school effects, student covariates and the data structure used (panel or cross section). They generally promote the three-level HLM and the URM as the most robust, but also acknowledge high agreement with many of the approaches and discuss various robustness properties of the other estimators.[4] Their preferred approaches all treat the teacher effects as a random component of the error term, thus relying on similar assumptions for the consistency of the estimated teacher effects.

We build on the work of Rose, Henry, and Lauen (2012) and Henry and Rose (2014) by providing more evidence specific to the URM, to fill the gap in a small literature on EVAAS that largely focuses on the MRM. Both the MRM and URM accommodate (some) incomplete student records and use test scores from several subjects and years, but they do so in very different ways. Our paper focusing on the URM also includes both simulations and the analysis of actual data but we consider different estimators and specifications and our simulations are designed somewhat differently, so our results diverge from theirs. While we confirm that random effects approaches such as HLM perform well under random assignment of students to teachers (a result found in Guarino, Reckase, and Wooldridge, 2015), we find that under nonrandom assignment based on prior achievement, approaches that allow

[4]Rose, Henry, and Lauen (2012) also note that their results (and hence conclusions) for the DOLS estimator may differ from those in Guarino, Reckase and Wooldridge (2015) due to simulation design. Their DOLS and URM implementation also appear to differ from that used in Guarino et al. (2015) and the present paper.

teacher assignment to be correlated with prior achievement are better suited to capturing true teacher effects than the URM. A key distinction here is the difference in the exogeneity assumptions of random and fixed effects estimators or, more simply, whether we consider the assignment of students to teachers to be random or nonrandom (that is, as a function of observed previous test scores or unobserved student characteristics).

When coefficients on the teacher assignment indicators are included along with prior test scores, students can be assigned to teachers based on their prior test scores or any covariates included in the regression (but not on omitted variables that are relevant to assignment). However, approaches that treat the teacher effects as a random component of the error term—and then use a prediction formula to uncover the teacher effects—rely on the assumption that students are *not* assigned to teachers based on included or omitted covariates. The URM assumes random teacher effects and may thus be inconsistent if teacher assignment is related to students' prior test scores. In contrast, OLS estimation of the regression of student achievement on teacher fixed effects and control variables, including lagged student achievement scores, is consistent even when nonrandom assignment based on lagged achievement generates correlation between the teacher assignment dummies and the control variables. In other nonrandom assignment scenarios in which both the URM and OLS are inconsistent (e.g., assignment based on unobserved student heterogeneity), OLS performs at least as well as the URM.

Our theoretical discussion below provides a detailed explanation of these things specific to the URM and also describes how the URM relates to more easily understood estimators. We also provide simulation evidence to illustrate implications of a key theoretical feature of the URM (random versus fixed teacher effects) under various teacher assignment scenarios. We further apply the URM and other estimators to our administrative data, using each estimator on a series of specifications that sequentially become more similar (in terms of included covariates) to the URM. We find that controlling for the same set of scores as

those included in the URM produces very similar results to the URM, despite the URM's complicated steps taken to incorporate students with partially missing records. Given that the EVAAS methods are currently (and have previously been) used in teacher evaluation programs in several states, we should understand the methods and how they relate to other more easily understood (and more easily implemented) approaches. Our study aims to facilitate such understanding, especially with respect to the EVAAS URM.

3 Value-Added Models

Teacher value-added models (VAMs) are generally derived from a formulation that posits that academic achievement at any point in time is a function of all current and past child, family, and school inputs (e.g., Hanushek, 1979):

$$A_{it} = f(E_{it}, ..., E_{i0}, X_{it}, ..., X_{i0}, c_i, u_{it}) \tag{1}$$

where A_{it} is current achievement at time t for student i, E_{it},...,E_{i0} represent current and past education (school) inputs, X_{it},...,X_{i0} represent current and past student or parent inputs, c_i is unobserved student heterogeneity (e.g., motivation or some form of time-invariant innate ability), and u_{it} is an idiosyncratic error term. Given that we cannot measure each of these elements during each time period—at least not in available data—researchers typically adopt a more parsimonious model with a simple (estimable) functional form. For example, with a set of simplifying assumptions, a standard reformulation of the general model is often something like the following estimating equation:

$$A_{it} = \tau_t + \lambda A_{it-1} + X_{it}\beta + E_{it}\gamma + c_i + e_{it} \tag{2}$$

where τ_t allows for a different intercept in each time period to capture time (e.g., year) effects, A_{it} is the current test score at time t, A_{it-1} is the lagged test score from the previous year, E_{it} is a vector of *observed* education inputs at time t (e.g., teacher assignment indicators), and X_{it} is a vector of observed individual student characteristics.

The simplifying assumptions that facilitate the transition from equation (1) to equation (2) include linearity and geometric decay in the parameters; see Todd and Wolpin (2003) and Guarino, Reckase, and Wooldridge (2015) for a detailed discussion and derivations. Individual student heterogeneity, c_i, is generally left in the error term in commonly used approaches, although some of this is likely captured by A_{it-1}. While there are methods to completely eliminate this term in panel data settings (e.g., adding student indicators, or fixed effects estimation), we seldom compute teacher value-added measures with enough years of data on the same students to accurately identify these individual student effects.[5] Rather, teacher effects are sometimes obtained using up to a few years of data on *teachers* (so multiple cohorts of students). With the URM, teacher effects are estimated using one cohort of students, and in some cases composites of these effects are computed across years or subjects.

Even with this relatively parsimonious model in (2), administrative data may be missing test scores or characteristics for some students, or some students may not be linked to teachers. In traditional regression analysis such as OLS estimation, student observations missing these data are omitted from the estimation sample (or must be imputed), but consistent estimates can still be obtained. For consistency, whether data on the outcome or the regressors are observed or missing for a student can be related to the observed covariates that we control for (e.g., the lagged score, A_{it-1}, or student characteristics, X_{it}) but not unobserved elements of the error term (see Wooldridge, 2010, Ch 19). This consistency condition is nearly identical in practice to the "missing at random" (MAR) assumption that the EVAAS URM

[5]Such approaches actually performed quite poorly in the simulations conducted in Guarino, Reckase, and Wooldridge (2015). See the paper for details on the reasons for this for each grouping/assignment scenario.

relies on. For the URM, the pattern of missing scores can be related only to the lagged test scores (underlying the composite score) that are fully observed; by design of the URM, this is the same lagged scores that we control for in our preferred OLS approach.

3.1 Common Methods for Estimating Teacher Effects

Given that the student heterogeneity term in equation (2) is generally ignored when estimating value-added models, the estimating equation for a given subject s can be written as:

$$A_{ist} = \tau_t + \lambda A_{ist-1} + X_{it}\beta + E_{ist}\gamma + v_{ist} \tag{3}$$

where $v_{ist} = c_i + e_{ist}$ is the composite error term. OLS on this equation will estimate teacher effects, $\hat{\gamma}$. We call this estimator $DOLS$, as done by Guarino et al. (2015), to reflect the OLS estimation of the teacher effects and acknowledge the dynamic (D) specification including the lag score on the right-hand side. This can easily be extended to incorporate multiple lagged scores in multiple subjects. With this approach, to consistently estimate the vector γ, we need teacher assignment (E_{ist}) to be uncorrelated with the student heterogeneity term, c_i. This means, for example, that principals cannot assign students with higher (or lower) unobserved ability to more effective teachers. However, it is important to note that the estimates of the vector γ are adjusted for correlation of teacher assignment with prior test scores and other observable characteristics and that this feature of the model may go a long way toward mitigating bias due to nonrandom assignment of students to teachers (Guarino, Reckase, and Wooldridge, 2015).

Two widely used estimation methods initially omit the teacher assignments (E_{ist}) from the regression in (3) partialling out the effects of past test scores and student covariates on the current test score. The student-level residuals from this regression are used to estimate teacher effectiveness. The problem is that these methods do not adjust for the potential cor-

relation between the teacher assignments and the lagged test scores or student characteristics.

The first of these approaches estimates the abbreviated version (omitting E_{ist}) of equation (3) via OLS and then calculates the teacher effects as the within-teacher averages of the student-level OLS residuals. We refer to this as the average residual (AR) method. Again, consistency (as the number of students per teacher grows) requires that teacher assignment is not based on the student heterogeneity. However, also note that any correlation between the lagged test score A_{ist-1} and teacher assignment is not being partialled out of the teacher effects, so assignment based on prior scores also becomes problematic.

A second such approach, which we will abbreviate to *EB*, involves empirical Bayes' estimation of this more parsimonious equation, obtaining the teacher effects from the shrunken residuals. The empirical Bayes' method is essentially a GLS or random effects approach, where the teacher effect estimates are effectively "shrunken" towards the mean teacher effect (Guarino et al., 2015).[6] The so-called shrinkage takes teachers' class sizes into account, and thus aims to reduce the noisiness of the estimates from a small number of observations contributing to the estimation of the teacher effects. Like the AR method, consistent estimation relies on teacher assignment being uncorrelated with student heterogeneity and student-level covariates contained in the model (including prior achievement). The latter assumption is also relevant to the EVAAS URM approach we focus on in this paper and describe in the next section.

[6] As described in Guarino et al. (2015), this method involves two stages, but is easily implemented in Stata with the "xtmixed" command specifying a random component at the teacher level, and then post-estimation using the "predict , reffects" command to get the teacher random effects. The first stage estimates the normal maximum likelihood (with the random teacher effects in the error term) and the second stage applies the shrinkage factor to these teacher effects.

3.2 EVAAS Univariate Response Model (URM)

Similar to the OLS and EB approaches discussed above, the URM estimates teacher effectiveness for a single grade and subject (e.g., 5th grade math). There are two key differences between the common approaches just described and the URM. First, the URM uses prior test scores from multiple years and subjects to mitigate the influence of measurement error and does not include other student characteristics affecting current achievement. Second, the URM allows for students to be missing some of these prior test scores. The URM's strategy for allowing incomplete test score data generates the complex nature of the approach, but the complicated steps do not necessarily develop a more robust estimator, as we will demonstrate below.

For instance, the consistency conditions regarding the nature of the missing test scores allowed by the URM are nearly identical to the assumptions needed for OLS estimates to be consistent. And, when there are no missing data, there is a direct relationship between the URM and simpler standard linear regression techniques. Consider the simplest case where students have no missing test score data, students are randomly assigned to teachers, teachers have identical class sizes, and estimation is based on one cohort of students for teachers. Then the teacher effect estimates from the URM are nearly identical to OLS estimates, and would be identical if OLS (instead of EB) were used in the final estimation step. When students are nonrandomly assigned to teachers based on the included prior test scores, the URM and OLS estimates diverge. OLS partials out this assignment mechanism and consistently estimates the teacher effects while the URM assumes this assignment mechanism is not present and consequently produces biased estimates of the teacher effects. One goal of this paper is to derive and demonstrate these relationships.

In the discussion that follows, we first provide a detailed explanation of the URM approach—expanding on the descriptions in Wright, White, Sanders, and Rivers (2010) and SAS Institute Inc. (2015b)—and then illustrate how the URM compares with standard lin-

ear regression methods.

The URM estimating equation for subject s is:

$$A_{ist} = \tau + \kappa \hat{A}_{ist} + E_{ist}\gamma + \zeta_{ist} \tag{4}$$

where, compared to equation (3), the intercept τ does not have a time subscript since only one year of data is used, the lag score and student covariates have been replaced by a "composite score" $\hat{A}_{ist}$, and now the error term ζ_{ist} includes estimation error from using estimated components in $\hat{A}_{ist}$. The γ contains the random effect for the student's teacher. This equation is estimated using empirical Bayes' to obtain the teacher effects γ. Although this appears relatively simple, the composite score $\hat{A}_{ist}$ is the result of a multi-step process using all available lagged test scores (Wright et al., 2010), so the model is not as parsimonious as it appears. The composite score is essentially a different approach to a control, using multiple lagged test scores to predict a student's current score, and this prediction serves as a sort of sufficient statistic for the student's past inputs.

The URM uses multiple steps to compute the composite score, with each step performed separately for every year of data (i.e., student cohort) that contributes to the estimated teacher effects. In practice, the URM does not pool over these multiple years in estimation of teacher effects. In our analysis though, we do such pooling to illustrate how the performance of estimators changes as more years of data on teachers is incorporated. Thus, to estimate teacher effectiveness during a three-year period (i.e., based on three cohorts of students), each of the initial steps—up to and including computing the composite score—is done separately for the first, second, and third years of data. Then the final step—empirical Bayes' estimation of the teacher effects—is performed pooling the three years of data.

In computing the composite scores, the URM allows for many prior test scores across different subjects and years. For clarity, we focus our discussion on an example where we are using one-year and two-year lagged test scores for both reading (r) and math (m). The

URM computes a composite score in a specific subject (math shown in the equation below) as a linear combination of demeaned versions of the lagged test scores:

$$\hat{A}_{imt} = \hat{\mu}_{mt} + \hat{\beta}_{mt-1}\ddot{A}_{imt-1} + \hat{\beta}_{mt-2}\ddot{A}_{imt-2} + \hat{\beta}_{rt-1}\ddot{A}_{irt-1} + \hat{\beta}_{rt-2}\ddot{A}_{irt-2}. \quad (5)$$

In this equation, $\ddot{A}_{ist-y}$ (for the one-year and two-year lagged scores in subject s) denotes a "demeaned" y-year lagged test score in subject s for student i,

$$\ddot{A}_{ist-y} = A_{ist-y} - \hat{\mu}_{st-y}. \quad (6)$$

In equations (5) and (6), the estimated means $\hat{\mu}_{st-y}$ are not the overall means of the test scores. Rather, each $\hat{\mu}_{st-y}$ (including y=0 for the current score) is the sum of two components: an average across teachers of the teacher-level mean score and an adjustment to account for students with missing test score data. We discuss each of these components in further detail below.[7]

The weights in the composite score equation, $\hat{\beta}_{st-y}$, are coefficient estimates that maximize the correlation between the lagged scores and current score. With no missing data, $\hat{\beta}$ is essentially a vector of OLS coefficient estimates from the regression of A_{imt} on an intercept, A_{imt-1}, A_{imt-2}, A_{irt-1}, A_{irt-2}, and teacher assignment indicators. So, this step would produce coefficients on the lags from a DOLS-type equation that includes lagged test scores in multiple subjects, where teacher assignment is partialled out of the coefficient estimates.

Rather than use regression, however, the URM takes a different approach to estimation

[7]The Stata code for estimation as described here is: `mi impute mvn` $\ddot{a}_{m0}$ $\ddot{a}_{m1}$ $\ddot{a}_{m2}$ $\ddot{a}_{r1}$ $\ddot{a}_{r2}$, `emonly`. Our description and implementation are based on the Wright et al. (2010) documentation which was in use when we began this study. More recently, Wright (2015, p.14) indicates that the URM uses an EM algorithm that is "modified to accommodate the nesting of students within teachers," but provides no further details on the modification except that it is no longer necessary to construct the $\ddot{a}_{sy}$. Rather, the A_{sy} can be used with the modified algorithm, though we are not aware of any available commands in Stata or SAS for this modified algorithm. Regardless, this change is a different way of accounting for students grouped within teachers that avoids some of the constructed scores and means, but would not have a meaningful impact on any of our theoretical or empirical results.

to allow for certain patterns of missing data. In general, the URM requires a minimum of three lagged scores and one of these must be the most recent lag in the same subject as the dependent variable. In our example, this means students must have records for A_{imt-1} and at least two scores out of the set of $\{A_{imt-2}, A_{irt-1}, A_{irt-2}\}$. The URM uses the EM Algorithm to estimate a variance-covariance matrix, $\mathbf{C}$, for calculating the coefficients $\hat{\beta}$ (rather than estimating these directly with a regression, which would omit observations with missing data). Still, for the EM Algorithm to consistently estimate the variance-covariance matrix, the pattern of missing scores can only be related to the fully observed variable (and not omitted or unobserved variables nor the values of the partially missing scores), which is the same lagged test score included in the DOLS specification.

The EM Algorithm estimation step of the URM is done separately for each year of data. It uses a transformation of the current and lagged test scores where the *teacher*-level means are subtracted from each score so that $\mathbf{C}$ is a "within-teacher" variance-covariance matrix. We denote these transformed scores used for the EM Algorithm estimation as:

$$\ddot{a}_{isy} = A_{ist-y} - \hat{\mu}_{jst-y} \tag{7}$$

where $\hat{\mu}_{jst-y}$ is the average of A_{ist-y} across the students i assigned to teacher j.

Then the within-teacher variance-covariance matrix obtained via the EM Algorithm, for each year, is:

$$\mathbf{C} = \left[\begin{array}{c|c} c_{\ddot{a}_{m0}\ddot{a}_{m0}} & \mathbf{c}_{\ddot{a}_{sy}\ddot{a}_{m0}} \\ \hline \mathbf{c}_{\ddot{a}_{m0}\ddot{a}_{sy}} & \mathbf{C}_{\ddot{a}_{sy}\ddot{a}_{sy}} \end{array}\right] = \left[\begin{array}{c|cccc} c_{\ddot{a}_{m0}\ddot{a}_{m0}} & c_{\ddot{a}_{m1}\ddot{a}_{m0}} & c_{\ddot{a}_{m2}\ddot{a}_{m0}} & c_{\ddot{a}_{r1}\ddot{a}_{m0}} & c_{\ddot{a}_{r2}\ddot{x}_{m0}} \\ \hline c_{\ddot{a}_{m0}\ddot{a}_{m1}} & c_{\ddot{a}_{m1}\ddot{a}_{m1}} & c_{\ddot{a}_{m2}\ddot{a}_{m1}} & c_{\ddot{a}_{r1}\ddot{a}_{m1}} & c_{\ddot{a}_{r2}\ddot{x}_{m1}} \\ c_{\ddot{a}_{m0}\ddot{a}_{m2}} & c_{\ddot{a}_{m1}\ddot{a}_{m2}} & c_{\ddot{a}_{m2}\ddot{a}_{m2}} & c_{\ddot{a}_{r1}\ddot{a}_{m2}} & c_{\ddot{a}_{r2}\ddot{x}_{m2}} \\ c_{\ddot{a}_{m0}\ddot{a}_{r1}} & c_{\ddot{a}_{m1}\ddot{a}_{r1}} & c_{\ddot{a}_{m2}\ddot{a}_{r1}} & c_{\ddot{a}_{r1}\ddot{a}_{r1}} & c_{\ddot{a}_{r2}\ddot{x}_{r1}} \\ c_{\ddot{a}_{m0}\ddot{a}_{r2}} & c_{\ddot{a}_{m1}\ddot{a}_{r2}} & c_{\ddot{a}_{m2}\ddot{a}_{r2}} & c_{\ddot{a}_{r1}\ddot{a}_{r2}} & c_{\ddot{a}_{r2}\ddot{x}_{r2}} \end{array}\right] \tag{8}$$

where the first matrix shows subdivided "blocks" of the matrix (to be referenced below), with

$\ddot{a}_{sy}$ referencing the vector of lagged test scores in both subjects. The second matrix, with the lines for the subdivided blocks, is fully expanded to show each element of $\mathbf{C}$; the diagonal elements are the variance terms and the off-diagonal elements are the covariance terms.

The URM uses the elements of $\mathbf{C}$ to compute the set of within-teacher coefficient estimates, $\hat{\beta}_{st-y}$, by plugging into the familiar formula:

$$\beta_p = \mathbf{C}^{-1}_{\ddot{a}_{sy}\ddot{a}_{sy},p}\mathbf{c}_{\ddot{a}_{sy}\ddot{a}_{m0},p} \tag{9}$$

where p has been added to index each pattern of observed scores. With complete data for all students, the p index is not needed, and this equation would be equivalent to the OLS estimator from the regression of $\ddot{a}_{m0}$ on $\ddot{a}_{m1}$, $\ddot{a}_{m2}$, $\ddot{a}_{r1}$, $\ddot{a}_{r2}$ (or, equivalently, with the original scores, from the regression of A_{mt} on A_{mt-1}, A_{mt-2}, A_{rt-1}, A_{rt-2}, an intercept, and teacher assignment indicators).[8] When students have incomplete records though, the formula in (9) allows us to separately estimate a unique vector of coefficients, $\hat{\beta}_p$, for each pattern of observed scores, using the subset of matrix $\mathbf{C}$ corresponding to the relevant observed scores. So, in our example, given that the first lag of the math score must be present, we would compute up to four vectors $\hat{\beta}_p$ to account for different missing scores. We could consider $p = 0$ for complete records, $p = 1$ for records missing A_{mt-2}, $p = 2$ for records missing A_{rt-1} and $p = 3$ for records missing A_{rt-2}. For students with $p = 0$ the full matrix is used, while for students with $p = 1$ (missing A_{mt-2}) the 3rd row and 3rd column are dropped.

The EM Algorithm estimation also produces means that contribute to the $\hat{\mu}_{st}$ in the composite score equation and the $\hat{\mu}_{st-y}$ underlying the transformed scores ($\ddot{A}_{ist-y}$) in (6). To be clear, in equations (5) and (6), the estimated mean is $\hat{\mu}_{st-y} = \hat{\mu}^{mtm}_{st-y} + \hat{\mu}^{EMm}_{st-y}$, which is not the overall mean of the lagged test score. The first term on the right-hand-side is the mean-of-teacher-means $\hat{\mu}^{mtm}_{st-y}$ for the y-year lagged score in subject s. In other words, the mean lagged

[8] Other equivalent representations for the case of full data include $\beta = (\ddot{a}'_{sy}\ddot{a}_{sy})^{-1}\ddot{a}'_{sy}\ddot{a}_{m0}$, where $\ddot{a}_{sy}$ contains $\ddot{a}_{m1}$, $\ddot{a}_{m2}$, $\ddot{a}_{r1}$, $\ddot{a}_{r2}$, or $\beta = (X'X)^{-1}X'A_{mt}$ where X includes A_{mt-1}, A_{mt-2}, A_{rt-1}, A_{rt-2}, an intercept and teacher assignment indicators.

test score is computed for each teacher and then the average over all teachers is taken.[9]

The second term on the right-hand-side is produced by the EM Algorithm. It is an adjustment to the mean of teacher means to account for missing data—i.e., students with incomplete records. Since the EM Algorithm estimation step uses demeaned test scores (specifically, the teacher-demeaned scores $\ddot{a}_{st-y}$), this term is zero when there is complete data for all students. But when some students are missing test scores (and thus not contributing to the mean-of-teacher-means for the missing score), the estimated $\hat{\mu}^{mtm}_{st-y}$ may be biased and the URM includes the mean $\hat{\mu}^{EMm}_{st-y}$ to reduce potential bias from missing lagged scores.

The transformation in (6) that subtracts these two mean components is similar to removing year effects, which would be done by instead subtracting the overall mean (or by including year dummies in a regression). Subtracting the mean-of-teacher-means ($\hat{\mu}^{mtm}_{st-y}$) instead ensures that the "average" teacher has a teacher effect of zero and the EM Algorithm component ($\hat{\mu}^{EMm}_{st-y}$) corrects for potential bias in the mean-of-teacher-means from students missing test scores (Wright et al., 2010).

Finally, we compute the so-called *composite score*, $\hat{A}_{imt}$, according to equation (5). The composite score is the sum of the "adjusted mean" of the current math score ($\hat{\mu}_{st} = \hat{\mu}^{mtm}_{st} + \hat{\mu}^{EMm}_{st}$) plus a weighted average of transformed lagged scores $\ddot{A}_{st-y}$, with the weights being the coefficient estimates, $\hat{\beta}_p$. The composite score is a prediction of the current score (A_{imt}) based on the student's past test scores and assuming the student has the "average" teacher in the current year (Wright et al., 2010).

After the composite scores are obtained, the final step in computing the teacher effects is the empirical Bayes' estimation of equation (4)—as mentioned above. Although this is the final step for obtaining teacher effect estimates, the multi-step process also complicates estimation of the variability of the estimates. The standard errors from this last step do not

[9] To the best of our knowledge—based on the description in Wright et al. (2010)—this average per teacher is across all of the teacher's students, even if the teacher has multiple classes. Regardless, this is not important for our theoretical or empirical results and conclusions.

account for the earlier estimation of elements of the composite score.

Note that this discussion has focused on estimating teacher effects for math teachers. If one wished to estimate teacher effectiveness in, say, reading, then the outcome variable would be the current reading score, and the composite score would constitute a predicted reading score. While the same lagged scores could be used to obtain the composite score, the estimated elements (i.e., the $\hat{\mu}_{st}^{mtm}$, $\hat{\mu}_{st}^{EMm}$, and $\hat{\beta}_{st}$) would be different because they would be based on predicting the current *reading* score, using the sample of students satisfying the corresponding data requirements. So, in this respect, the URM is similar to the common VAM approaches that estimate teacher effects separately by subject (and grade).

3.2.1 Relating the EVAAS URM to other approaches

Unlike traditional regression-based VAM methods, the EVAAS approach handles at least some missing data patterns. It also uses empirical Bayes' shrinkage in the final step in order to account for teachers having different numbers of students. But is EVAAS very different from the standard regression estimators? In practice, differences in the estimated teacher VAMs may be minor. In fact, in the simplest scenario the two approaches yield nearly identical (up to a constant) teacher effect estimates.

In the simplest setting, there are no missing data and only one year of data is used (i.e., one cohort of students). Either shrinkage is not used or the number of students per teacher is identical, in which case shrinkage simply multiplies all of the teacher VAMs by the same constant. With a single year of data, a simple extension of DOLS to allow other lagged test scores comes from OLS estimation of the equation

$$A_i = X_i\beta + E_i\gamma + v_i, \tag{10}$$

where X_i includes all lagged test scores in various subjects and E_i is the vector of teacher

assignment dummies. For simplicity, we drop the subscripts indicating subject and year. This specification is a modified version of our DOLS specification in (3), now augmented with additional lagged test scores in the same and other subjects and omitting student characteristics, reflecting the variables generally included in the URM approach.

From the Frisch-Waugh partialling-out theorem, the OLS coefficients on the lagged test scores, $\hat{\beta}$, can be obtained in three steps:

(i) Regress A_i on E_i and obtain the residuals, $\ddot{A}_i$. Now, $\ddot{A}_i = A_i - E_i\hat{\eta}$ where, because the E_i are teacher assignment dummies, $\hat{\eta}_j$ is the average of the A_i (current test score) for teacher j. Therefore, $\ddot{A}_i$ is student i's test score deviated from the average test score for the student's teacher.

(ii) Regress each lagged test score in X_i on E_i and collect the vectors of residuals, $\ddot{X}_i$. Just as with $\ddot{A}_i$, each element of $\ddot{X}_i$ is one of student i's lagged test scores deviated from the mean for student i's teacher.

(iii) Run the regression of $\ddot{A}_i$ on $\ddot{X}_i$ and obtain $\hat{\beta}$.

In other words, when the regression is restricted to a single year, and there are no missing data, the OLS and URM estimates of β are identical; the URM simply partials out teacher assignment in a separate step, rather than using the full regression in (10).

As described earlier, the next step in the URM is to construct the composite score in equation (5). But the composite score $\hat{A}_i$ can be written as

$$\hat{A}_i = X_i\hat{\beta} + \hat{\psi}, \tag{11}$$

where $\hat{\psi}$ depends on $\hat{\beta}$ and the overall means of the test scores. Now, the equation used to obtain the teacher effects is

$$A_i = \kappa\hat{A}_i + E_i\gamma + \omega_i, \tag{12}$$

where the error term ω_i includes estimation error because $\hat{A}_i$ depends on $\hat{\beta}$.

Without missing data, we know, by the algebra of OLS, what will happen if we apply OLS to equation (12): $\hat{\kappa} = 1$ and $\hat{\gamma}$ will be identical to what is obtained from the long regression in (10). The argument is simple. We know the OLS estimates minimize the sum of squared residuals, and we know the $\hat{\beta}$ from the URM is identical to the $\hat{\beta}$ from OLS. So one cannot do any better by choosing $\hat{\kappa}$ different from unity and $\hat{\gamma}$ as the DOLS coefficients. The additive constant in (11) changes nothing because the DOLS regression, with a full set of teacher dummies, effectively estimates an intercept.

The URM approach is not to apply OLS to equation (12)—otherwise it would just be DOLS. Instead, the URM method applies empirical Bayes' to (12), which treats the teacher effects as a random component of the error term and then obtains the teacher effects with a prediction formula that involves shrinking the estimates of γ towards the average teacher effect. When the coefficient on the composite score is estimated by empirical Bayes', the coefficient is not unity, which breaks the equivalence and causes bias. This bias stems from the random effects estimator (not the shrinkage) underlying empirical Bayes' estimation, a result that has been shown more generally by Guarino et al. (2015), but is less obvious here due to the URM's complicated multistep method. The random effects estimator underlying empirical Bayes' assumes E_i is uncorrelated with the composite score estimated in (11). The bias from this assumption is precisely what was shown in Guarino et al. (2015), only now the lagged test score(s) is replaced with the composite score. Hence, it is the underlying random effects estimator applied to (12) that biases estimates of both κ and γ; the shrinkage typically has a minor effect (Guarino et al., 2015). Conversely, when students are not assigned to teachers based on any included or omitted test scores or characteristics—so E_i is not correlated with γ or other components of the error term—the URM estimates are nearly identical to the OLS estimates from (12). Or, if equation (12) were estimated by OLS instead of empirical Bayes', then the URM and OLS estimates from (10) would be the same

(regardless of any correlation between E_i and γ).

4 Simulation

4.1 Simulation Design

We conduct simulations to assess the performance of the DOLS, EB, AR, and URM estimators under various student grouping and assignment scenarios. This allows us to know the "true" teacher effect (which we generate), and then evaluate the ability of each of the estimators to capture this effect—something not possible with administrative data. Much of our simulation design is similar to that from Guarino et al. (2015), which focused on evaluating the performance of EB along with DOLS and AR. We generate data for three cohorts of 800 students each, creating a current score and two lagged scores for each student. For our analysis, we focus on a single grade, so using one observation per student, but three cohorts of students per teacher. The simulations are designed with elementary grades in mind, so we can think of this setting as 5th grade students and teachers. Class size is set to 20, for a total of 40 teachers.

To generate the test scores, we first obtain a baseline score (i.e., the first grade tested) drawn from a standard normal distribution. Each of the three subsequent test scores, A_{it}, is then generated according to the equation

$$A_{it} = \lambda A_{i,t-1} + \gamma_{it} + c_i + u_{it}, \tag{13}$$

where $A_{i,t-1}$ is lagged achievement, γ_{it} is the teacher contribution to the current score (the true teacher effect), c_i is the time-constant unobserved student effect, and u_{it} the idiosyncratic error. The decay parameter, λ, is set to either 0.5 (substantial decay of student achievement)

or 1 (no decay).[10] The correlation between the baseline score and the student fixed effect is 0.5. The three random parameters are drawn from normal distributions: student fixed effect $c_i \sim N(0, .5^2)$, teacher effect $\gamma \sim N(0, .25^2)$, and the idiosyncratic error $u_{it} \sim N(0, 1)$ (so their respective proportions of the total variance in test scores are 19%, 5%, and 76%).

To look at nonrandom sorting of students, we make the distinction between *grouping* (how students are grouped into classrooms) and *assignment* (how students are assigned to teachers), allowing for students to be, say, grouped based on prior achievement levels but then randomly or nonrandomly assigned to teachers. We look at grouping based on the lagged score (referred to as dynamic grouping), the original baseline score (a form of "static" grouping referred to as baseline grouping), and the student individual heterogeneity (another form of static grouping, referred to as heterogeneity grouping). We look at three different assignment mechanisms for each of these grouping scenarios: random assignment, positive assignment (e.g., better students to better teachers), and negative assignment (e.g., struggling students to better teachers). The nonrandom assignment cases are not perfectly separating students in rank order of, say, lagged achievement; rather assignment is noisy with the noise being drawn from a standard normal distribution. We conduct 100 Monte Carlo repetitions for each grouping-assignment-parameter scenario.

Given that a notable feature of the URM is the handling of missing data, we also simulate data where some students are missing test scores, following the method used by Wright (2004). We first rank students based on their prior test score (calling this the $TrueRank$). Then we generate a random student rank ($RandomRank$). To decide which scores to set to missing, we create a linear combination of these: $SortRank = a * (TrueRank) + (a - 1) * (RandomRank)$. Sorting on this value, we then replace the lagged score with missing for the bottom 20 percent of students. For randomly missing scores, we set $a = 0$. For non-randomly missing scores (that still satisfy the MAR assumption discussed earlier), we set $a = .4$.

[10] The decay parameter reflects the fact that a student's prior score may not be fully additive into the current score, whether due to changes in the test, what the test measures in one year versus the next, or student "learning loss" from one test period to the next.

We examine the performance of four of the estimators discussed above (DOLS, AR, EB, EVAAS URM). For the first three estimators we consider a "common" specification, similar to equation (3), where the covariates include a lagged test score, and in the case of DOLS, teacher assignment indicators. (We do not incorporate effects for student characteristics into the simulation.) For the URM, we base the composite score on this same lagged test score as well as a two-year lagged test score.

As discussed above, we also estimate specifications that "mimic" the EVAAS URM approach, using DOLS, AR, and EB, to illustrate where divergences in the performance of the estimators is coming from. Hence, for the simulations, this means including both the one-year and two-year lagged test scores in the estimating equation. For all estimators and specifications, we estimate the "5th grade" teacher effects first using one cohort (year) of data, and then estimate them pooling over three cohorts.

Our first metric for evaluating the performance of these estimators is the Spearman rank correlation between the estimates and the true teacher effects, to examine their ability to uncover the correct rankings of the true effects.[11] For an additional viewpoint, we also provide the mean squared error (MSE), which illustrates the tradeoff between bias and variance. However, given the extensive policy emphasis on ranking and categorizing teachers by the continuum of estimated effectiveness, we focus much of our discussion on the rank correlations which provide a metric for identifying relative effectiveness, which is what policy seems to be after.

4.2 Simulation Results

We first assess the ability of each of the estimators to uncover the true teacher effect, looking at the correlations between the estimated and true effects. For our main results, we focus

[11] We also have results using simple (Pearson) correlations, which follow the same patterns as the rank correlations.

on the "small" teacher effects, which account for 5 percent of the variation in test scores, in the case where $\lambda = .5$ (substantial decay of student achievement). In practice, the URM uses one year of data (i.e., one cohort of students) to estimate teacher effectiveness, so Table 1 provides the rank correlations between the true teacher effects and the estimated teacher effects (Panel A) in this setting, along with the mean squared error (Panel B). For this and the following two tables, all 10 grouping-assignment scenarios are explored. The estimators considered first are DOLS, AR, and EB on the "common" specification that controls only for one lag score (in addition to the teacher effects in the case of DOLS). The next set of columns are based on DOLS, AR, and EB estimation of specifications that also include a two-year lagged score, to "mimic" the information in the composite score of the URM.

Table 1 shows that under random grouping and random assignment, the rank correlations are 0.69 for all estimators, and nonrandom grouping does not cause large departures from this, as long as assignment to teachers is random. The estimators actually perform best in the positive assignment cases, in particular when grouping is based on the student heterogeneity, with rank correlations ranging from .78 to .80, a result arising from bias that expands the distribution of estimated teacher effects, making it easier to distinguish between teachers (see Guarino, Reckase, and Wooldridge (2015) for a more detailed discussion of this result). Conversely, the estimators perform worst when students are grouped on heterogeneity and then negatively assigned to teachers, with correlations ranging from .41 to .43.

Also evident in Table 1 is the close relationship between the EVAAS URM and using EB to estimate a specification with the same lagged test scores, as the correlations in the URM and EB-mimic columns are nearly identical. Further, we see that under dynamic grouping with positive or negative assignment, although all estimators perform worse relative to random assignment, DOLS performs substantially better than AR, EB, or URM. This was previously shown in Guarino et al. (2015) for EB and AR, and we show this extends to the URM. This result arises from the fact that these approaches are not correctly partialling

Table 1: Correlations and MSE (1 cohort of students and no missing data)

		"common" 1-yr lag score				"mimic" 1-yr and 2-yr lag scores		
Grouping	**Assignment**	**DOLS**	**AR**	**EB**	**URM**	**DOLS**	**AR**	**EB**
PANEL A - Spearman rank correlations								
Random	Random	0.69	0.69	0.69	0.69	0.69	0.69	0.69
Dynamic	Random	0.70	0.70	0.70	0.70	0.70	0.70	0.70
	Positive	0.67	0.49	0.53	0.53	0.68	0.50	0.53
	Negative	0.70	0.53	0.58	0.58	0.70	0.53	0.57
Baseline	Random	0.67	0.67	0.67	0.68	0.68	0.68	0.68
	Positive	0.75	0.72	0.73	0.71	0.73	0.69	0.71
	Negative	0.50	0.49	0.50	0.55	0.55	0.53	0.55
Heterogeneity	Random	0.64	0.64	0.64	0.65	0.64	0.65	0.65
	Positive	0.80	0.79	0.79	0.79	0.79	0.78	0.79
	Negative	0.41	0.41	0.41	0.43	0.43	0.43	0.43
PANEL B - Mean squared error								
Random	Random	0.057	0.057	0.032	0.032	0.057	0.056	0.032
Dynamic	Random	0.059	0.057	0.031	0.031	0.058	0.056	0.031
	Positive	0.060	0.070	0.044	0.044	0.059	0.069	0.044
	Negative	0.060	0.067	0.041	0.041	0.059	0.066	0.041
Baseline	Random	0.063	0.062	0.033	0.032	0.060	0.058	0.032
	Positive	0.062	0.060	0.029	0.030	0.060	0.058	0.031
	Negative	0.070	0.071	0.045	0.042	0.065	0.067	0.043
Heterogeneity	Random	0.073	0.071	0.037	0.036	0.071	0.068	0.036
	Positive	0.072	0.068	0.030	0.029	0.070	0.065	0.029
	Negative	0.079	0.080	0.050	0.049	0.076	0.077	0.049

Notes: Panel A provides the Spearman rank correlations with the true teacher effects, Panel B the mean squared error. These results are based on simulations with small teacher effects and 1 cohort of students with $\lambda = .5$.

out the assignment mechanism from the teacher effects. EB and the URM both are closer to DOLS than AR, though, because as the number of students per teacher gets larger, the empirical Bayes' estimates of the teacher effects (underlying EB and URM) will get closer to DOLS (see Guarino et al. (2015) for a more detailed discussion of this result that the random effects estimates will converge to the fixed effects estimates as the sample size increases).

The rank correlations in Panel A reflect the bias of the estimators in how well they correctly order teachers. Panel B provides another performance metric—the mean squared

error (MSE)—which reflects the bias and variance of the estimators ($MSE = variance + bias^2$). With one cohort of students, we see that the URM and EB have nearly identical MSE across assignment scenarios, which are generally smaller than AR and DOLS, reflecting the smaller variance of these estimators. However, in some cases (such as positive or negative assignment based on the lagged test score), this smaller variance comes at the cost of having more error in ranking teachers.

Table 2: Correlations and MSE (3 cohorts of students and no missing data)

		"common" 1-yr lag score				"mimic" 1-yr and 2-yr lag scores		
Grouping	**Assignment**	**DOLS**	**AR**	**EB**	**URM**	**DOLS**	**AR**	**EB**
PANEL A - Spearman rank correlations								
Random	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Dynamic	Random	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Positive	0.84	0.66	0.76	0.76	0.84	0.66	0.76
	Negative	0.83	0.68	0.77	0.77	0.83	0.68	0.77
Baseline	Random	0.82	0.82	0.82	0.83	0.83	0.83	0.83
	Positive	0.88	0.87	0.88	0.87	0.87	0.85	0.87
	Negative	0.65	0.65	0.65	0.71	0.72	0.70	0.71
Heterogeneity	Random	0.81	0.81	0.81	0.82	0.82	0.82	0.82
	Positive	0.89	0.89	0.89	0.89	0.89	0.89	0.89
	Negative	0.52	0.52	0.52	0.55	0.56	0.55	0.55
PANEL B - Mean squared error								
Random	Random	0.021	0.021	0.016	0.017	0.020	0.021	0.016
Dynamic	Random	0.021	0.022	0.016	0.017	0.021	0.021	0.016
	Positive	0.021	0.034	0.023	0.024	0.021	0.034	0.023
	Negative	0.021	0.031	0.022	0.023	0.021	0.031	0.022
Baseline	Random	0.023	0.024	0.017	0.018	0.022	0.023	0.016
	Positive	0.024	0.022	0.015	0.015	0.021	0.021	0.014
	Negative	0.032	0.033	0.032	0.028	0.027	0.029	0.027
Heterogeneity	Random	0.026	0.027	0.019	0.020	0.025	0.026	0.019
	Positive	0.036	0.033	0.023	0.022	0.034	0.030	0.021
	Negative	0.043	0.045	0.040	0.039	0.040	0.042	0.038

Notes: Panel A provides the Spearman rank correlations with the true teacher effects, Panel B the mean squared error. These results are based on simulations with small teacher effects and 3 cohorts of students with $\lambda = .5$.

Although using one cohort of students is convenient, in practice multiple cohorts are often used, so we also present results from using three cohorts of students (i.e., three years of data

on teachers) to estimate teacher effectiveness. Given that this is increasing the amount of information on teachers (and teacher effects do not vary by year in our simulation), we expect the performance of all of the estimators to improve. The rank correlations in Table 2 show this improved performance, but the results also follow the same relative performance patterns across scenarios and estimators. The correlations under random grouping and random assignment are now larger at .84. In the case of grouping based on student heterogeneity with positive assignment to teachers, the correlations are now .89 for all estimators. When students are instead negatively assigned to teachers (based on heterogeneity), the correlations are .52–.56. Under this scenario, the correlations for the "mimic" specification estimators are slightly larger than those from the "common" specifications; this result comes from the amount of decay in student achievement with $\lambda = .5$, so adding additional lag scores helps. Again we see the nearly identical performance of the URM and EB-mimic. The issue of poor performance of AR, EB, and URM under nonrandom assignment based on the lagged score remains. Again, the URM and EB estimators perform more similarly to DOLS than AR exhibiting the convergence of the random effects approach (EB, URM) to the fixed effects approach (DOLS). AR performs the worst because the assignment mechanism is not partialled out at all.

When we turn to Panel B, we see that all estimators perform very similarly in terms of MSE now, as expected. With more data on teachers, the MSE converges as the variance of all estimators gets smaller, but the bias does not shrink. So the efficiency gain from using the URM or EB diminishes, while the advantages of using DOLS to correctly rank teachers remains, as we discussed for Panel A.

All of the simulation results thus far have been based on the ideal situation with no missing data. The rank correlations in Table 3 are based on the analogous sorting and assignment scenarios, only now with 20 percent of students missing the one-year lagged test score. When we introduce missing data, AR, EB, and DOLS tend to do slightly worse (rank correlations fall by about .02–.06). The performance of the URM behaves differently;

sometimes improves, sometimes worsens, and in some cases by large magnitudes in either direction.

Under random grouping and assignment, the URM performs slightly better than DOLS. We see similar performance under the other random assignment scenarios (with various nonrandom grouping). The URM again shows slight advantage under most of the positive assignment scenarios, with the largest of these being for dynamic grouping. However, the URM performs significantly worse under many of the negative assignment scenarios, with the largest gap occurring when grouping is based on the one-year lag score (the URM correlations is .32 compared to .64 for DOLS). With three cohorts of students, we see some convergence in performance of the estimators as all of them perform better with the additional data. However, the URM still does significantly worse under negative assignment based on the one-year lag score (now with a correlation of .46 compared to .82 for DOLS).

Overall, the URM does show slight advantage in some of the grouping and assignment scenarios, especially when only one cohort of students is used. However, when students with lower lagged test scores are assigned to better teachers—which is certainly a plausible assignment scenario—the URM does much worse at ranking teachers.

4.3 Sensitivity of Simulation Results

While some sensitivity analyses were presented with the main results (e.g., using one versus three cohorts of students, or having missing versus non-missing data), we also conducted simulations with various modifications. These alternate scenarios include increasing the number of repetitions to 500, having larger teacher effects, choosing $\lambda = 1$, generating other missing data scenarios, varying class sizes, using additional lagged test scores in the "same" and "other" subjects, and scenarios with grouping and assignment based on a composite of these scores. (All sensitivity results are available upon request.)

Table 3: Correlations (1 and 3 cohorts of students, with missing data)

		"common" 1-yr lag score				"mimic" 1-yr and 2-yr lag scores		
Grouping	**Assignment**	**DOLS**	**AR**	**EB**	**URM**	**DOLS**	**AR**	**EB**
PANEL A - 1 cohort of students								
Random	Random	0.65	0.65	0.65	0.67	0.65	0.65	0.65
Dynamic	Random	0.63	0.63	0.64	0.61	0.64	0.64	0.64
	Positive	0.64	0.46	0.50	0.74	0.64	0.46	0.50
	Negative	0.63	0.49	0.52	0.32	0.64	0.49	0.52
Baseline	Random	0.66	0.66	0.66	0.70	0.67	0.67	0.68
	Positive	0.71	0.69	0.70	0.72	0.69	0.65	0.67
	Negative	0.47	0.47	0.47	0.54	0.51	0.50	0.51
Heterogeneity	Random	0.60	0.61	0.61	0.61	0.61	0.61	0.61
	Positive	0.78	0.77	0.77	0.80	0.77	0.76	0.77
	Negative	0.38	0.37	0.39	0.36	0.40	0.39	0.41
PANEL B - 3 cohorts of students								
Random	Random	0.81	0.81	0.81	0.84	0.81	0.82	0.81
Dynamic	Random	0.82	0.82	0.82	0.82	0.82	0.82	0.82
	Positive	0.82	0.63	0.72	0.89	0.82	0.63	0.72
	Negative	0.82	0.66	0.74	0.46	0.82	0.66	0.74
Baseline	Random	0.80	0.80	0.80	0.83	0.81	0.81	0.81
	Positive	0.86	0.84	0.86	0.86	0.85	0.82	0.84
	Negative	0.62	0.61	0.61	0.69	0.69	0.66	0.68
Heterogeneity	Random	0.79	0.79	0.79	0.80	0.79	0.79	0.79
	Positive	0.89	0.88	0.89	0.90	0.89	0.88	0.88
	Negative	0.49	0.49	0.49	0.48	0.52	0.52	0.52

Notes: This table provides the Spearman rank correlations with the true teacher effects when 20 percent of students are (randomly) missing the one-year lag test score. These results are based on simulations with small teacher effects and $\lambda = .5$.

Increasing the number of repetitions to 500 does not produce any meaningful changes to the rank correlations. In the case of larger teacher effects with $\gamma \sim N(0, .6^2)$ and $c_i \sim N(0, .5^2)$, the teacher effect and the student heterogeneity each account for about 21% of the total variation in test scores. With the teacher effects accounting for more of the variation in the test scores, we naturally expect the performance of the estimators to improve, and we do find this to be the case. But the results follow the same general patterns discussed for the small teacher effects case. Choosing $\lambda = 1$ also had little impact on the results, which exhibit the same general patterns. The main difference is that the performance of estimators on the

"mimic" specifications is no different than for the "common" specifications since there is no motivation for including the two-year lag score when $\lambda = 1$.

The missing data scenario presented in Table 3 involved 20 percent of students chosen randomly to be missing the one-year lag test score. When instead we chose students somewhat non-randomly (setting $a = .4$) to be more likely to have a missing test score if they were lower achieving students, we still find similar patterns of results. This is not surprising given that this case still satisfies the MAR assumption that the estimators rely on, since the selection is on an observed (and included) score. When we instead chose to set the two-year lag score to missing for these scenarios, the results differ. The performance differences across estimators are much smaller. The URM again does slightly better in many cases, but DOLS does slightly better under both types of nonrandom assignment based on dynamic grouping (positive and negative assignment).

To vary class size, we randomly assigned 36 teachers class sizes of 10, 20, or 30 students (12 teachers for each class size). For the simulations with three cohorts of students, class size was the same for a teacher in all cohorts. The rank correlations for all estimators are slightly smaller for all scenarios except for the negative assignment scenarios. In these cases, all estimators improve. For assignment on the lagged score, the performance of DOLS remains similar, but the other estimators improve, bringing them closer to DOLS. Overall though, the patterns of performance are similar to the simulation with uniform class sizes of 20 students.

Incorporating further lagged scores or lagged scores in other subjects does not contribute substantively to our evaluation of the theoretical implications of sorting or assignment for the URM, as these constitute the same issues as having one vs. two lags. Thus, our main results focused on the simple case of two lags to facilitate transparency in our simulation design and results. But we also performed simulations with an additional lag as well as with multiple lags in another subject, and further incorporated an additional sorting mechanisms

based on a combination of these scores (with more weight on the one-year and two-year lags in the "original" subject).

Adding the three-year lag score does not affect performance under the random assignment scenarios. However, under positive assignment the rank correlations decrease (more so for AR, EB), while under negative assignment they increase (more so for DOLS, URM). These results follow the same pattern as going from including only a one-year lag to also including a two-year lag. Under positive assignment, all of the estimators are biased upward, while they are biased towards zero under negative assignment. Adding the three-year lag reduces the bias (in both cases), so the performance is converging. This result is more pronounced for baseline grouping because the three-year lag score is the baseline score in our simulation.

The scores in the "other" subject were generated in the same fashion as our original scores, with a baseline score drawn from the same distribution as the original baseline score, and with a correlation between these scores of .7 (chosen based on our administrative data). We see similar patterns to adding the three-year lag score, with no meaningful changes for random assignment, and now only small decreases for positive assignment and small increases for negative assignment (primarily only under heterogeneity or composite grouping).

When we extend dynamic grouping to a grouping scenario based on a composite of all of these prior scores, estimators still perform similarly under random assignment. However, as expected, the performance of DOLS is affected similarly to the other estimators when the specification does not include all of the scores underlying the grouping-assignment mechanism. So for all estimators, we see higher rank correlations under positive assignment and lower rank correlations under negative assignment. And, DOLS still performs better than the other estimators even when we do not control for all of the lag scores.

5 Administrative Data

5.1 Data

We use administrative data on students in grades 5 and 6 during years 2002–2007 in a large urban anonymous district.[12] Similar to our example used in the EVAAS URM discussion, we focus on math scores as the outcome and use one-year and two-year lagged math and reading scores as covariates in some specifications. The data contain information on student race/ethnicity, days absent, gender, disability, limited English proficiency (LEP), and free- or reduced-price lunch eligibility (FRL). We exclude students who are not linked to math teachers, students who are assigned to classes (i.e., teacher/year groups) with fewer than 10 students, and students who were retained. All estimations also require that students have, at a minimum, a current math score and a one-year lagged math score.

Average scores for the 5th and 6th grade samples are provided in Table 4 for the students with data satisfying the minimum sample inclusion requirements just described; these estimation samples cover years 2002–2007. The first set of descriptives in Panel A are for the sample of 5th grade students, while Panel B contains the descriptives for the 6th grade sample. As an illustration of how the samples could change depending on which lagged test scores are included, consider adding a two-year lagged math score in a regression. This would mean 3.1% of the 5th grade students are omitted. For 6th grade, the sample falls by 3.4% with the addition of two-year lagged math. This indicates that including a longer history of scores does impose data restrictions, though the URM is able to relax these restrictions somewhat.

[12] Our data sharing agreement does not allow us to name the district or state. To maintain anonymity of the district, we do not include observation counts or information on student demographics.

Table 4: Descriptive statistics for students in sample, by grade

	Mean	Std Dev	Min	Max
Panel A: Grade 5				
Math score	1638.62	232.26	569	2456
Reading score	1572.35	314.13	474	2713
1-yr lag Math	1485.78	254.84	569	2330
2-yr lag Math	1344.95	287.95	375	2225
1-yr lag Reading	1523.12	317.68	295	2638
2-yr lag Reading	1297.28	350.45	86	2514
Panel B: Grade 6				
Math score	1652.63	242.93	770	2492
Reading score	1635.41	302.95	539	2758
1-yr lag Math	1634.20	220.28	569	2456
2-yr lag Math	1460.65	247.73	569	2330
1-yr lag Reading	1550.82	306.45	474	2713
2-yr lag Reading	1504.67	313.31	86	2638

5.2 Results

With the administrative data, we estimate teacher effects separately for 5th and 6th grade, focusing on math teachers only (so we use math scores as our outcome variable). Similar to the approach for the simulations, we use AR, DOLS, and EB to estimate several specifications. The first two specifications, which we refer to as "common" specifications, are based on equation (3). The first specification controls for the one-year lagged math score, year effects, and other student-level covariates (days absent, race/ethnicity, disability, LEP, FRL-eligibility, and female). The second specification is augmented with a two-year lagged math score also. The last two specifications are designed to be more similar to the EVAAS URM. The third specification omits student covariates but includes the same lagged scores as the composite score computed for the URM, hence attempting to "mimic" the information used in the URM estimation. The fourth specification uses the composite score itself as the only covariate (so when using EB estimation, this is identical to the URM). We compute estimates using one year of data or pooled over two years of data, covering the years 2002-2007. We then examine agreement among the estimators in each year and present results on aver-

age agreement during this time period.

In Table 5, we provide average Spearman correlations between the URM estimates and those from each of the other estimator/specification combinations. Within each specification, the rank correlations do not change significantly when pooling over an additional year of data for estimation and also do not differ substantially between estimators. In column [1], the correlations show that agreement with the URM is slightly better in the 6th grade analysis for all estimators, and there we also see that agreement is highest for EB, slightly lower for DOLS, and lowest for AR.

Table 5: Spearman rank correlations, comparing EVAAS URM to other estimators

	1-yr lag Math, Student Char. [1]	1-yr & 2-yr lag Math, Student Char. [2]	1-yr & 2-yr lags in Math & Reading [3]	Composite score [4]
Panel A: 5th grade				
1-year estimates				
DOLS	0.918	0.971	0.997	0.999
AR	0.922	0.972	0.995	0.998
EB	0.920	0.972	0.998	1.000
2-year estimates				
DOLS	0.918	0.971	0.997	0.999
AR	0.921	0.971	0.994	0.997
EB	0.920	0.972	0.998	1.000
Panel B: 6th grade				
1-year estimates				
DOLS	0.941	0.982	0.995	0.997
AR	0.931	0.964	0.987	0.990
EB	0.944	0.984	0.998	1.000
2-year estimates				
DOLS	0.945	0.982	0.995	0.997
AR	0.935	0.963	0.986	0.990
EB	0.948	0.984	0.998	1.000

Notes: This table provides the average rank correlation between the URM estimate and other estimator/specifications.

When we add a two-year lagged math score to the specification (column [2]), the rank correlations all increase substantially, to around .97 for 5th grade and slightly higher around .98 for 6th grade (with the exception of AR, which is lower at .96 for 6th grade). In column

[3] we omit student characteristics and use two lag scores each in reading and math, and now find even greater agreement with the URM estimates (correlations >.99).[13] In column [4], we use the composite score as the only regressor, and now the rank correlations are even higher. (The rank correlations for EB are exactly 1 because this is the URM approach itself.)

Within each grade/specification combination, the EB rank correlations are at least as large as those for DOLS or AR, which indicates that the estimation approach matters somewhat. However, the specification seems to be more important in our data. Agreement with the URM increases for all estimators as we get closer to using the same specification as the URM (moving left to right from columns [1]-[4]); when we use the composite score as the only regressor, all of the rank correlations are very close to 1. We find fairly high rank correlations between the URM and estimates from specifications that include student characteristics, consistent with the results found by Ballou, Sanders, and Wright (2004) for the MRM.

The results in column [3] also show that the differences between the URM and the regression based approaches using the same lag scores are not large. The complicated nature of the URM stems from taking extra steps to include students with certain patterns of partially missing test score records, since regression-based methods omit these students from estimation. Given that consistent estimation for DOLS and the URM requires very similar (if not identical) assumptions regarding the way in which data are missing, it is not surprising that the two approaches reach similar results. The estimates from simple DOLS estimation of a similar specification with teacher indicators correlates very highly (>.99) with the complicated multi-step EVAAS URM estimation.[14]

[13] We also estimated a specification that would fall between columns [2] and [3] of Table 5; the specification included student characteristics along with the two lag scores each in reading and math. The rank correlations for this specification were .98 for all estimators in both grades, falling appropriately between those shown in columns [2] and [3].

[14] A simple approach to handling missing data, which relies on the same MAR assumption as DOLS, is to replace a missing value (test score) with zero and include a dummy variable corresponding to that variable. The dummy takes on a value of one when such a replacement is made and zero otherwise. This avoids dropping the observation from from the regression, but the observation still does not contribute to estimation of the coefficient for the missing score. With this method, we get rank correlations nearly identical to those in column [3].

Although we cannot know the true teacher effects in the case of administrative data, the high agreement between DOLS and the URM along with EB and AR suggests that there may not be substantial nonrandom assignment in our data.[15] Our simulation results showed that with complete data, DOLS is robust to nonrandom assignment on lagged scores while the URM (along with EB and AR) is not, but we do not see DOLS ranking teachers differently here. Further, when there is substantial missing data on the one-year lag test score, we found larger differences in the rank correlations across estimators (with the URM better in some cases), but we also do not see those divergences here either. So the missing data patterns that are present in the administrative are not driving important differences in the estimated teacher effects.

For another illustration related to a policy context, Table 6 shows the average percent of teachers for which each of the other estimators would disagree with the URM on their classification of teachers in the top decile of the distribution of estimated teacher effects. So this could represent a scenario where the top 10 percent of teachers received a pay increase or bonus. The disagreement rates range from 0.3%–2.6%, with the smallest for EB estimation of the specification that "mimics" the URM, which is expected. In this case, during the 2002-2007 period only a handful of teacher effects were classified in the top 10 percent with the URM estimates, but classified as below the 90th percentile with the EB-mimic estimates. The analogous results in column [3] for DOLS show disagreement rates on the top decile are .7% for 5th grade and 1.2%–1.6% for 6th grade.

In all of our analysis, we see high agreement between our DOLS estimates using standard linear regression and the estimates from the more complicated EVAAS URM. As expected, the agreement between EB and the URM is at least as strong as that between DOLS and the URM—and in many cases slightly stronger—which suggests the estimation approach does

[15] Another scenario where DOLS and the URM would be expected to perform similarly is when there are large numbers of students per teacher, since then fixed effects and random effects estimators are expected to be similar. The similarity with AR though, further supports that there is not substantial deviation from random assignment in our data.

Table 6: Disagreement with the URM in classification of teachers above the 90th percentile

	1-yr lag Math, Student Char. [1]	1-yr & 2-yr lag Math, Student Char. [2]	1-yr & 2-yr lags in Math & Reading [3]	Composite score [4]
Panel A: Grade 5				
1-year estimates				
DOLS	2.6%	1.5%	0.7%	0.7%
AR	2.6%	1.4%	0.9%	0.8%
EB	2.6%	1.4%	0.2%	0.0%
2-year estimates				
DOLS	2.5%	1.4%	0.7%	0.7%
AR	2.5%	1.4%	1.0%	0.8%
EB	2.5%	1.4%	0.2%	0.0%
Panel B: Grade 6				
1-year estimates				
DOLS	2.1%	1.1%	1.2%	1.1%
AR	2.0%	2.0%	1.6%	1.5%
EB	2.0%	0.8%	0.3%	0.0%
2-year estimates				
DOLS	2.0%	1.2%	1.2%	1.0%
AR	1.9%	2.0%	1.6%	1.5%
EB	1.8%	0.8%	0.3%	0.0%

Notes: This table provides the average percent of teachers whose classification changes from the top 10 percent in the distribution of EVAAS URM estimated teacher effects to below the top 10 percent in the distribution of teacher effects based on the other estimator/specification combinations. The average is taken as the simple average of the percent misclassified in each year 2002-2007 for the 1-year estimates and 2003-2007 for the 2-year estimates.

matter somewhat in our data. And, although we cannot conclude which method is "better" based on the administrative data alone, we do find that our results based on administrative data are consistent with our simulation results.

6 Summary and Conclusions

We have shown how, in a simplified setting, the multi-step EVAAS URM estimation approach relates very closely to simple OLS estimation using the same lagged test scores. While this exact relationship is more difficult to see when we extend to settings with missing data or multiple years, we show how similar the estimates are, and under what conditions they are expected to diverge, using both simulations and administrative data.

Our simulation evidence with complete data shows that the URM exhibits similar performance patterns to those seen with empirical Bayes' estimation in Guarino et al. (2015). This is not surprising given that, as we show, the URM and EB estimators are very similar when there is no missing data. While the URM and EB perform similarly to DOLS under the ideal conditions of random assignment and random grouping, DOLS is most robust to nonrandom assignment, especially assignment based on the lagged score, which is certainly a plausible assignment mechanism. When we simulate data where 20 percent of students are missing the one-year lag test score, the URM performance fluctuates more than the other estimators, performing slightly better than DOLS in some sorting/assignment scenarios, but also much worse under negative assignment based on the lagged score. In sensitivity analyses, we found that the these performance differences are much smaller when the missing test score is instead the two-year lag test score.

Our results based on administrative data show similarity between the URM/EB and DOLS estimates, regardless of specification, which suggests that that there may not be substantial sorting in this district, and that the incomplete data patterns are not driving

important differences in results.

So although our simulations showed that OLS generally does as well—or better—than the more complicated EVAAS URM in recovering true teacher effects under many data scenarios, our analysis of administrative data suggests the extent of the differences may not be extremely problematic in practice. This is perhaps reassuring given that the EVAAS methods are already used in several states and districts for teacher evaluation purposes, in some cases for high-stakes decision making. However, taking into consideration the superiority of DOLS-type estimators in dealing with potential bias due to nonrandom assignment along with the proprietary nature and lack of transparency of the EVAAS methodology, our findings suggest that policymakers would be better served and better able to verify results using the conceptually and computationally simpler DOLS approach.

Acknowledgements

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

References

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System (EVAAS). *Educational Researcher*, 37(2), 65-75.

Backes, B., Cowan, J., Goldhaber, D., Koedel, C., Miller, C., & Xu, Z. (2018). The common core conundrum: To what extent should we worry that changes to assessments will affect test-based measures of teacher performance? *Economics of Education Review*, 62, 48-65.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.

Ballou, D., & Springer, M.G. (2015). Using student test scores to measure teacher performance: Some problems in the implementation of evaluation systems. *Educational Researcher*, 44(2), 77-86.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *The American Economic Review*, (9), 2593-2632.

Doherty, K.M., & Jacobs, S. (2015). State of the states 2015: Evaluating teaching, leading, and learning. Accessed August 7, 2016 at http://www.nctq.org/dmsView/StateofStates2015

Goldhaber, D., & Chaplin, D. D. (2015). Assessing the "Rothstein falsification test": Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, 8(1), 8-34.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics*

and Public Policy, 1(1), 28-39.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2015). An evaluation of empirical Bayes' estimation of value-added teacher performance measures. *Journal of Educational and Behavioral Statistics*, (40), 190-222.

Guarino, C., Reckase, M. & Wooldridge, J. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*, 10(1), 117-156.

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

Harris, D. N. (2009). Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives. *Education*, 4(4), 319-350.

Henry, G., & Rose, R. (2014). Are value-added models good enough for teacher evaluations? Assessing commonly used models with simulated and actual data. Investigaciones de Economía de la Educación 9, 383-405. Accessed August 8, 2016 at https://ideas.repec.org/h/aec/ieed09/09-20.html

Ishii, J., & Rivkin, S. G. (2009). Impediments to the estimation of teacher value-added. *Education Finance and Policy*, 4(4), 520-536.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kane, T. & Staiger, D. (2008). Estimating teacher impacts on student achievement: an experimental evaluation, Working Paper 14607, National Bureau of Economic Research.

Koedel, C. & Betts, J. (2009) Value-added to what? How a ceiling in the testing instrument influences value-added estimation, Working Paper 14778, National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education*, 6(1), 18-42.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis*, 25(3), 287-298.

Lockwood, J. R. & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics*, (1), 223-252.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics*, 32(2), 125-150.

Mariano, L.T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 35(3), 253-279.

McCaffrey, D. F., & Lockwood, J. R. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics*, 5(2A), 773-797.

McCaffrey, D. F., Lockwood, J. R., 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.

Rose, R., Henry, G., & Lauen, D. (2012). Comparing value-added models for estimating individual teacher effects on a statewide basis: Simulations and empirical analyses. Consortium for Educational Research and Evaluation North Carolina. Retrieved August 2, 2016, from http://cerenc.org/wp-content/uploads/2011/10/Full-VAM_report_FINAL_8-27-12.pdf

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

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38, 9-23.

Sanders, W. (2006). Comparisons among various educational assessment value-added models. Presented at The Power of two: National Conference on Value-Added, Columbus, OH, October 16. SAS® White Paper. Cary, NC: SAS Institute.

SAS Institute Inc. (2011). SAS EVAAS for K–12. Retrieved August 3, 2016 from http://scee.groupsite.com/file_cabinet/files/459928/download/SAS_EVAAS_for_K-12.pdf?m=1319821954

SAS Institute Inc. (2014). SAS EVAAS for K–12: Fact Sheet. Retrieved August 2, 2016, from http://www.sas.com/content/dam/SAS/en_us/doc/productbrief/sas-evaas-k12-104570.pdf

SAS Institute Inc. (2015a). SAS EVAAS for K–12: Solution Overview. Retrieved August 2, 2016, from http://www.sas.com/content/dam/SAS/en_us/doc/overviewbrochure/sas-evaas-k12-104570.pdf

SAS Institute Inc. (2015b). SAS EVAAS for K–12 Statistical Models. Retrieved August 2, 2016, from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/sas-evaas-k12-statistical-models-107411.pdf

Steinberg, M.P., & Donaldson, M.L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the Post-NCLB era. *Education Finance and Policy*, 11(3): 340-359.

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, 2nd ed. Cambridge, MA: MIT Press.

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. SAS® White Paper. Cary, NC: SAS Institute Inc.

Wright, S. P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS® EVAAS® statistical models. SAS® EVAAS® Technical Report. Cary, NC: SAS Institute Inc.

Wright, S. P. (2010). An investigation of two nonparametric regression models for value-added assessment in education. SAS® EVAAS® Technical Report. Cary, NC: SAS Institute Inc.

Wright, S. P (2015). Educational Value-Added analysis of covariance models with error in the covariates. Chapter 10 in R. W. Lissitz and H. Jiao (eds), Value Added Modeling and Growth Modeling With Particular Application to Teacher and School Effectiveness. Charlotte, NC: Information Age Publishing.