The Education Policy Center
AT MICHIGAN STATE UNIVERSITY

WORKING PAPER #18

Can Value-Added Measures of Teacher Education Performance Be Trusted?

Cassandra M. Guarino, Indiana University
Mark D. Reckase, Michigan State University
Jeffrey M. Wooldridge, Michigan State University

December 12, 2012

Note: This paper was originally posted March 2011, revised April 2011, May 2012 and July 17, 2013

The content of this paper does not necessarily reflect the views of The Education Policy Center or Michigan State University

epc@msu

Can Value-Added Measures of Teacher Performance Be Trusted?

Author Information

Cassandra M. Guarino
Associate Professor of Educational Leadership and Policy Studies
Indiana University
4220 W. W. Wright Education Building, 201 N. Rose Avenue
Bloomington, IN 47405-1006
Email: guarino@indiana.edu Phone: (812) 856-2927

Mark D. Reckase
University Distinguished Professor of Measurement and Quantitative Methods
Michigan State University
461 Erickson Hall, 620 Farm Lane
East Lansing, MI 48824
Email: reckase@msu.edu Phone: (517) 355-8537

Jeffrey M. Wooldridge
University Distinguished Professor of Economics
Michigan State University
110 Marshall-Adams Hall
East Lansing, MI 48824-1038
Email: wooldri1@msu.edu Phone: (517) 353-5972

Original version: March 2011
Revised: April 2011
Revised: May 2012
Current version: 12/12/12

Acknowledgment: This work was supported by grants R305D10002 and R305B090011 from the Institute of Education Sciences in the U.S. Department of Education. We thank Francis Smart, Brian Stacy, and Eun Hye Ham for programming and research assistance, and Steven Haider, Ken Frank, Barbara Schneider, Anna Bargagliotti, Steven Dieterle, Brian Stacy, Francis Smart, Cory Koedel, Doug Harris, Jacob Vigdor, Jonah Rockoff, Kelly Vosters, and Andrew Bibler for helpful comments.

Abstract

We investigate whether commonly used value-added estimation strategies produce accurate estimates of teacher effects under a variety of scenarios. We estimate teacher effects in simulated student achievement data sets that mimic plausible types of student grouping and teacher assignment scenarios. We find that no one method accurately captures true teacher effects in all scenarios, and the potential for misclassifying teachers as high- or low-performing can be substantial. However, a dynamic OLS estimator is more robust across scenarios than other estimators. Misspecifying dynamic relationships can exacerbate estimation problems.

Can Value-Added Measures of Teacher Performance Be Trusted?

Cassandra M. Guarino
(Corresponding Author)
Associate Professor of Educational Leadership and Policy Studies
Indiana University
4220 W. W. Wright Education Building, 201 N. Rose Avenue
Bloomington, IN 47405-1006 USA
Phone: (812) 856-2927
Email: guarino@indiana.edu

Mark D. Reckase
University Distinguished Professor of Measurement and Quantitative Methods
Michigan State University
461 Erickson Hall, 620 Farm Lane
East Lansing, MI 48824 USA
Phone: (517) 355-8537
Email: reckase@msu.edu

Jeffrey M. Wooldridge
University Distinguished Professor of Economics
Michigan State University
110 Marshall-Adams Hall
East Lansing, MI 48824-1038 USA
Phone: (517) 353-5972
Email: wooldri1@msu.edu

Acknowledgment: This work was supported by grants R305D10002 and R305B090011 from the Institute of Education Sciences in the U.S. Department of Education. We thank Francis Smart, Brian Stacy, Michelle Maxfield, and Eun Hye Ham for programming and research assistance, and Steven Haider, Ken Frank, Barbara Schneider, Anna Bargagliotti, Steven Dieterle, Brian Stacy, Francis Smart, Cory Koedel, Doug Harris, Jacob Vigdor, Jonah Rockoff, Kelly Vosters, and Andrew Bibler for helpful comments.

Abstract: We investigate whether commonly used value-added estimation strategies produce accurate estimates of teacher effects under a variety of scenarios. We estimate teacher effects in simulated student achievement data sets that mimic plausible types of student grouping and teacher assignment scenarios. We find that no one method accurately captures true teacher effects in all scenarios, and the potential for misclassifying teachers as high- or low-performing can be substantial. However, a dynamic OLS estimator is more robust across scenarios than other estimators. Misspecifying dynamic relationships can exacerbate estimation problems.

<1 Introduction>

Accurate indicators of educational effectiveness are needed to advance national policy goals of raising student achievement and closing socioeconomically based achievement gaps. If constructed and used appropriately, such indicators for both program evaluation and the evaluation of teacher and school performance could have a transformative effect on the nature and outcomes of teaching and learning. Measures of teacher quality based on value-added models of student achievement (VAMs) are gaining increasing acceptance among policymakers as a possible improvement over conventional indicators, such as classroom observations or measures of educational attainment or experience. They are already in use to varying degrees in school districts[1] and widely reported in the research literature.

Intuitively, VAMs are appealing; they track relative levels of achievement from one year to the next for individual students and parse growth into pieces believed to represent the separate contributions made by teachers and schools as well as individual-specific factors. Moreover, given that standardized testing is now ubiquitous in U.S. school systems, VAMs can be inexpensive to implement relative to other forms of teacher evaluation such as classroom observation, and their use has been encouraged by Race to the Top (U.S. Department of Education, 2009). As a teacher evaluation tool, VAM-based measures are sometimes viewed as less subjective than judgments based on observations by principals or portfolios of accomplishments. Given the increasing visibility of VAM-based estimates of teacher and school quality, and the possible inclusion of teacher performance incentives in the upcoming

[1] McGuinn (2012) reports that at least 43 states currently require annual teacher evaluations and 32 incorporate student performance measures. In addition, in some districts, the popular press has computed and published teacher value-added scores online. The Los Angeles Times has released Los Angeles Unified School District VAM-based ratings for individual public school teachers in the Los Angeles Unified School District in grades 3 through 5 since 2010 (see: http://www.latimes.com/news/local/teachers-investigation/ downloaded 11/21/12)), and in New York City, after a protracted court battle, the district has was required to make teacher evaluation measures available to the public. The Wall Street Journal has since released them for teachers in grades 4 through 8 (see: http://projects.wsj.com/nyc-teachers/ downloaded 11/21/12).

reauthorization of NCLB, it is imperative that such measures be well constructed and understood.

Disagreement exists, however, as to the best way to construct VAMs and to their optimal application. Numerous methods have been developed (e.g., Sanders and Horn, 1994; Ballou, Sanders, and Wright, 2004; Kane and Staiger, 2008; Raudenbush, 2009), and studies that compare estimates derived from different models have found substantial variability across methods (McCaffrey et al., 2004). Moreover, concerns remain that our understanding of these models is as yet limited and that incentives built around them may do more harm than good, with teachers' unions, in particular, reluctant to allow their constituents to be judged on the basis of measures that are potentially biased or imprecise.

A central issue involved in establishing the validity of measures and inferences based on VAMs is whether VAMs effectively isolate the "true" contribution of teachers and schools to achievement growth or instead confound these effects with the effects of other factors that may or may not be within the control of teachers and schools. Given that neither students nor teachers are randomly assigned to schools and that students are not randomly assigned to teachers within schools, disentangling the causal effects of schooling from other factors influencing achievement is far from straightforward. Studies that have attempted to corroborate results from VAMs have led to somewhat different conclusions, and questions about the usefulness of VAMs linger.[2]

In this paper, we investigate the ability of commonly used estimation strategies to produce accurate estimates of teacher effects. Our main research question is the following: How

[2] Kane and Staiger (2008) compare experimental VAM estimates for a subset of Los Angeles teachers with earlier non-experimental estimates for those same teachers and find that they are similar, suggesting that they are valid. Chetty et al. (2011) find that student achievement responds in expected ways to the entry and exit of teachers with differing value-added to a school. On the other hand, studies that examine the intertemporal stability of estimates find a fair amount of variation from year to year (e.g., Aaronson, Barrow, and Sanders, 2007). Rothstein (2010) devises falsification tests that challenge the validity of VAM-based measures of teacher performance in North Carolina, although Goldhaber and Chaplin (2012) and Kinsler (2012) have shown that such tests can reject even when estimates are bias free.

well do different estimators perform in estimating teacher effects under a variety of known conditions, including those in which particular underlying assumptions are violated?

We focus our study on estimators that are commonly used in research and policy applications involving teacher effects. We first outline the assumptions that must be met for each estimator to have desirable statistical properties in the context of a conventional theoretical framework. We then apply the estimators to the task of recovering teacher effects in simulated student achievement data generated under different types of student grouping and teacher assignment scenarios. We then compare the estimated teacher effects to the true teacher effects embedded in the data.

Our investigations yield several important findings. No one estimator performs well under all plausible circumstances, but some are more robust than others. Surprisingly, certain estimation approaches known to be inconsistent in the structural modeling framework fare better than expected. We find that some of the most popular methods are neither the most robust nor ideal choice. Our simulations highlight the pitfalls of misspecifying the dynamic relationship between current and prior achievement. In addition, we find that substantial proportions of teachers can be misclassified as “below average” or “above average” as well as in the bottom and top quintiles of the teacher quality distribution, even in the best-case scenarios.

In interpreting our findings, it is important to emphasize that they result from data generation processes that represent controlled conditions and incorporate many of the assumptions underlying the relatively simple conceptual model upon which value-added estimation strategies are based. These simplifications are the strength of our research design. Undoubtedly real-life educational conditions are more complex, potentially obscuring the sources of difference among them. Applying the various estimators to controlled, plausible scenarios is the best way to discover fundamental flaws and differences among them when they should be expected to perform at their best.

The paper is organized as follows. In Section 2, we outline a structural framework for value-added models. Section 3 discusses each estimator in turn and its underlying assumptions. We describe different mechanisms for grouping students and assigning teachers to classrooms in Section 4. Section 5 describes the simulation procedures and estimation strategies we employ. The simulation results in Section 6 investigate the ability of the various value-added estimators of teacher performance to uncover true effects under our different data generating scenarios. By systematically comparing VAM-based estimates resulting from different estimators to the true effects embedded in the various data generating processes, we are able to identify estimation strategies most likely to recover true effects under particular conditions.

<2 Conceptual Framework Underlying Value-Added Modeling>

The derivation of particular VAMs typically rests on the specification of a structural "education production function," in which achievement at any grade is modeled as a function of child, family, and schooling inputs. In its most general formulation, learning is a process that is considered to be both dynamic and cumulative – that is, past experiences and past learning contribute to present learning. Thus the model—often referred to as the generalized *cumulative effects model* (CEM)—includes all relevant past child, family, and school inputs (Hanushek, 1979, 1986; Boardman & Murnane, 1979; Todd & Wolpin, 2003; Harris, Sass, & Semykina, 2011). This model can be expressed as:

$$A_{it} = f_t\,(E_{it},\ldots, E_{i0},X_{it},\ldots,X_{i0},c_i, u_{it}) \qquad (1)$$

where A_{it} is the achievement of child i in grade t, E_{it} represents school-related inputs, X_{it} represents a set of relevant time-varying child and family inputs, c_i captures the unobserved time-invariant student effect (representing, for example, motivation, some notion of sustained ability, or some persistent behavioral or physical issue that affects achievement), and the u_{it} represent the idiosyncratic shocks that may occur in any given period. In this very general formulation, the functional form is unspecified and can vary over time.

Moving to an empirical model poses large challenges due to the lack of information regarding most past and even many current inputs to the process and the manner in which they are related to one another—that is, functional form, interactions, lagged responses, feedback, and so on. Inferring the causal effects of teachers and schools is therefore difficult. If children were randomly assigned to teachers and schools, many omitted variable issues would be considerably mitigated. However, random assignment does not typically characterize school systems, and, indeed, is not necessarily desirable. Random assignment of children to schools deprives parents of the ability to find schools that they believe to be best suited for their children through both residential sorting and school choice. Random assignment to teachers within schools deprives principals of one of their most important functions: to maximize overall achievement by matching the individualized skills of teachers to those students most likely to benefit from them. Thus random assignment—while helpful from an evaluation standpoint—could result in suboptimal learning conditions if particular teacher and school characteristics interact in a beneficial way with student characteristics in the learning process.

Clearly, however, knowledge of the effectiveness of particular schools, teachers, or programs in promoting learning is essential if we are to foster successful instructional approaches and curtail the use of ineffective ones. Causal measures of performance at the school, teacher, and program level are needed to identify instructional strategies that contribute to high performance. In the context of nonrandom assignment and omitted variables, statistical methods are the only tools available with which to infer effects, but they rely on strong assumptions. In the next sections, we describe the assumptions used to derive models that are empirically feasible to estimate.

<2.1 The General Linear Formulation>

A distributed lag version of the cumulative effects model that assumes linearity is the typical and potentially tractable starting point for structural modeling. Equation (1) becomes

$$A_{it} = \alpha_t + E_{it}\beta_0 + E_{i,t\text{-}1}\beta_1 + \ldots + E_{i0}\,\beta_t + X_{it}\gamma_0 + X_{i,t\text{-}1}\gamma_1 + \ldots + X_{i0}\gamma_t + \eta_t c_i + u_{it} \qquad (2)$$

where we take E_{it} to be a row vector of observed education inputs at time t – including teacher or school characteristics, or, say, teacher indicators – and X_{it} to be a vector of observed time-varying individual and family characteristics such as health status, household income, and so on. The term α_t allows for a separate intercept in each time period, which would be appropriate if, for example, the score scales are set to be different for different grade levels by the testing program. The period t = 0 corresponds to the initial year in school (which is generally kindergarten or could be pre-kindergarten in states where this is a common public school option). This formulation has several assumptions embedded in it: linearity, a functional form that is constant over time (except for the intercept and possibly the coefficient on c_i), and an additive, idiosyncratic shock, u_{it}, that accounts for all unobserved time-varying current and past factors.

Note that the formulation in (2) does not explicitly recognize the possible presence of interactions among teachers, between teachers and students, or among students and is therefore a limited conceptualization of the educational learning process. It is possible to build in these complexities, although it is rarely done in practice, except for the occasional inclusion of peer characteristics.

Exogeneity assumptions on the inputs are needed to estimate the parameters in the linear CEM. A common starting point – termed *sequential exogeneity* by Chamberlain (1992) – assumes that the expected value of the time-varying unobservables, u_{it}, conditional on all relevant time-varying current and past inputs and the unobserved child effect, is zero:

$$E(u_{it} | E_{it}, E_{i,t-1}, \ldots, E_{i0}, X_{it}, X_{i,t-1}, \ldots, X_{i0}, c_i) = 0 \,. \tag{3}$$

In practical terms, (3) requires that the time-varying unobservables that affect achievement are uncorrelated with observed school and family inputs—both current and past. However, it is plausible that this assumption could be violated. For example, u_{it} can contain factors such as unobserved parental effort that respond to the assignment of school inputs such as when a parent provides more help for a student who is assigned to a poor teacher or a large class.

Importantly, (3) is an assumption about correlation between inputs and the time-varying unobservables, u_{it}, and it is silent on the relationship between student heterogeneity, c_i, and the observed inputs. Many estimation approaches either ignore the presence of c_i or assume it is uncorrelated with observed inputs – in other words, they assume what we would call *heterogeneity exogeneity*. If c_i is correlated with observed inputs, standard pooled regression and generalized least squares approaches are generally inconsistent regardless of what we assume about the relationship between u_{it} and the inputs. Several approaches can be used to deal with unobserved heterogeneity in equation (2) – most commonly, fixed effects and first-differencing methods – each with a set of assumptions and drawbacks. If we are not wedded to a structural model as in equation (2), past test scores can be included in regression equations as proxies for the unobserved heterogeneity. In fact, this is an important motivation for the dynamic regression method described in Section 3.

Beyond the issue of unobserved heterogeneity, there are other obstacles to estimating equation (2). The linear CEM in this form is rarely estimated due to data limitations. If, for example, we have testing data on third through sixth grade for each child and want to allow for the possibility that all previous teachers affect current outcomes (in this case, the E_{it} vector may be composed of teacher dummy variables), we need to have data linking students to their teachers in second and first grades, as well as kindergarten. In addition to the onerous data requirements, high correlations among inputs across time periods can limit the ability of any of these estimators to isolate specific contemporaneous or past effects and make estimation of the linear CEM unattractive.

<2.2 Geometric Distributed Lag Restrictions on the Linear CEM>

To solve the data limitations issue and conserve on parameters in the general linear CEM, researchers typically impose restrictions on the distributed lag coefficients. A simple and commonly applied restriction is a geometric distributed lag (GDL), which imposes geometric decay on the parameters in (2) for some $0 \leq \lambda \leq 1$:

$$\beta_s = \lambda^s \beta_0, \; \gamma_s = \lambda^s \gamma_0, \quad s = 1,\ldots,T \tag{4}$$

This means that the effects of all past time-varying inputs (schooling-related as well as child- and family-related) decay at the same rate over time and their influence on current achievement decreases in the specified manner as their distance from the present increases. With these restrictions, after subtracting $\lambda A_{i,t\text{-}1}$ from both sides of (2) and performing substitutions and simple algebra, we obtain a much simpler estimating equation:

$$A_{it} = \tau_t + \lambda A_{i,t\text{-}1} + E_{it}\beta_0 + X_{it}\gamma_0 + \pi_t c_i + e_{it} \tag{5}$$

where

$$e_{it} = u_{it} - \lambda u_{i,t\text{-}1} \tag{6}$$

Equation (5) has several useful features. First, the right hand side includes a single lag of achievement and only contemporaneous inputs. This is a much more parsimonious estimating equation than the general model (2) because past inputs do not appear. Consequently, data requirements are less onerous than those for the linear CEM, and parameter estimation of (5) is less likely to suffer from the multicollinearity that can occur among contemporaneous variables and their lags.

It is important to see that the decay structure in the GDL equation means that any distributed lag effects are determined entirely by λ and β_0. In other words, once we know the effect of contemporaneous inputs (β_0) and the persistence parameter (λ), the effects of lagged inputs are determined. Undoubtedly this is a highly restrictive assumption, but (5) is fairly common in the education literature. It is important to note, however, that the rate at which knowledge decays may differ for different students or for different subpopulations of students (Entwistle and Alexander, 1992; Downey, Hippel, and Broh, 2004). Although allowing persistence parameters to vary by individuals or groups is possible in (5), this is rarely, if ever, done in the literature on teacher effects.

In deriving estimators based on equation (5), we must consider the exogeneity of inputs in this equation, including possible correlation with c_i as well as correlation with the time-varying unobservables e_{it}. As shown in equation (6), e_{it} depends on the current and lagged error from equation (2). If we maintain the sequential exogeneity assumption (3) in the structural CEM, u_{it} is uncorrelated with E_{it}. In that case, simple algebra gives

$$Cov(E_{it}, e_{it}) = -\lambda\, Cov(E_{it}, u_{i,t-1}). \qquad (7)$$

Equation (7) shows explicitly that in order to treat E_{it} and X_{it} as exogenous in (5) – that is, uncorrelated with the time-varying unobservables e_{it} – we need to impose an assumption stronger than the sequential exogeneity in the structural equation (2) (unless $\lambda = 0$, which seems unlikely). In this case, the weakest exogeneity condition is that E_{it} is uncorrelated with $u_{it} - \lambda\, u_{i,t\text{-}1}$, which could be true even if we do not assume E_{it} is uncorrelated separately with $u_{i,t\text{-}1}$ and u_{it}. However, for certain estimation strategies discussed below, the imposition of a stronger exogeneity assumption on the CEM, namely *strict exogeneity*, is needed and is clearly sufficient for Cov(E_{it}, e_{it}) = 0. A straightforward way to state the strict exogeneity assumption is

$$E(u_{it}|\, E_{iT}, E_{i,T\text{-}1},\ldots, E_{i0,}\, X_{iT}, X_{i,T\text{-}1},\ldots, X_{i0}, c_i) = 0. \qquad (8)$$

The difference between assumptions (8) and (3) is that (8) includes the entire set of observed inputs, including *future* inputs [this is why the t in (3) is replaced with T in (8)]. Assumption (8) implies that the error term e_{it} in (5) is uncorrelated with inputs at time t and all other time periods.

In addition to possible correlation between the covariates and e_{it}, however, we must recognize that it is virtually impossible for c_i to be uncorrelated with $A_{i,\, t\text{-}1.}$ Moreover, we often expect c_i to be correlated with the inputs. A simplistic approach to dealing with issues stemming from the presence of the lagged dependent variable is to assume that it does not matter – that is, assume that $\lambda = 0$, which implies complete decay (i.e., no persistence). In this case, (5) reduces to what is often referred to as a “level-score” equation. As a special case of the CEM, the level-score approach is unattractive because $\lambda = 0$ is unrealistic. But level-score regressions have been used with experimental data – that is, when the inputs are randomly assigned – because then the structural CEM approach is not necessary for identifying teacher effects (see, for example, Dee, 2004). For estimating teacher value added, random assignment means that one can compare mean achievement scores across teachers, and that is exactly what level-score regressions do in that setting.

Another simple but widely used formulation sets $\lambda = 1$ (no decay), which leads to subtracting $A_{i,\ t-1}$ from both sides of (5), thereby achieving a so-called "gain-score" formulation:

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + X_{it}\gamma_0 + \pi_t c_i + e_{it}. \qquad (9)$$

We now turn to describing different estimators used to estimate VAMs along with their statistical properties.

<3 Commonly Used Estimators and their Underlying Assumptions>

This section describes six commonly used estimation methods and the assumptions underlying their use. One important point is that the justification for many of these approaches appeals to large-sample properties because several of the estimators have no tractable finite-sample properties (such as unbiasedness) under any reasonable assumptions. Appealing to asymptotic analysis is hardly ideal, especially for applications where the inputs are teacher assignments. In this scenario, the large-sample approximation improves as the number of students per teacher increases. But in many data sets, the number of students per teacher is somewhat small – fewer than 100 – making large-sample discussions tenuous. Nevertheless, asymptotic theory is the unifying theme behind the estimators that are applied in VAM contexts and provides a framework within which to identify underlying assumptions.

<3.1 Dynamic Ordinary Least Squares>

If we ignore $\pi_t c_i$ in equation (5), then we might take a seemingly naïve approach and simply estimate a dynamic regression. In other words, we estimate λ, β_0, and γ_0 using a pooled OLS regression. We refer to this estimator as "dynamic ordinary least squares" (DOLS), where "dynamic" indicates that we have included a lagged test score.

Consistency of the DOLS estimator for β_0, γ_0, and λ – which are the parameters in the structural model – hinges on strict exogeneity of the inputs (with respect to $\{u_{it}\}$) and no serial correlation in $\{e_{it}\}$. Since $e_{it} = u_{it} - \lambda\, u_{i,t\text{-}1}$, to claim that the $\{e_{it}\}$ are serially uncorrelated, we must place restrictions on the original errors $\{u_{it}\}$. First, we must assume they follow an $AR(1)$ process, namely $u_{it} = \rho u_{i,t\text{-}1} + r_{it}$ where $\{ r_{it} \}$ is serially uncorrelated, and, second, we must assume that $\rho = \lambda$, which is often called the "common factor" (CF) restriction. The CF restriction amounts to assuming that past shocks to learning decay at the same rate as learning from family- and school-related sources. This is by no means an intuitive assumption. In any case, under the CF restriction the transformed errors $e_{it} = u_{it} - \lambda\, u_{i,t\text{-}1}$ in (5) are the same as the serially uncorrelated r_{it}.

In addition, the presence of $\pi_t c_i$ generally causes inconsistency because c_i is correlated with $A_{i,t\text{-}1}$. Further, c_i might be correlated with the inputs E_{it}, which happens if students are assigned educational inputs based on time-constant unobservables. Controlling for background variables can mitigate the problem, but good proxies for c_i may be hard to come by; those easily available (for example, gender or race) are not suitable proxies for factors such as motivation or cognitive ability.

Even if, technically speaking, DOLS is inconsistent for the structural parameters, it could nevertheless provide relatively accurate estimates of β_0 under certain circumstances. For example, if the $\pi_t c_i$ are sufficiently "small," ignoring this component of the composite error term $v_{it} = \pi_t c_i + u_{it}$ might not be costly. Even with substantial heterogeneity, the lagged test score may serve as a good proxy for c_i, resulting in good estimators of β_0 even though λ may be poorly estimated. An attractive feature of DOLS is that controlling for $A_{i,t\text{-}1}$ explicitly allows for the kinds of dynamic assignment of students to inputs based on prior test scores.

<3.2 The Average Residual Approach>

The DOLS regression requires including numerous teacher assignment dummies along with the lagged test score and possibly other covariates. A simpler, two-step alternative is fairly common in the VAM literature. First, one regresses A_{it} on $A_{i,t-1}$ and any other covariates, X_{it}, obtaining the residuals, say $\hat{v}_{it}$. Then, these residuals are averaged within teacher to obtain the teacher VAMs. (This is algebraically the same as obtaining the coefficients from regressing the $\hat{v}_{it}$ on the teacher assignment dummies.) This two-step approach, which we call the "average residual" (AR) approach, is implemented by several authors,[3] including Chetty et al. (2011), McCaffrey et al. (2010), West and Chingos (2009), and Kane and Staiger (2008). It is also used for evaluation purposes in districts in various states (see the Wisconsin VARC estimator, Value-Added Research Center 2010).

The popularity of AR may hinge on the computational simplicity of obtaining average residuals by teacher, avoiding the need to run regressions with large sets of teacher dummies as is done with DOLS. Like DOLS, the AR approach can be applied if we use the gain score, ΔA_{it}, as the dependent variable rather than A_{it}.

An important drawback to the AR approach, and one that seems to be largely ignored, is that it does not partial out the teacher assignment from the lagged test score and other controls. Thus, the AR approach is generally biased and inconsistent for estimating the teacher effects if E_{it} is correlated with $A_{i,t-1}$. By contrast, because DOLS includes $A_{i,t-1}$ and E_{it} in the same regression, it properly accounts for any correlation that may exist between them – such as when teacher assignment depends on past achievement.

<3.3 Pooled OLS on the Gain Score>

[3] These authors often apply a shrinkage factor to the two-step estimates after the fact, as described below in the section on empirical Bayes.

Estimation based on equation (9), where the gain score, ΔA_{it}, is used as the dependent variable and the explanatory variables are the contemporaneous inputs, is advantageous if the assumption $\lambda = 1$ holds. If we can ignore the presence of c_i or successfully introduce proxies for it, pooled OLS (POLS) is a natural estimation method and is used in various applications (e.g., Ballou, Sanders, and Wright, 2004).

A more subtle point is that when we view (9) as an estimating equation derived from the structural model (2), consistency of POLS relies on the same kind of strict exogeneity assumption we discussed in connection with condition (8): assignment of inputs at time t, E_{it}, cannot be correlated with the time-varying factors affecting achievement at time $t - 1$, $u_{i,t-1}$. If the inputs are strictly exogenous in the CEM then E_{it} is uncorrelated with e_{it}, and POLS is consistent provided the inputs are uncorrelated also with the unobserved heterogeneity. In applying POLS, the presence of heterogeneity, and possibly heteroskedasticity and serial correlation in e_{it}, means that inference should be made robust to arbitrary heteroskedasticity and serial correlation in the composite error $\pi_t c_i + e_{it}$.

When there are only teacher dummies E_{it} in equation (9) the POLS estimator is the same as computing the average gain score for each teacher.

<3.4 The Instrumental Variables/Arellano and Bond Approach>

Rather than ignore c_i, panel data estimators can be used to account for it in various ways. For example, a combination of first differencing and instrumental variables can be used to account for unobserved heterogeneity, again assuming that π_t is a constant. We can eliminate c_i by first differencing (5) to obtain:

$$\Delta A_{it} = \chi_t + \lambda\Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta X_{it}\gamma_0 + \Delta e_{it}. \tag{10}$$

Generally, this differenced equation cannot be consistently estimated by OLS because $\Delta A_{i,t-1}$ is correlated with Δe_{it}. Nevertheless, under strict exogeneity of inputs $\{E_{it}\}$ and $\{X_{it}\}$, Δe_{it} is uncorrelated with inputs in any time period, and so it is possible to use lagged values of E_{it} and X_{it} as instrumental variables for $\Delta A_{i,t-1}$. (ΔE_{it} and ΔX_{it} act as their own instruments under strict exogeneity.) If we use more than one lag – as is often required to make the instruments sufficiently correlated with the changes – this IV approach increases the data requirements because we lose an additional year of data for each lag we include among the instruments. For example, if we use the lagged changes, $\Delta E_{i,t-1}$ and $\Delta X_{i,t-1}$, as IVs, we lose one year of data because these depend on $E_{i,t-2}$ or $X_{i,t-2}$, respectively. Thus, this estimator is rarely applied in practice. Instead, the estimator proposed by Arellano and Bond (1991) (AB), which chooses instruments for the lagged gain score from available achievement lags, is more often used (e.g., Koedel and Betts, 2011).

The AB approach is limited by its requirement that there be no serial correlation in the $\{e_{it}\}$, thus imposing the common factor restriction described above. Formally stated, an assumption that implies no serial correlation in the errors and strictly exogenous inputs is:

$$E(e_{it}| A_{i,t-1}, A_{i,t-2},\ldots, A_{i0}, E_{iT}, E_{i,T-1},\ldots, E_{i0}, X_{iT}, X_{i,T-1},\ldots, X_{i0}, c_i) = 0, \qquad (11)$$

which maintains that e_{it} is unpredictable given past achievement and the entire history of inputs. The usefulness of assumption (11) is that it implies that $\{A_{i,t-2}, \ldots,A_{i0}\}$ are uncorrelated with e_{it}, and so these are instrumental variable candidates for $\Delta A_{i,t-1}$ in (11). Typically, $\{A_{i,t-2}, \ldots,A_{i0}\}$ is sufficiently correlated with $\Delta A_{i,t-1}$, as long as λ is not "close" to one. With achievement scores for four grades, and teacher assignments for the last three, equation (10) can be estimated using two years of gain scores.

Generally, care is needed when instrumenting for $\Delta A_{i,t-1}$ when λ is close to one. In fact, if there were no inputs and $\lambda = 1$, the AB approach would not identify λ. Simulation evidence in Blundell and Bond (1998) and elsewhere verifies that the AB moment conditions produce noisy estimators of λ when λ is near one. We should remember, though, that our main purpose here is in estimating school input effects (in our case, teacher effects), β_0, rather than λ. For that purpose, the weak instrument problem when λ is near unity may not cause the AB approach to suffer too severely.

If we wish to allow for the possibility of dynamic assignment and *not* assume strict exogeneity of the inputs in (2), then ΔE_{it} requires instruments as well, and this is a tall order. In (10), Δe_{it} depends on $\{u_{it}, u_{i,t-1}\}$ and so, if we hope to relax strict exogeneity of the inputs in (2), we must choose our IVs from $\{A_{i,t-2}, \ldots, A_{i0}, E_{i,t-2}, \ldots, E_{i0}, X_{i,t-2}, \ldots, X_{i0}\}$. This approach imposes substantial data requirements.

It should be noted that AB, although it explicitly recognizes the presence of c_i, may not perform better than any of the OLS estimators. As we have noted, estimating λ when it is unity can be costly when using the first-differenced equation (10). Moreover, in cases where assignment is based on lagged test scores, DOLS controls directly for the lagged test score, whereas AB controls for the lagged change in the test score (and then uses further lags for instruments). Thus, the AB method may not produce good VAM estimates in reasonable assignment scenarios.

<3.5 Random Effects on the Gain Score>

If we assume that π_t in equation (9) is constant and that $\{e_{it}\}$ is homoskedastic and serially uncorrelated, the composite error term, $c_i + e_{it}$ has a random effects (RE) variance-covariance structure, and RE estimation of (9) is attractive because it is then the feasible GLS estimator. Thus, under an ideal set of assumptions in (9), RE is asymptotically more efficient than POLS. For the purposes of estimation, the POLS estimator ignores the serial correlation in the composite error, $c_i + e_{it}$. As we discussed earlier, we can make POLS inference fully robust to any kind of heteroskedasticity and serial correlation in $c_i + e_{it}$ (even if we allow for the time-varying π_t), but we may be giving up some asymptotic efficiency. Like POLS, consistency of RE generally relies on λ = 1; otherwise a lagged value of the test score has been effectively omitted from the right hand side of (9). Even assuming that λ = 1, RE, like POLS, assumes the student heterogeneity is uncorrelated with inputs. Nevertheless, RE uses the heterogeneity exogeneity assumption, along with strict exogeneity, more efficiently than POLS.[4]

<3.6 Fixed Effects on the Gain Score>

If, instead of ignoring or proxying for c_i, we allow for unrestricted correlation between c_i and the inputs E_{it} (and X_{it}), we can eliminate c_i in the gain score equation via the use of student-level fixed effects (FE) estimation (at least when π_t is constant). Of course we require at least two grades per student to apply this method. For consistency, the FE estimator requires a form of strict exogeneity of E_{it} and X_{it} because FE employs a time-demeaning transformation that

[4] When POLS and RE are both consistent, RE can still improve upon POLS in terms of efficiency even if $\{e_{it}\}$ is serially correlated or contains heteroskedasticity. Efficiency gains using RE in such settings are not guaranteed, but it is often more efficient that POLS because it accounts for serial correlation to some extent, even if not perfectly. This is the motivation behind the generalized estimating equations (GEE) literature (see, for example, Zeger, Liang, and Albert 1988 or Wooldridge 2010, Chapter 12). Also, π_t not being constant does not cause inconsistency of RE (or POLS), although RE would not be the efficient GLS estimator with time-varying π_t. (Fully robust inference for RE is available; see Wooldridge (2010, Chapter 10)). One could instead use an unrestricted GLS analysis that would allow any kind of variance-covariance structure for $\pi_t c_i$+ e_{it}. We do not explore that possibility here, however, as it is rare in applications.

requires that the e_{it} are uncorrelated with the time-demeaned inputs.[5] As with the other methods, the strict exogeneity assumption stated in (8) is sufficient. When inputs related to classroom assignments are thought to be based largely on time-constant factors, FE is attractive, whereas POLS and RE will suffer from systematic bias. If inputs are uncorrelated with the shocks and heterogeneity, however, FE is typically less efficient than RE, and can be less efficient than POLS, too. If in equation (5) $\lambda \neq 1$, equation (9) effectively omits the lagged dependent variable, and strict exogeneity fails if teacher assignment depends on the past test score or heterogeneity. Generally, consistency of the FE estimator is ensured only if λ is equal to 1, although it can be consistent if the inputs are randomly assigned.

Although FE is rarely used in practice to estimate teacher performance measures, we include it here for didactic purposes and note that it is sometimes used in value-added models designed to assess program effects at the teacher or school level.

<3.7 Empirical Bayes and Related Estimators>

A popular estimation approach to teacher VAMS is the so-called "empirical Bayes" (EB) method, application of which results in so-called "shrinkage estimators." The EB estimators are essentially the same as the VAMs obtained from the mixed model that is at the foundation of the Tennessee Value Added Assessment System (TVAAS) estimator originally developed by Sanders (for example, Ballou, Sanders, and Wright 2004) as well as related applications of hierarchical linear modeling. Briefly, the teacher effects are modeled as random outcomes and then the best linear unbiased predictors are obtained as functions of the unknown variance parameters; estimates of the variance parameters are inserted to form the final shrinkage estimates. As is well known (for example, Morris, 1983), the EB VAM estimates when only

[5] See Wooldridge (2010, Chapter 10) for a general discussion. For the same reason, a lagged dependent variable cannot be included on the right-hand side.

teacher effects are included are simply the pooled OLS estimates shrunk toward the overall mean teacher effect, using a shrinkage factor that varies by class size. If class size is constant across teachers and time periods, as in the simulations we conduct, the EB estimators of the teacher effects are the same up to a common scale factor as POLS when we impose $\lambda = 1$. When we allow $\lambda \neq 1$ in estimation, the EB estimates are similar to shrinking the AR estimates to the overall mean – a common practice in applications. The difference is that with EB λ would be estimated using a GLS procedure that allows student random effects (because of the panel structure) and treats the teacher effects as random variables. Like the AR approach, the EB approach is generally inconsistent when teacher assignment depends on the lagged test score or unobserved heterogeneity. Plus, when applied to panel data, the EB approach loses its theoretical appeal in the presence of lagged test scores because the explanatory variables can no longer be strictly exogenous. Given its equivalence to POLS when $\lambda = 1$ and near equivalence to AR when $\lambda \neq 1$ in our simulations, we do not report separate EB estimates in this paper.[6]

<3.8 Summary of Estimation Approaches>

In summary, estimation of the parameters of the structural cumulative effects model, even after we impose the geometric distributed lag restriction to arrive at equation (5), requires numerous additional assumptions. Although it is clear that every estimator has an Achilles heel (or more than one area of potential weakness), it is unclear whether violations of some assumptions cause more severe problems in estimating teacher effects than violations of others under plausible conditions. Some of these assumptions are required for consistent estimation of parameters such as λ, when, in our case, the parameters of interests are the teacher effects. It is thus an empirical question as to which estimator will perform best for our purpose across various

[6] In Guarino, Maxfield, Reckase, Thompson, and Wooldridge (2012), we discuss the empirical Bayes estimator in depth.

data generating mechanisms. The interesting question is which of the methods we have discussed does a better job recovering teacher effects under different conditions, and that is what this study aims to answer.

<4 Situating Theory in Context>

Until now, we have discussed assumptions underlying value-added models and estimation strategies in relatively abstract terms. We now describe the types of educational scenarios that form the basis of our simulation design and how they might be expected to violate exogeneity assumptions.

We consider the process of matching students to teachers to be composed of two separate decisions—the grouping of students into classrooms and the assignment of classrooms to teachers. Grouping students in classrooms on the basis of their perceived ability, often called “tracking,” is not uncommon and can take a number of forms. Students will likely be grouped together on the basis of either (1) their prior test score, $A_{i,t\text{-}1}$,[7] (2) their baseline level of achievement or ability upon entering school, A_{i0}, or (3) their potential for learning gains, c_i. For the purpose of the simulations we design, we consider these three cases. Of course, in reality, it may be possible for other factors to play a role in grouping, such as the recommendations of the prior teacher, the requests of parents, and behavioral considerations, although such grouping considerations will be less systematically applied.

We will refer to ability grouping based on prior test scores as “dynamic tracking,” following terminology used by Rothstein (2010). The second and third types of grouping, both forms of “static tracking,” are likely less common but might occur when, for example, schools either formally or informally assess the level of learning or the growth potential of children upon

[7] Here for simplicity we refer to just one prior test score. However, principals might average over a series of prior test scores.

entering school, group the children accordingly, then keep more or less the same groups of children together for several grades.

Ability grouping in one form or another is likely to occur on a reasonable scale within educational systems. In the empirical literature, the phenomenon of ability grouping has been investigated primarily through techniques such as those developed by Aaronson, Barrow, and Sander (2007), which compare the average within classroom standard deviation in prior test scores with that produced by artificially regrouping students into classrooms—either randomly or perfectly sorted. Most such studies find that actual average standard deviations are closer to the random scenario than the perfectly sorting scenario. Dieterle et al. (2012), however, use a more fine-grained approach to the analysis of achievement data and find that substantial numbers of schools engage in dynamic tracking, a fact that can easily be obscured by the aggregated statistics.

Tracking does not, in and of itself, induce correlation between unobserved factors affecting student performance and teacher effects but serves as a precondition for the possibility of correlation. We distinguish the practice of tracking— grouping of students together on the basis of some performance or ability criterion—from the practice of assigning these groups of students to teachers in nonrandom ways. In this study, we use the term “grouping” for the practice of placing students in classrooms and the term “assignment” for the action of assigning classrooms to teachers.

In our study, the assignment of classrooms to teachers takes three primary forms: random assignment, assignment in which there is a positive correlation between teacher effects and student performance (that is, when better students are assigned to better teachers), and assignment in which there is a negative correlation between teacher effects and student performance (that is, when worse students are assigned to better teachers). We summarize different combinations of grouping and assignment mechanisms that might be encountered in educational settings in Table 1, along with acronyms that we use in the remainder of the paper.

It is important to recognize that a mixture of these grouping and assignment methods can be used in any given district or even within a given school. However, for the sake of clarity in understanding and evaluating the performance of various estimators, we keep the scenarios distinct when we conduct our simulations and assume that all schools simultaneously use the same process.

Generally, the random assignment of groups of students (regardless of how the groups may be formed) to available teachers is not a violation of either strict exogeneity or heterogeneity exogeneity and thus may not cause problems for standard estimation methods. The students may be grouped using dynamic or static assignment provided the teachers are randomly assigned to the groups. Of course, grouping may have other consequences, such as inducing correlation within classrooms in the unobserved factors affecting performance. But this is different from failure of exogeneity.

The systematic assignment of high-performing students to either high- or low-performing teachers, on the other hand, can violate exogeneity assumptions. Dynamic grouping coupled with positive or negative assignment virtually always causes failure of strict exogeneity, because if the teacher assignment is correlated with past scores, then teacher assignment must be correlated with the innovations (errors) that affect past scores. In addition, if student heterogeneity c_i exists then dynamic grouping with nonrandom assignment violates heterogeneity exogeneity, too: part of past performance depends on c_i. It should be noted that Dieterle et al. (2012) find empirical evidence to suggest that ability grouping with *positive* assignment may occur to a nontrivial degree in school systems.

The two cases of *static* grouping differ in important ways. For example, suppose students are grouped on a baseline score upon entry to school and then assigned to teachers nonrandomly in all subsequent grades. While this is a case of nonrandom assignment, for some estimation approaches there is no violation of relevant exogeneity assumptions. As an illustration, in the gain score equation (9), the baseline score does not appear. Therefore, even if teacher assignment is determined by the base score, if it is independent of the student heterogeneity c_i and the errors e_{it}, then pooled OLS estimation consistently estimates β_0 (and the other parameters). Of course, this assumes that $\lambda = 1$ has been correctly imposed. If $\lambda < 1$, then the gain-score equation effectively omits the lagged test score, and this lagged score will be correlated with the base score, causing bias in any of the usual estimators applied to (9).

Static assignment based on c_i causes problems for estimating equations such as (9) unless $\pi_t c_i$ is removed from the equation. When π_t is constant, the fixed effects and first-differencing transformations do exactly that. Therefore, assigning students to teachers based on the student heterogeneity does not cause problems for these types of estimators applied to (9). But other estimators, particularly POLS and RE, will suffer from omitted variable bias because E_{it} is correlated with c_i. Static assignment based on student growth also causes problems for DOLS because DOLS ignores c_i in estimating (5).

Until now, we have focused on the assignment of students to teachers within schools. Another key consideration, however, is the sorting of students and teachers across schools. If higher achieving students are grouped within certain schools and lower achieving students in others, then the teachers in the high-achieving schools, regardless of their true teaching ability, will have higher probabilities of high-achieving classrooms. Similarly, if higher ability teachers are grouped within certain schools and lower ability teachers in others, then students in the schools with better teachers will realize higher gains. If both high ability teachers and high performing students are then grouped together within schools, the nonrandom sorting issue is exacerbated.

In designing our simulation scenarios, we therefore consider three distinct "school sorting" cases. In our baseline case, both students and teachers are randomly placed in schools. Thus there is no systematic difference in average test scores or average true teacher effects across schools. In sensitivity analyses, we explore two nonrandom cases: (1) students are sorted into schools according to their baseline levels of learning but teachers are still randomly placed in schools (thus there is a significant difference in average test scores across schools but not in average teacher effects) and (2) students are randomly placed in schools but teachers are sorted into schools based on their true effects (thus, there are systematic differences in average teacher effects across schools but not in average test scores).

In our investigation of the performance of various estimators under different sorting, grouping, and assignment scenarios, we focus on how well the estimators meet the needs of policymakers, considering how VAM-based measures of teacher effectiveness might typically be used in educational settings. If districts wish only to rank teachers in order to identify those who are high or low performing, then estimators that come close to getting the rankings right are the most desirable. For the purposes of structuring rewards and sanctions or identifying teachers in need of professional development, districts may wish primarily to distinguish high and low performing teachers from those who are closer to average; if so, it is important that the estimators accurately classify teachers whose performance falls in the tails of the distribution.

Our study investigates the performance of various estimators with respect to these criteria, using summary measures described in the next section.

<5 Methods>

Our investigations consist of a series of simulations in which we use generated data to study how well each estimator recovers true effects under different scenarios. These scenarios are captured in data generating processes (DGPs) that vary the mechanisms used to assign students to teachers in the ways discussed in the previous section. To the generated data we apply the set of estimators discussed in Section 3. We then compare the resulting estimates of the teacher VAMs with the true underlying effects.

<5.1 Data Generating Processes>

To isolate fundamental problems, we restrict the DGPs to a relatively narrow set of idealized conditions. We assume that test scores are perfect reflections of the sum total of a child’s learning (that is, with no measurement error) and that they are on an interval scale that remains constant across grades. We assume that teacher effects are constant over time and that unobserved child-specific heterogeneity has a constant effect in each time period. We assume there are no time-varying child or family effects, no school effects, no interactions between students and teachers or schools, and no peer effects. We also assume that the GDL assumption holds – namely, that decay in schooling effects is constant over time. In addition, we assume no serial correlation. Finally, there are no time effects embedded in our DGPs.

Our data are constructed to represent three elementary grades that normally undergo standardized testing in a hypothetical district. To mirror the basic structural conditions of an elementary school system for, say, grades 3 through 5 over the course of three years, we create data sets that contain students nested within teachers nested within schools, with students followed longitudinally over time. Our simple baseline DGP is as follows:

$$A_{i3} = \lambda A_{i2} + \beta_{i3} + c_i + e_{i3} \qquad (12)$$

$$A_{i4} = \lambda A_{i3} + \beta_{i4} + c_i + e_{i4}$$

$$A_{i5} = \lambda A_{i4} + \beta_{i5} + c_i + e_{i5}$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of child *i* entering third grade, A_{i3} is the achievement score of child *i* at the end of third grade, λ is a time constant persistence parameter, β_{it} is the teacher-specific contribution to growth (the true teacher value-added effect), c_i is a time-invariant child-specific effect, and e_{it} is a random deviation for each student. Because we assume independence of e_{it} over time, we are maintaining the common factor restriction in the underlying cumulative effects model. We assume that the time-invariant child-specific heterogeneity c_i is correlated at about 0.5 with the baseline test score A_{i2}.[8]

In the simulations reported in this paper, the random variables A_{i2}, β_{it}, c_i, and e_{it} are drawn from normal distributions, where we adjust the standard deviations to allow different relative contributions to the scores. It is somewhat challenging to anchor our estimates of teacher effect sizes to those in the literature, however, because reported teacher-related variance components range from as low as 3 percent to as high as 27 percent and obtained through different estimation methods (e.g., Nye et al. 2004, McCaffrey et al. 2004, Lockwood et al. 2007). Estimates in the smaller end of the range—i.e., around 5 percent—are more frequently reported. In our own investigations of data from a set of districts, however, we found rough estimates of teacher effects tending toward 20 percent of the total variance in gain scores but highly variable across

[8] Other work by the authors (Reckase et al., 2013) finds that when test scores are generated as in (13) such correlation—which seems realistic—is necessary to achieve data that conform to the parameter estimates derived from observed achievement distributions.

districts. Thus in our simulations, we explore two parameterization schemes. In the first, the standard deviation of the teacher effect is .25, while that of the student fixed effect is .5, and that of the random noise component, e_{it}, is 1. The components represent approximately 5, 19, and 76 percent of the total variance in gain scores, respectively. In the second parameterization, the standard deviation of the teacher effect is .6, while that of the student fixed effect is .6, and that of the random noise component is 1, representing approximately 21, 21, and 58 percent of the total variance in gain scores, respectively. Thus, in the latter scenario, teacher effects are relatively more important and should be easier to estimate.[9]

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

- 10 schools
- 3 grades (3rd, 4th, and 5th) of scores and teacher assignments, with a base score in 2nd grade
- 4 teachers per grade at each school (thus 120 teachers overall)
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools (except in sensitivity analyses)

To create different scenarios, we vary certain key features: the grouping of students into classes, the assignment of classes of students to teachers within schools, and the amount of decay in prior learning from one period to the next. We generate data using each of the 10 different mechanisms for the assignment of students outlined in Table 1.[10] We vary the persistence parameter λ as follows: (1) $\lambda = 1$ (no decay or complete persistence) and (2) $\lambda = .5$ (fairly strong

[9] See the Appendix for a summary of the simulation parameters used.

[10] We introduce a small amount of noise into each grouping process. The assignment noise parameter used in these simulations has a standard deviation of 1.

decay).[11] Thus, we explore 10 × 2 = 20 different baseline scenarios in this paper.[12] We use 100 Monte Carlo replications per scenario in evaluating each estimator.

<5.2 Methods for Estimating Teacher Effects>

We estimate the teacher effects using modified versions of the estimating equations (5) and (9). The modified equations reflect the simplifications determined by our DGPs. Specifically, we remove the time-varying intercept because our data have no time effects, we have no time-varying child and family effects, and we assume that $\pi_t = 1$. It is useful to write the equations as

$$\Delta A_{it} = E_{it}\beta_0 + c_i + e_{it} \tag{13}$$

$$A_{it} = \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + e_{it} \tag{14}$$

$$\Delta A_{it} = \lambda \Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta e_{it} \tag{15}$$

where E_{it} is the vector of 119 teacher dummies (with one omitted because every estimation method includes an intercept, either explicitly or by accounting for c_i). These equations are written to reflect the DGPs used to obtain the simulated data and also indicate the estimating equations used by the various procedures. The heterogeneity, c_i appears in all simulations, and so it is part of the error term in equations (13) and (14) regardless of the estimation method used. Pooled OLS (POLS) uses (13) as the estimating equation but ignores the composite nature of the error term $v_{it} = c_i + e_{it}$. RE is based on (13) but it exploits the presence of c_i in a GLS analysis. The fixed effects (FE) estimator also starts with (13) but eliminates c_i using the within-student transformation. Dynamic OLS (DOLS) is OLS applied to (14) ; as with POLS, c_i is ignored and treated as just another part of the error term. The AB estimator is instrumental variables applied

[11] Rough estimates of λ in real data cover a wide range of values. Andrabi et al. (2011) find persistence rates of .5 or lower in Pakistani schools.

[12] Several more scenarios that relax various assumptions are added in sensitivity analyses discussed in a later section of the paper.

to (15), which is in differenced form and so is free of the heterogeneity. Finally, the average residual (AR) approach is essentially based on (14) (ignoring c_i), but it only nets out the lagged test score from the current test score.For each of the 100 iterations pertaining to each DGP we estimate the teacher effects using each of the six estimation methods summarized above. We use the statistical software Stata for all data generation and estimation.

<5.2 Summary Statistics for Evaluating the Estimators>

For each iteration and for each of the six estimators, we save the estimated individual teacher effects, which are the coefficients on the teacher dummies, and also retain the true teacher effects. To study how well the methods uncover the true teacher effects, we adopt some simple summary measures. The first is a measure of how well the estimates preserve the rankings of the true effects. We compute the Spearman rank correlation, $\hat{\rho}$, between the estimated teacher effects, $\hat{\beta}_j$, and the true effects, β_j, and report the average $\hat{\rho}$ across the 100 iterations.

Because we compute teacher effect estimates for three grades in the same analysis in order to include panel data estimators in our evaluation, we standardize the teacher effect estimates within each grade before computing the rank correlations. We do so to adjust for differences in the estimated mean teacher effects and their standard deviations across grades that result from changes in average achievement gains due to the various ways in which the data are generated. This post-estimation standardization is only needed because we summarize the rank correlations across all three grades for ease of inspection. In our simulations, there is no difference in the distribution of true teacher effects across grades—therefore we impose the same condition on the teacher effect estimates to avoid artificially tamping down rank correlations in certain scenarios.

Second, we compute two measures of misclassification. The first is the percentage of above average teachers (in the true quality distribution) who are misclassified as below average in the distribution of estimated effects. The second focuses on the tails of the quality distribution. We determine which teachers are estimated to be in the bottom 20 percent and then display the proportion of teachers at each percentile of the true effect distribution who are classified in this category using graphs.

<6 Simulation Results>

Below we present results for our six estimation approaches under different simulated conditions. We first discuss in depth the findings for a baseline set of scenarios in which students and teachers are randomly sorted into schools and λ = 1. We then discuss departures from the patterns exhibited in the baseline scenarios when λ = .5 and when teacher effects are larger. Several additional sensitivity analyses are then summarized in the final results section. These include the introduction of nonrandom sorting of students and teachers across schools, measurement error, serial correlation, and student mobility across schools.

<6.1 Baseline Scenarios with No Decay>

Results for the case in which students and teachers are randomly sorted into schools and $\lambda = 1$ are shown in the left side of Table 2. The underlying parameterization scheme used here is the one in which teacher effects represent only five percent of the total variance in gain scores. Each cell in Table 2 contains two numbers specific to the particular estimator-scenario combination. The first is the average rank correlation between the estimated and true teacher effects over the 100 replications. The second is the average percentage of above average teachers who are misclassified as being below average.

We expect all estimators to work well when students and teachers are both randomly assigned to classes – the RG-RA scenario defined in Table 1. Of course, the estimated teacher effects still contain sampling error, and so we do not expect to rank or classify teachers perfectly using these estimates. We find that DOLS, AR, POLS, and RE yield rank correlations near .87 or higher, with RE producing a rank correlation of about .88. The findings indicate that RE is preferred under RG-RA with no decay, something that econometric theory leads us to expect because RE is the (asymptotically) efficient estimation method. However, POLS, AR, and DOLS produce very similar results. FE and AB have rank correlations well under .65, with the correlation for AB being the worst at .59. The relatively poor performance of the FE and AB estimators is not due to any systematic bias in the estimated teacher effects; in fact, the FE estimator is unbiased (and consistent) in this scenario. (Technically, we can only say that the AB estimator is consistent.) Rather, the relatively poor performance of FE and AB is due to their large sampling variances: both estimators unnecessarily remove a student effect in this scenario. In addition, AB unnecessarily estimates a coefficient on the lagged test score, and uses instrumental variables in doing so.

The DOLS, AR, POLS, and RE estimators are also better at classifying the teachers than the other two methods, incorrectly classifying an above average teacher as being below average about 15% of the time. The misclassification rates for FE and AB, on the other hand, are 26% and 27%, respectively. Clearly, the estimation error in the teacher effects using FE and AB has important consequences for using those estimates to classify teachers.

The potential for misclassification is explored further in Figure 1 for selected scenarios and estimators. The true teacher percentile rank is represented along the x-axis, and the y-axis represents the proportion of times in which a teacher at a particular true percentile is classified in the bottom quintile of the distribution on the basis of his or her estimate. Thus, a perfect estimator would produce the step function traced on the graph, with $y = 1$ when x ranges from 0 to 20 and $y = 0$ when x ranges from just above 20 to 100. The first graph in Figure 1 shows the superiority of DOLS, POLS, AR, and RE over FE in the RG-RA scenario with lambda equal to one. However, it should be noted that even for these estimators under these idealized conditions, identification of the “worst” teachers appears subject to a nontrivial amount of error.

Nonrandom student grouping mechanisms have relatively minor consequences for RE, DOLS, AR, and POLS provided the *teachers are randomly assigned* to classrooms – whether the students are grouped according to their prior scores (DG-RA), baseline scores (BG-RA), or heterogeneity (HG-RA).[13] Their rank correlations all remain above .8 in these scenarios. Generally, nonrandom grouping of students causes POLS and RE to do less well in terms of precision – especially when grouping is based on student heterogeneity – most likely because the student grouping induces cluster correlation within a classroom.[14] Nevertheless, they continue to yield relatively high correlations, ranging from .80 to .85. The methods that remove the student effect, FE and AB, behave similarly to the RG/RA scenario, which means they continue to do a much worse job in ranking and classifying teachers. Because the teacher assignments are not systematically related to student heterogeneity, FE and AB still unnecessarily remove a student effect and do worse because of the large sampling variances (not bias).

When students are *nonrandomly grouped* in classrooms and the classrooms are also *nonrandomly assigned* to teachers, the properties of the estimation procedures change markedly

[13] All random assignment scenarios are shown in shaded cells in the tables.

[14] The consequences of clustering are easy to understand with a method such as POLS. The POLS estimates are simply within-teacher averages of the student gain scores. The sampling method that gives the most precise estimates of a population average is random sampling, where observations are independent (as in the RG-RA design). With cluster correlation each new student effectively adds less information because that student’s outcome is correlated with other students’ outcomes in the cluster.

– and the behavior of the estimators depends critically on the nature of the nonrandom assignment. Across all the DG scenarios, DOLS is the most robust estimator, with its lowest correlation being a healthy .86. Although DOLS leaves c_i in the error term, DOLS controls directly for the assignment mechanism in this scenario – that is, lagged test scores. (And, relative to the variance of the idiosyncratic error u_{it}, the variance of the student heterogeneity component is not especially large in this simulation.) The other estimators all experience systematic biases that move the rank correlations in different ways depending upon whether assignment is positive or negative.

POLS and RE do well in ranking teachers in the DG-PA scenario – exceeding even the performance of DOLS – but they do poorly in the DG-NA scenario (with both rank correlations equal to only .30). Both POLS and RE leave c_i in the error term, and this heterogeneity term is highly, positively correlated with the lagged test score. Consequently, in the DG-PA and DG-NA scenarios the teacher assignment is correlated with the error term, and this correlation results in systematic bias. The bias actually helps with ranking the teachers in the positive assignment case but it hurts in the negative assignment case. Generally, the rank correlations are typically improved when the distance of the estimated VAMs from the mean teacher is inflated; conversely, the rank correlations suffer when that distance is compressed. Under DG-PA, better teachers are assigned to students with higher previous test scores, and therefore, on average, to students with higher c_i. This assignment mechanism implies an upward bias for good teachers and a downward bias for poor teachers. The resulting amplification in the bias actually makes it easier to rank teachers. However, under negative assignment (DG-NA), the bias works in the opposite direction: all estimated teacher effects are compressed toward the overall mean, making ranking much more difficult.[15]

The AR method is clearly inferior to DOLS for both DG-PA and DG-NA because, as we discussed in Section 3.2, AR does not partial out the effect of lagged test scores from the teacher assignment indicators. In the DG-NA case the rank correlation for AR is only .63 and 28% of above-average teachers are misclassified, compared with .87 and 16%, respectively, for DOLS.

In contrast to DOLS, POLS, and RE, the FE estimator performs very poorly under positive assignment, actually resulting in a negative rank correlation (−.31). The reason for this dismal performance is that FE is systematically biased toward zero, a phenomenon that can be most easily explained in the case of two grades per student (but the argument holds more generally). With only two grades, the FE estimator is the same as the OLS estimator obtained from the first-differenced (FD) equation, as in (10). When better teachers are assigned to students that had higher previous test scores – the DG-PA scenario – the change in teacher assignment for good teachers is *negatively* correlated with the change in the idiosyncratic error. This causes a downward bias in the FD (FE) estimator, much like the bias caused by including a lagged dependent variable in first-differencing estimation – see, for example, Wooldridge (2010, Chapter 11). For teachers with low ability, the bias goes the other way. Consequently, in the DG-PA case the FE estimated teacher effects are biased toward zero, again making it difficult if not impossible to sort out the different teachers. In our simulations the bias is so extreme that the FE estimates are negatively correlated with the actual teacher effects.

[15] This point is easiest to see in the simple case where there are two teachers, with the omitted teacher in the regression being the less effective of the two. When the chance of getting the better teacher is positively correlated with the error, the coefficient on the better teacher indicator will be upward biased. When assignment is negative, the coefficient on the better teacher will be biased downward.

The Arellano and Bond estimator does little to help in the DG-PA case. It instruments for the lagged change in the test score, but it is the change in teacher assignment that is endogenous in the FD equation (10). Plus, AB does not properly partial out the lagged level of the test score from teacher assignment; instead, it includes the lagged change, not the lagged level. AB would likely work better if assignment in grade *t* depended only on the test score two grades earlier, but this assignment scenario seems unrealistic.

Overall, the estimators that remove the student-level heterogeneity – FE and AB – perform very poorly in the DG-PA scenario. FE misclassifies above average teachers as being below average in 57% of all cases. The poor performance of FE is highlighted in second graph in Figure 1, which vividly illustrates how the best teachers are more likely to be classified as underperforming than the worst ones. In this type of scenario – with students grouped on the basis of past test scores and assigned to teachers whose performance tends to match their own – FE and AB are so biased as to be at best unhelpful, and possibly harmful, for distinguishing among teachers. In the DG-PA case, the problem is mostly due to bias, as can be seen by comparing the rank correlations in the DG-RA case to those in the DG-PA case.

With dynamic grouping and negative assignment the FE estimator (though not AB) works pretty well – second only to DOLS. But this is essentially a fluke, just like the strong performance of RE in the DG-PA case. In the DG-NA case, teacher assignment in the FD equation is positively correlated with the change in the error, and this positive correlation produces a bias away from zero in the estimated teacher effects. As discussed earlier, an amplified bias actually helps when the goal is to rank teachers. With DOLS we do not need to know whether the assignment is positive or negative: its performance is the same in the DG-PA and DG-NA cases (and better than FE), which is why DOLS is preferred to FE.

Nonrandom teacher assignment coupled with either of the two *static* grouping mechanisms (BG or HG) also poses challenges for several of the estimators. When $\lambda = 1$ and the grouping of students is determined by the baseline score, the DOLS and AR estimators fluctuate the least across the two scenarios with nonrandom assignment, with DOLS performing slightly better. AR is biased because it does not net out the base score, A_{i2}, from the teacher assignment. However, the bias only hurts the teacher rankings in the BG-NA case.

Technically, DOLS is biased in the BG cases with nonrandom assignment because it effectively controls for the wrong explanatory variable, $A_{i,t-1}$; it should control for the base score, A_{i2}. This claim can be shown, with $\lambda = 1$, by writing A_{it} as a function of all past inputs, shocks, and the initial (second-grade) score, A_{i2}. The resulting equation includes A_{i2} with a time-varying coefficient. We can think of $A_{i,t-1}$ acting as an imperfect proxy for this linear function of A_{i2} – a proxy that seems to work pretty well, which is not too surprising given the strong persistence in the level score

POLS and RE fluctuate greatly across the BG-PA and BG-NA scenarios, performing well in the BG-PA case and not so well in the BG-NA case (with rank correlations of .53 and .64, respectively). In the BG-PA case, the bias is amplified, for reasons similar to those described earlier: the student heterogeneity is positively correlated with the base score and better teachers are assigned to teachers with a higher base score. In the BG-NA case, the negative correlation induces a bias toward zero. If the base score and student heterogeneity were uncorrelated, POLS and RE would be consistent and better performing. [16] Because FE and AB properly remove time-constant heterogeneity, their performances are stable across the BG-PA and BG-NA scenarios.

[16] The simulation findings reported here modeled correlation between the base score and the student fixed effect, as explained in the methods section. It is unlikely that no such correlation would exist. However, we did explore that possibility in analyses not shown. POLS and RE did perform better in the BG scenarios, but, although some correlations and misclassification rates were affected, the general patterns revealed in the findings were the same whether the base score and student fixed effect were correlated or uncorrelated.

Nevertheless, the imprecision in the estimates that comes from removing a relatively small student effect causes them to perform notably worse than DOLS and AR.

The second type of static grouping mechanism (HG) combines students based on the value of c_i, the time invariant student-specific growth potential. When $\lambda = 1$, c_i is a permanent component of the gain score, which means it is a student-specific trend in the level-score equation. In other words, c_i is added, in each period, to the previous score (along with the teacher effects and shocks) to get the current score. When the students with the highest growth potential are grouped with the best teachers (HG-PA), the bias in DOLS, AR, POLS, and RE (estimators that ignore c_i) leads them to rank and classify the teachers well. But negative assignment causes them to do much worse. The last line in Table 2 shows that, in the HG-NA scenario, no estimator does very well: the highest rank correlation is .61 (FE) and the lowest misclassification rate is 26% (FE), with AB and DOLS not being far behind.

When $\lambda = 1$ and assignment is based on c_i, the FE estimator is the asymptotically efficient estimator among estimators of the teacher effects that are consistent in the presence of arbitrary correlation between c_i and teacher assignment. This theoretical result means that we expect FE to have a stable performance across the different HG scenarios, and this expectation is confirmed by the HG scenarios in Table 2. But stability across the scenarios is not necessarily a good thing, especially when it is accompanied by large sampling variances. For ranking teachers, FE performs much worse than several of the other estimators in the HG-RA and HG-PA cases. As mentioned earlier, for ranking purposes having estimates that are amplified relative to the true effects is actually helpful. Even in the HG-NA case, the unbiasedness of FE leads to only modest improvements over DOLS, AR, and RE. Evidently, the bias in DOLS, AR, and RE in the HG-

NA case is not so great as to make FE, with its larger sampling variance, more than a marginal improvement.

The third graph in Figure 1 illustrates the decline in performance of RE and DOLS relative to the scenario depicted in the first graph, and shows that the FE estimator has a slight edge in the frequency with which it places teachers in the lowest quintile.

So far, even though we have discussed only the case of nonrandom sorting of students and teachers across schools, and we have assumed no decay in learning, we can summarize some useful insights. First, the findings show that, even under these idealized conditions, some estimators perform very poorly under certain assignment mechanisms – even estimators that use the correct restriction $\lambda = 1$ in estimation. Generally, estimators that are intended to be robust to static assignment do poorly under dynamic assignment.

A useful finding, in looking across all assignment mechanisms, is that DOLS does best: it is superior under dynamic grouping with nonrandom teacher assignment and still does well for ranking teachers under most static assignment mechanisms. We can understand the relatively good performance of DOLS under the dynamic assignment mechanisms by noting that if the DGP did not include a student effect, the DOLS VAM estimators would be consistent: the teacher dummies E_{it} are correlated with $A_{i,t\text{-}1}$ but the latter is controlled for in the DOLS regression. AR includes the lagged test score in the first-step regression but does not partial out its correlation with the teacher dummies. Even though $\lambda = 1$, POLS and RE are systematically biased because they leave the heterogeneity (which is correlated with $A_{i,t\text{-}1}$) in the error term. And, as discussed earlier, FE and AB induce correlation between the transformed error and teacher assignment, rendering teacher assignment endogenous in the transformed equations. While bias can improve teacher rankings in some cases, it is very costly in others.

<6.2 Baseline Scenarios with Strong Decay>

The performance of several estimators deteriorates when we change the value of λ from 1 to $.5$. The right side of Table 2 shows simulation findings when students and teachers are randomly assigned to schools and $\lambda = .5$. Importantly, because POLS, RE, and FE act as if $\lambda = 1$, these estimators are now applied to an equation with misspecified dynamics, regardless of the assignment mechanism. Because POLS, RE, and FE use the gain score as the dependent variable, we can think of equation (13) as containing an omitted variable, $A_{i,t-1}$, with coefficient $-.5$ on it;[17] this is important for understanding the findings in Table 2.

Interestingly, dynamic misspecification has little effect on the quality of the teacher rankings in the RG-RA scenario. FE suffers the most, but even its rank correlation only falls from .62 to .57. Some of the other estimators perform slightly worse with $\lambda = .5$, but with 100 simulations the small differences could be explained by simulation error.

Due to the negative sign on the omitted lagged test score in the gain-score equation, the fluctuations in the rank correlations across the PA and NA scenarios are reversed from the situation where $\lambda = 1$. For example, the POLS estimator, which did very well in the DG-PA case when $\lambda = 1$, is essentially useless when $\lambda = .5$ (rank correlation = .11)[18]. By contrast, in the HG-NA case the POLS estimator is the best by some margin, giving a rank correlation of .75 compared with .56 for DOLS. Figure 2 further illustrates that the preferred estimator changes across the different scenarios. Because we can never know which scenario is relevant – and in practice assignment is likely to be even more complicated – it is difficult to know how to choose

[17] This is because the gain score specifications subtract $A_{i,t-1}$ from both sides of the equation, thus leaving $(\lambda - 1)A_{i,t-1}$ in the error term.

[18] When $\lambda = .5$, in all simulation runs the RE and POLS estimates are identical. This equivalence happens because of the substantial negative serial correlation in the composite error term caused by omitting the variable $-.5A_{i,t-1}$. The negative serial correlation is so severe that the usual RE variance estimate of c_i is always negative. In such cases, Stata sets the variance to zero, which then leads to the POLS estimate.

the best estimator. Nevertheless, taken as a whole, the simulations reported in Table 2 point to several conclusions. While DOLS is not uniformly better across all of the grouping, assignment, and decay assumptions, it is nearly so. DOLS is preferred under dynamic grouping as it outperforms the other estimators by a large margin, with the exception of AR which is often close to DOLS. Unlike DOLS, and to a lesser extent AR, the other estimators show much more sensitivity to the value of λ. The robustness of DOLS makes it the recommended approach among the group of estimators considered in this study. However, we should note that the potential for misclassification in these simple DGPs, even using DOLS, can approach levels that might be considered unacceptable for policy purposes.

<6.3 Baseline Scenarios with Large Teacher Effects>

We also conducted a set of simulations where teacher effects represent a much larger relative share of the total variance in student gains. These are reported in Table 3. As to be expected, when the size of the teacher effects is raised relative to the student effect and shock, rank correlations improve and misclassification rates decline somewhat. The same overall patterns observed in the “small” teacher effects case continue to hold. The relative superiority of DOLS over AR in the DG scenarios and over POLS and RE in the scenarios with strong decay is still evident when teacher effects are large. The FE and AB estimators improve their rank correlations in many scenarios when teacher effects are large but remain the least effective estimators overall. Although concerns over inaccuracy in the estimates and rankings are mitigated when teacher effects are large, the same lessons regarding which estimator to use in particular contexts apply, and the overall conclusion that DOLS is more robust across scenarios holds.

<6.4 Sensitivity Analyses>

We subjected our simulations to several sensitivity analyses, relaxing many of the assumptions used to form our baseline DGPs. First, we looked at the impact of nonrandom sorting of students and teachers *across* schools. In these cases, DOLS continued to show consistently high correlations across assignment scenarios and to outperform AR in the DG-PA and DG-NA scenarios, although all correlations were slightly lower than in the random school sorting cases. POLS and RE deteriorated slightly when students were nonrandomly sorted across schools, probably because there is less information for estimating teacher VAMS when students have similar abilities. An unresolved puzzle is that FE and AB deteriorated substantially when teachers were nonrandomly sorted across schools.

We also ran a full set of simulations with $\lambda = .75$ (more moderate decay), without any surprises. This implies a less severe form of dynamic misspecification for estimators such as POLS and RE than the $\lambda = .5$ case. It is not surprising that the performance of POLS and RE was essentially between the $\lambda = 1$ and $\lambda = .5$ cases. The DOLS estimator was hardly affected by the value of λ. Nor was AR, but it was still generally outperformed by DOLS.

We also added classical measurement error to the test scores in our DGPs. This did little to affect the patterns reported above. DOLS continued to yield the best relationship to the true teacher effects across scenarios.

In addition, we ran simulations in which serial correlation was introduced in the errors. We did this to relax the common factor restriction upon which DOLS and AB relied for consistency. AB performed worse when $\lambda = 1$, but DOLS remained largely unaffected and retained its status as the most robust estimator across scenarios.

Finally, we examined the performance of the estimators when student mobility across schools was present. When we allowed 10 percent of students to switch schools in each year, FE and AB improved in the random and heterogeneity grouping scenarios – likely because having more variation in students within a school that comes from allowing random mobility helps with precision. But FE and AR still suffer from severe bias in the dynamic grouping-nonrandom assignment scenarios. DOLS remained the most robust estimator. For all sensitivity analyses, details are available from the authors upon request.

<7 Conclusions and Future Directions>

Simulated data with known properties permits the systematic exploration of the ability of various estimation methods to recover the true parameters used to generate the data – in our case teacher effects. This study has taken the first step in evaluating different value-added estimation strategies under conditions in which they are most likely to succeed. Creating somewhat realistic but idealized conditions facilitates the investigation of issues associated with the use of particular estimators. If they perform poorly under these idealized conditions, they will almost certainly do worse in real settings.

A main finding is that no one method is guaranteed to accurately capture true teacher effects in all contexts even under these relatively idealized conditions, although some are more robust than others. Because we consider a variety of DGPs, student grouping mechanisms, and teacher assignment mechanisms, it is not surprising that no single method works well in all hypothetical contexts. Both the teacher assignment mechanism and the nature of the dynamic relationship between current and past achievement play important roles in determining how well the estimators function.

A dynamic specification estimated by OLS—what we have called DOLS – was, by far, the most robust estimator across scenarios. Only in one scenario – heterogeneity-based grouping with negative assignment – did it fail to produce useful information with regard to teacher effects. However, none of our estimators was able to surmount the problems posed by this scenario – not even estimators designed to eliminate bias stemming from unobserved heterogeneity – and it is likely a less realistic scenario than others we considered.

In all other situations, DOLS provided estimates of some value. The main strength of this estimator lies in the fact that, by including prior achievement on the right-hand side, it controls either directly or indirectly for grouping and assignment mechanisms. In the case of dynamic grouping coupled with non-random assignment, it explicitly controls for the potential source of bias. It should be further noted that even in a dynamic grouping situation with mixed assignment, in which some principals use negative assignment and some use positive assignment, DOLS would work well. In the case of baseline and heterogeneity grouping, the effect of controlling for prior achievement is less direct but still somewhat effective in that both those grouping mechanisms are correlated with prior achievement.

These findings suggest that choosing estimators on the basis of structural modeling considerations may produce inferior results by drawing attention to relatively unimportant concerns and away from key concerns. The DOLS estimator is never the prescribed approach under the structural cumulative effects model with a geometric distributed lag (unless there is no student heterogeneity), yet it is often the best estimator. Even when we introduced serial correlation and measurement error, DOLS was not particularly harmed. One can think of the DOLS estimator as a regression-based version of a dynamic treatment effects estimator. That is not to say that the general cumulative effects model is incorrect. It merely reflects the fact that efforts to derive consistent estimators by focusing on particular concerns of structural modeling (for example, heterogeneity, endogenous lags) may obscure the fact that controlling for the assignment mechanism even in specifications that contain other sources of endogeneity is essential. Approaches that attend to less important features of the structural model, when coupled with nonrandom assignment, may yield estimators that are unduly constrained and thus poorly behaved. The poor performance of the AB estimator exemplifies this. By differencing for heterogeneity and using instrumental variables to remove bias from the estimation of lambda, it loses much of its ability to estimate teacher effects precisely. Another example is the AR approach, which coupled with “shrinkage” (as in commonly used methods frequently labeled empirical Bayes) is currently very popular in the research literature, misses the opportunity to control for nonrandom assignment. The findings in this paper suggest that flexible approaches based on dynamic treatment effects (for example, Lechner, 2008; Wooldridge, 2010, Chapter 21) may be more fruitful than those based on structural modeling considerations.

Finally, despite the relatively robust performance of DOLS, we find that even in the best scenarios and under the simplistic and idealized conditions imposed by our data generating process, the potential for misclassifying above average teachers as below average or for misidentifying the “worst” or “best” teachers remains nontrivial, particularly if teacher effects are relatively small. Applying the commonly used estimators to our simplified DGPs results in misclassification rates that range from at least seven to more than 60 percent, depending upon the estimator and scenario.

It is clear from this study that certain VAMs hold promise: they may be capable of overcoming many obstacles presented by non-random assignment and may yield valuable information, providing assignment mechanisms are known or can be deduced from the data. Our findings indicate that teacher rankings can correlate relatively well with true rankings in certain scenarios and that, in some cases, misclassification rates may be relatively low. Given the context-dependency of the estimators’ ability to produce accurate results, however, and our current lack of knowledge regarding prevailing assignment practices, VAM-based measures of teacher performance, as currently applied in practice and research, must be subjected to close scrutiny regarding the methods used and interpreted with a high degree of caution.

Methods of constructing estimates of teacher effects that we can trust for high-stakes evaluative purposes must be further studied, and there is much left to investigate. This paper does not address the degree to which test measurement error, school effects, time-varying teacher effects, different types of interactions among teachers and students, and compensating or reinforcing contemporaneous family effects alter the performance of the estimators.

Clearly, although value-added measures of teacher performance hold some promise, more research is needed before they can confidently be implemented in high-stakes policies. Our findings suggest that teacher effect estimates constructed using DOLS may be useful in answering research questions that employ them in regression specifications. The degree of error in these estimates, however, may make them *less trustworthy* for the specific purpose of evaluating individual teachers. It may be argued that including these measures in a comprehensive teacher evaluation along with other indicators could provide beneficial information and represent an improvement over the status quo. However, it would be unwise to use these measures as the sole basis for sanctions. Even if such measures are released to the public simply as information—as has been the case in Los Angeles and New York City—the potential for inaccuracy, and thus for damage to teachers' status and morale, creates risks that could outweigh the benefits. If such measures are accurate, then publicizing or attaching incentives to them may motivate existing teachers to increase efforts or induce individuals with high performance potential into the teaching profession. If, however, such measures cannot be trusted to produce fair evaluations, existing teachers may become demoralized and high potential individuals considering teaching as a profession may steer away from entering the public school system.

Given that the accuracy of VAM-based measures of teacher performance can vary considerably across contexts and that the potential for bias if particular methods are applied to the wrong situations is nontrivial, we conclude that it is premature to attach high stakes to these measures until their properties have been better understood.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25(1): 95-135.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc. 2011. Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics* 3(3): 29-54.

Arellano, Manuel, and Stephen Bond. 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies* 58(2): 277-297.

Ballou, Dale, William Sanders, and Paul Wright. 2004. Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics* 29(1): 37-65.

Blundell, Richard, and Stephen Bond. 1998. Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87(1): 115-143.

Boardman, Anthony E., and Richard J. Murnane. 1979. Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education* 52: 113-121.

Buddin, Richard. 2011. Measuring teacher and school effectiveness at improving student achievement in Los Angeles elementary schools. MRPA Paper No. 31963.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper No. 17699.

Dee, Thomas S. 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics* 86(1): 195-210.

Dieterle, Steven G., Cassandra M. Guarino, Mark D. Reckase, and Jeffrey M. Wooldridge. 2012. How do principals group and assign students to teachers? Finding evidence in administrative data and the implications for value-added. Michigan State University Education Policy Center Working Paper No. 30.

Downey, Douglas B., Paul T. Von Hippel, and Beckett A. Broh. 2004. Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review* 69(5): 613-635.

Entwisle, Doris R., and Karl L. Alexander. 1992. Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *American Sociological Review* 57(1): 72-84.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul Thompson, and Jeffrey M. Wooldridge. 2012. An evaluation of empirical Bayes' estimation of value-added teacher performance measures. Michigan State University Education Policy Center Working Paper No. 31.

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

Hanushek, Eric A. 1986. The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* 24(3): 1141-1177.

Harris, Douglas N., Tim R. Sass, and Anastasia Semykina. 2011. Value-added models and the measurement of teacher productivity. Unpublished paper, Florida State University.

Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working paper No. 14607.

Koedel, Cory, and Julian R. Betts. 2011. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy* 6(1): 18-42.

Lechner, Michael. 2008. Matching estimation of dynamic treatment models: Some practical issues. In *Advances in Econometrics Volume 21*, edited by Daniel Millimet, Jeffrey Smith, and Edward Vytlacil, pp. 289-333. Amsterdam: Emerald Group Publishing Limited.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. 2004. Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics* 29(1): 67-101.

McGuinn, Patrick. 2012. *The state of teacher evaluation reform: State Education Agency capacity and the implementation of new teacher-evaluation systems*. Washington, DC: Center for American Progress.

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

Raudenbush, Stephen W. 2009. Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy* 4(4): 468-491.

Guarino, Cassandra M., Michelle Maxfield, Mark D. Reckase, Paul Thompson, and Jeffrey M. Wooldridge

Reckase, Mark D., Eun Hye Ham, Cassandra M. Guarino, and Jeffrey M. Wooldridge. 2013. What can be learned from simulation studies of value-added models? Unpublished paper, Michigan State University.

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

Sanders, William L., and Sandra P. Horn. 1994. The Tennessee value-added assessment system (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education* 8(3): 299-311.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *The Economic Journal* 113(485): 3-33.

Santos, Fernanda and Robert Gebeloff. 2012. Teacher quality widely diffused, ratings indicate. *The New York Times*. Available http://www.nytimes.com/2012/02/25/education/teacher-quality-widely-diffused-nyc-ratings-indicate.html. Accessed 4 May 2012.

US Department of Education. 2009. *Race to the top program: Executive summary*. Available http://www2.ed.gov/programs/racetothetop/executive-summary.pdf. Accessed 8 September 2010.

Value-Added Research Center. 2010. *NYC teacher data initiative: Technical report on the NYC value-added model*. Available http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf. Accessed 15 May 2012.

West, Martin, and Matthew Chingos. 2009. Teacher effectiveness, mobility, and attrition in Florida. In *Performance Incentives: Their Growing Impact on American K–12 Education*, edited by Matthew G. Springer, pp. 251-271. Washington, DC: Brookings Institution Press.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* 2nd edition. Cambridge, MA: MIT Press.

Zeger, Scott L., Kung-Yee Liang, and Paul S. Albert. 1988. Models for longitudinal data: A generalized estimating equation approach. *Biometrics* 44(4): 1049-1060.

Table 1: Grouping and Assignment Acronyms

Acronym	**Process for grouping students in classrooms**	**Process for assigning students to teachers**
RG-RA	Random	Random
DG-RA	Dynamic based on prior test scores	Random
DG-PA	Dynamic based on prior test scores	Positive correlation between teacher effects and prior student scores (better teachers with better students)
DG-NA	Dynamic based on prior test scores	Negative correlation between teacher effects and prior student scores
BG-RA	Static based on baseline test scores	Random
BG-PA	Static based on baseline test scores	Positive correlation between teacher effects and baseline student scores
BG-NA	Static based on baseline test scores	Negative correlation between teacher effects and baseline student scores
HG-RA	Static based on heterogeneity	Random
HG-PA	Static based on heterogeneity	Positive correlation between teacher effects and student fixed effects
HG-NA	Static based on heterogeneity	Negative correlation between teacher effects and student fixed effects

Table 2: Results from 100 replications with random sorting of students and teachers across schools and small teacher effects. Row 1: Average rank correlation. Row 2: Percentage of above average teachers misclassified as below average.

Small Teacher Effects	λ=1						λ=.5					
Estimator Assignment Mechanism	DOLS	AR	POLS	RE	FE	AB	DOLS	AR	POLS	RE	FE	AB
RG-RA	0.87 15%	0.88 15%	0.87 15%	0.88 15%	0.62 26%	0.59 27%	0.87 15%	0.87 15%	0.85 16%	0.85 16%	0.57 28%	0.59 27%
DG-RA	0.87 15%	0.87 15%	0.80 19%	0.85 16%	0.58 27%	0.53 29%	0.87 15%	0.87 15%	0.76 21%	0.76 21%	0.48 31%	0.53 28%
DG-PA	0.86 16%	0.80 19%	0.90 14%	0.90 14%	-0.31 57%	-0.01 48%	0.87 15%	0.83 18%	0.11 47%	0.11 47%	-0.44 61%	-0.08 51%
DG-NA	0.87 16%	0.63 28%	0.30 42%	0.30 42%	0.73 20%	-0.48 63%	0.87 15%	0.69 26%	0.89 14%	0.89 14%	0.72 21%	0.74 20%
BG-RA	0.87 15%	0.87 15%	0.83 18%	0.85 17%	0.62 25%	0.60 27%	0.87 16%	0.86 16%	0.83 18%	0.83 18%	0.57 28%	0.60 27%
BG-PA	0.89 14%	0.88 15%	0.90 13%	0.91 12%	0.61 25%	0.59 27%	0.90 13%	0.89 13%	0.62 28%	0.62 28%	0.60 26%	0.59 27%
BG-NA	0.84 18%	0.75 23%	0.53 33%	0.64 28%	0.60 26%	0.58 27%	0.77 21%	0.71 25%	0.86 16%	0.86 16%	0.47 31%	0.58 27%
HG-RA	0.84 18%	0.84 18%	0.81 19%	0.84 17%	0.62 25%	0.59 27%	0.84 18%	0.84 18%	0.84 18%	0.84 18%	0.57 28%	0.59 27%
HG-PA	0.91 13%	0.91 13%	0.90 13%	0.90 13%	0.62 25%	0.60 27%	0.91 13%	0.91 13%	0.87 15%	0.87 15%	0.51 29%	0.60 27%
HG-NA	0.58 31%	0.54 33%	0.37 40%	0.53 33%	0.61 26%	0.58 27%	0.56 32%	0.52 34%	0.75 22%	0.75 22%	0.59 26%	0.59 27%

Table 3: Results from 100 replications, with random sorting of students and teachers across schools and large teacher effects. Row 1: Average rank correlation. Row 2: Percentage of above average teachers misclassified as below average.

Large Teacher Effects	λ=1						λ=.5					
Estimator Assignment Mechanism	DOLS	AR	POLS	RE	FE	AB	DOLS	AR	POLS	RE	FE	AB
RG-RA	0.97 7%	0.97 7%	0.97 7%	0.97 7%	0.69 23%	0.68 23%	0.97 7%	0.97 7%	0.95 9%	0.95 9%	0.63 26%	0.68 23%
DG-RA	0.97 8%	0.97 7%	0.94 10%	0.97 8%	0.69 23%	0.66 24%	0.97 8%	0.97 7%	0.92 12%	0.92 12%	0.61 26%	0.68 23%
DG-PA	0.96 8%	0.85 15%	0.96 9%	0.96 9%	0.30 37%	-0.22 54%	0.97 8%	0.89 14%	0.69 24%	0.69 24%	0.00 47%	0.56 29%
DG-NA	0.97 8%	0.84 18%	0.82 19%	0.82 19%	0.76 19%	0.38 35%	0.97 7%	0.87 16%	0.96 9%	0.96 9%	0.73 20%	0.77 19%
BG-RA	0.97 7%	0.97 7%	0.96 8%	0.97 7%	0.70 23%	0.69 23%	0.97 8%	0.97 7%	0.95 9%	0.95 9%	0.63 25%	0.69 23%
BG-PA	0.97 7%	0.95 9%	0.97 7%	0.97 7%	0.69 23%	0.68 23%	0.97 7%	0.96 8%	0.89 14%	0.89 14%	0.63 25%	0.68 23%
BG-NA	0.96 8%	0.93 11%	0.90 14%	0.93 11%	0.68 23%	0.68 23%	0.94 10%	0.93 12%	0.93 11%	0.93 11%	0.57 28%	0.68 23%
HG-RA	0.96 9%	0.96 9%	0.94 10%	0.96 9%	0.69 23%	0.68 23%	0.95 9%	0.95 9%	0.95 10%	0.95 10%	0.63 25%	0.68 24%
HG-PA	0.97 8%	0.97 8%	0.96 8%	0.97 8%	0.69 23%	0.68 23%	0.97 8%	0.97 8%	0.95 10%	0.95 10%	0.58 28%	0.69 23%
HG-NA	0.88 16%	0.86 18%	0.81 20%	0.88 15%	0.69 23%	0.68 23%	0.87 17%	0.85 18%	0.90 14%	0.90 14%	0.65 24%	0.68 23%

Appendix

Table A.1 Simulation parameters for DGP: $A_{it} = \lambda A_{i,t-1} + \beta_{it} + c_i + \varepsilon_{it}$

	“Small” teacher effects		“Large” teacher effects	
A_{i2} (base score)	$Normal(0,1)$		$Normal(0,1)$	
λ (persistence)	1	.5	1	.5
β_{it} (true teacher effect)	$Normal(0,.25)$		$Normal(0,.6)$	
c_i (fixed student effect)	$Normal(0,.5)$		$Normal(0,.6)$	
ε_{it} (random deviation)	$Normal(0,1)$		$Normal(0,1)$	
$Corr(A_{i2}, c_i)$	.5		.5	
Assignment noise under nonrandom grouping	$Normal(0,1)$		$Normal(0,1)$	
Teacher effect percentage of variance	5%		21%	

Figure 1. Small Teacher Effects, λ = 1
(thick solid = perfect classification, solid = DOLS, dash = POLS, cross = RE, dot = FE, asterisk = AR)

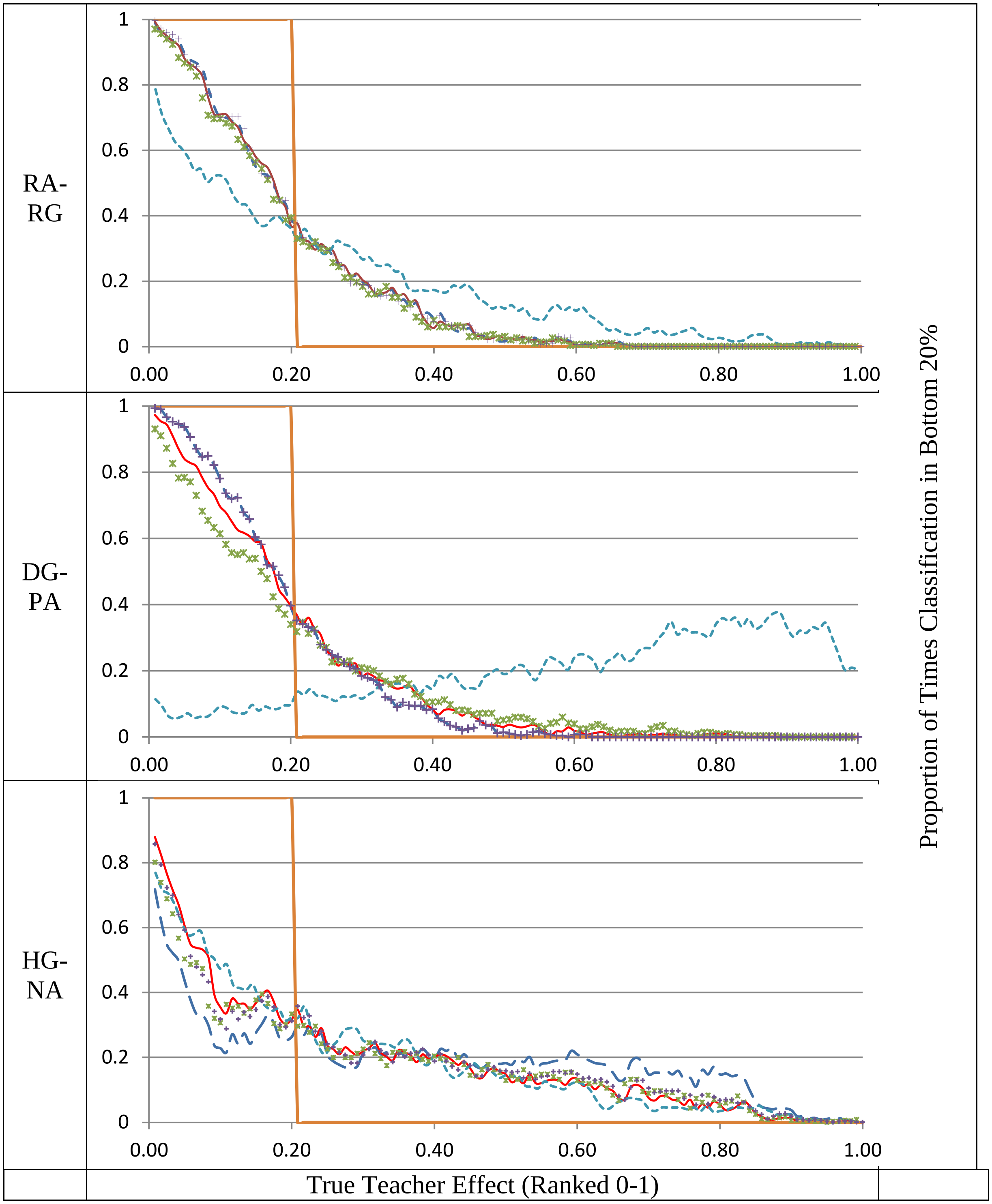
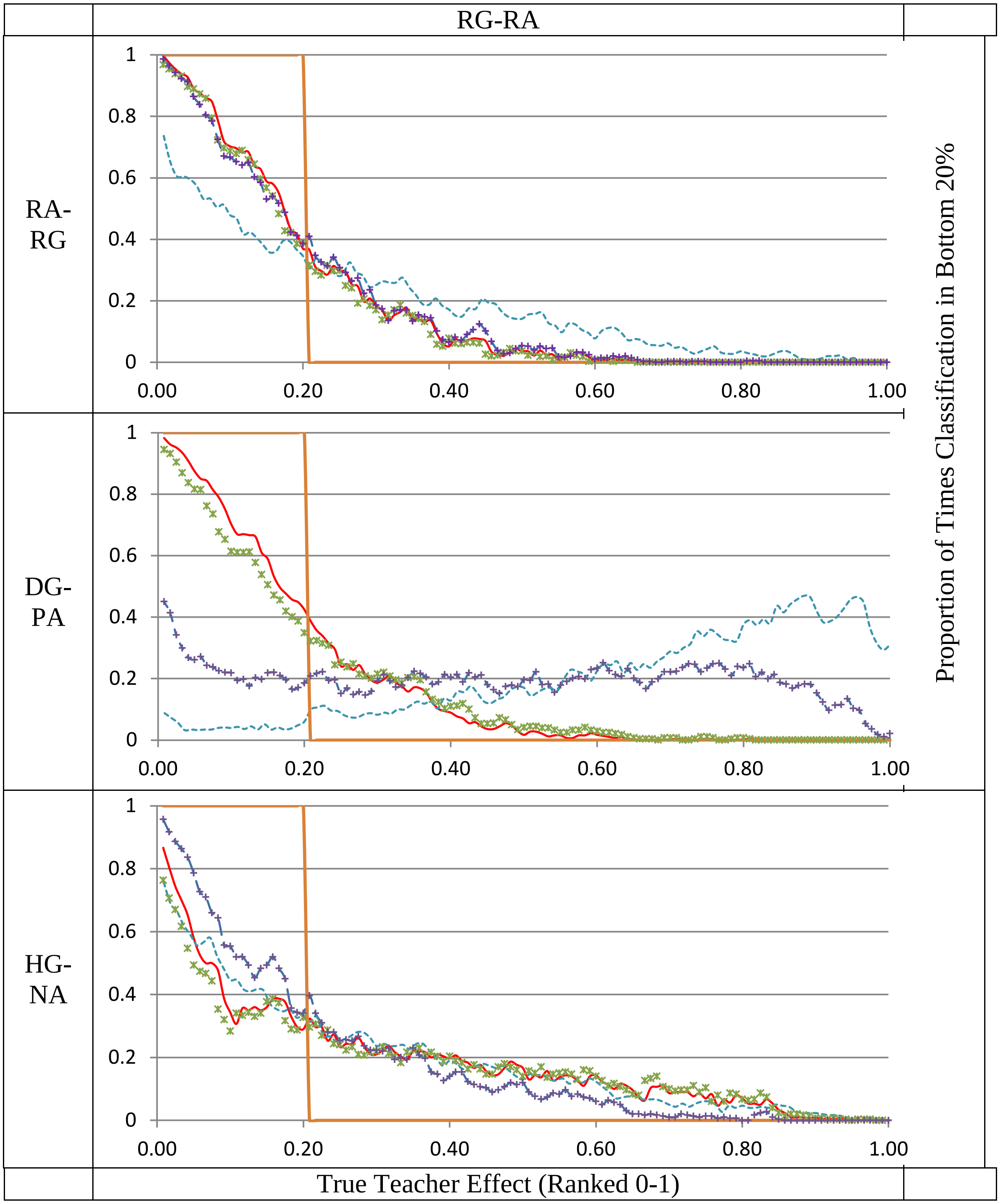


Figure 1. Small Teacher Effects, λ = .5
(thick solid = perfect classification, solid = DOLS, dash = POLS, cross = RE, dot = FE, asterisk = AR)


RG-RA
RA-RG
DG-PA
HG-NA
Proportion of Times Classification in Bottom 20%
True Teacher Effect (Ranked 0-1)