

Can Value-Added Measures of Teacher Performance Be Trusted?

By Cassandra M. Guarino, Mark D. Reckase, Jeffrey M. Wooldridge

Draft Version 03/19/11

Acknowledgment: This work was supported by grant no. R305D10002 from the Institute for Education Sciences in the U.S. Department of Education. Expert programming assistance was provided by Francis Smart. Research assistance was provided by Brian Stacy and Eun Hye Ham. Helpful comments were provided by Steven Haider, Ken Frank, Anna Bargagliotti, Steven Dieterle, Brian Stacy, Francis Smart, Cory Koedel, and Doug Harris.

Abstract: *Measures of teacher quality based on value-added models of student achievement are gaining increasing acceptance among policymakers as a possible improvement over conventional indicators based on classroom observations, educational attainment, or experience. Value-added models have been widely applied in educational research, and the use of value-added measures of performance as a component of teacher evaluation has been strongly encouraged by federal initiatives such as Race to the Top. Much controversy exists as to the validity of these measures for this purpose, however, given that students are not randomly assigned to teachers. This paper investigates whether commonly used value-added estimation strategies can produce accurate estimates of teacher effects and is the first study to show the manner in which bias is introduced. We estimate teacher effects in simulated student achievement data sets that mimic different plausible types of student grouping and teacher assignment scenarios. We then compare the estimates with the true teacher effects embedded in the data. We find that no one method accurately captures true teacher effects in all plausible assignment scenarios and that the potential for misclassifying high- and low-performing teachers can be substantial. We also find that misspecifying the dynamic relationship between current and prior achievement can exacerbate estimation problems. Certain estimation approaches predicted to be inconsistent in a structural modeling framework fare better than expected. Given that the ability of the models and estimators we examine to produce accurate teacher performance measures is context-dependent and that the potential for misclassification is nontrivial, we conclude that it is premature to attach stakes to such measures at this time.*

1. Introduction

Accurate indicators of educational effectiveness are needed to advance national policy goals of raising student achievement and closing social/cultural 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 districts¹ and widely applied in the research literature. Intuitively, VAMs are appealing; they model growth in learning from one year to the next for individual students and parse that growth into pieces believed to represent the separate contributions made by their teachers and schools as well as their own 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 reauthorization of NCLB, it is imperative that such measures be well constructed and understood.

Much controversy exists, however, as to the best way to construct VAMs and to their optimal application. Numerous methods have been developed (e.g., Sanders & Horn, 1994; Sanders et al., 1997; McCaffrey et al., 2004; Raudenbush, 2009), and studies that compare estimates derived from different

¹ In some districts, the popular press has computed and published teacher value-added scores. For example, in September 2010, the Los Angeles Times, after analyzing data obtained from Los Angeles Unified School District officials under California's Public Records Act, created a website in which any member of the public can look up VAM-based ratings for individual public school teachers in grades 3 through 5. See: <http://www.latimes.com/news/local/teachers-investigation/> (downloaded 10/12/10). In New York City, the courts have recently ruled that the district may disclose teacher evaluation measures to the public, but the decision is likely to be held up in an appeal. See: <http://www.nytimes.com/2011/01/11/education/11data.html>.

models have found substantial variability across methods (McCaffrey et al., 2004). Concerns remain that our understanding of these models is as yet limited and that incentives built around them may cause 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.

There are two central issues involved in establishing the validity of measures and inferences based on VAMs. The first 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. The few studies that have attempted to validate VAMs have drawn different conclusions (e.g., Kane & Staiger, 2008; Rothstein, 2008)², and questions about the validity of VAMs linger.

The second central issue concerns the accuracy of measures of students' achievement growth. Educational tests that under-represent the full range of desired skills and knowledge, and that substantially shift emphasis on particular constructs from year to year, will underestimate students' growth in achievement (Reckase, 1985; Reckase & Li, 2007) and may lead to statistical bias in indicators estimated using VAMs (Martineau, 2006). Moreover, different vertical scaling approaches (methods for placing test results from different grade levels on the same scale) can lead to different estimates of value-added outcomes for schools and teachers (Briggs & Weeks, 2009).

This paper is the first in a series of papers by the authors that aims to resolve these controversies. In this paper, we focus on the first of the two problems mentioned above and investigate the ability of various estimation strategies to produce accurate estimates of teacher effects. Our main research question is the following: How well do commonly used estimators perform in estimating teacher effects under a variety of known conditions, including those in which particular underlying assumptions are violated?

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. Rothstein (2008) devises falsification tests that challenge the validity of VAM-based measures of teacher performance in North Carolina.

We focus our study on five estimators that are commonly used in the research literature and in policy applications involving teacher effects. We answer our research question by first outlining the assumptions that must be met for each estimator to have good 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 linked to teachers. We generate different data sets that mimic different types of student grouping and teacher assignment scenarios. We then estimate teacher effects using each of the five estimation techniques and compare the estimates to the true teacher effects embedded in the data.

The paper is organized as follows. In Section 2, we outline a theoretical framework for value-added models based on a well-known structural cumulative effects model. The model, although it represents a somewhat simplified version of the learning process, suggests a process by which student test scores are generated and thus serves as a basis for our empirical investigation using simulations. In addition, the theoretical framework serves as a guide to the possible performance of various estimation strategies by clarifying the conditions needed for the estimators to produce consistent³ and efficient estimates of value-added parameters.

Section 3 discusses each estimator in turn and its underlying assumptions. An important component of our study is that we apply all estimators to all of our simulation conditions—even those in which they are not necessarily expected to perform well according to suppositions derived from the structural model. In this way, we can study whether some methods appear to be more robust than others across different scenarios.

We describe different mechanisms for grouping students and assigning teachers to classrooms in Section 4. As described there, we consider random, static, and dynamic tracking schemes for grouping students into classes as well as random and nonrandom teacher assignment to classrooms. We also consider random and nonrandom sorting of students and teachers into schools.

³ For the most part, the estimators we discuss—and that are widely used—rely on large sample sizes (e.g., large numbers of students) for consistency.

Section 5 describes the simulation procedures and estimation strategies we employ. The simulations 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. The simulation process serves as an explanatory tool permitting us to examine whether VAMs provide the results they should under known conditions and to study how they fail when the assumptions are violated in a known way. 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 specific conditions.

Our investigations yield several important findings, some surprising and others less so. A main finding is that no one estimator performs well under all plausible circumstances. Our most surprising finding is that certain estimation approaches known to be inconsistent in the structural modeling framework fare better than expected. Our simulations also 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.

An important caveat to apply to our findings is that they result from data generation processes that incorporate many of the primary assumptions underlying a relatively simple conceptual model. Thus, we subject the estimators to idealized conditions. Undoubtedly real-life educational conditions are more complex, and the estimators will likely perform less well when applied to real data. Detecting the flaws in various estimators under idealized conditions, however, is the best way to discover basic differences among them. Thus, the simplifications built into our research design are the strength of the design.

2. A Common Approach to Value-Added Modeling

The theoretical foundation for VAMs 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; Todd & Wolpin, 2003; Harris, Sass, & Semykina, unpublished draft):

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

where A_{it} is the achievement of child i in grade t , the X terms represent a set of relevant child and family inputs from time 0 to time t , E represents school-related inputs, the term 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 give 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, feedback and lagged responses, etc. 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 instructional programs in promoting learning is essential if we are to foster the use of successful strategies and curtail the use of ineffective approaches. Teachers and schools need to know who is performing effectively and why—that is, what instructional strategies are contributing to high performance. To do so, causal measures of performance at the school, teacher, and program level are needed. 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 a common and potentially tractable starting point for structural modeling. A_{it} is achievement for student i in grade t (for concreteness, measured at the end of the school year) and depends on inputs during school year t and possibly past inputs:

$$A_{it} = \alpha_t + E_{it}\beta_0 + E_{i,t-1}\beta_1 + \dots + E_{i0}\beta_t + X_{it}'\gamma_0 + X_{i,t-1}'\gamma_1 + \dots + X_{i0}'\gamma_t + \eta_t c_i + u_{it} \quad (2)$$

where E_{it} is a (row) vector of observed education inputs at time t – including teacher or school characteristics, or, say, teacher indicators – and X_{it} is a vector of observed time-varying individual and family characteristics such as health or disability status, socioeconomic status, and so on. The term α_t allows for a separate intercept in each time period. This is appropriate if, for example, the reporting score scales for tests at different grade levels are not the same. 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 the following assumptions embedded in it:

- The functional form is linear in the parameters.
- All parameters except the intercept and the coefficient on c_i are constant over time. For example, β_0 measures the effects of contemporaneous school inputs on achievement in every grade.
- Any family inputs prior to $t = 0$ are captured in c_i .
- All unobserved current and past factors that vary over time are in the additive, idiosyncratic shock, u_{it} .

Note that this formulation of the model does not explicitly recognize the possible presence of interactions among teachers, between teachers and students, or among students (as in peer effects) 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.

Exogeneity assumptions on the inputs are needed to estimate the parameters in the linear CEM. A common starting point 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:

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

Chamberlain (1992) referred to the condition in (3) as a *sequential exogeneity assumption*. 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. Initially, this may seem reasonable given that the timing of most school input decisions, such as teacher and class size assignments, are made at the end of the previous school year, and cannot be based on changes in a student’s situation over the course of the school year. However, u_{it} can contain factors such as unobserved parental effort that respond to the

assignment of school inputs. For example, a parent may provide more help for a student who is assigned to a poor teacher or a large class.

We should emphasize that (3) is an assumption about correlation between inputs and the time-varying unobservables, u_{it} , affecting A_{it} . It does not address the relationship between student heterogeneity (e.g., motivation or some innate component of ability), c_i , and the observed inputs. For lack of a better name, we will call lack of correlation between c_i and the inputs “heterogeneity exogeneity.” This is an important distinction because some estimation approaches either effectively ignore the presence of c_i or assume it is uncorrelated with observed inputs – in other words, they assume heterogeneity exogeneity. If, however, c_i is correlated with observed inputs—which seems likely—then 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)—for example, proxy variables, fixed effects, first-differencing—each with a set of assumptions and drawbacks. Accounting for heterogeneity, however, is not the primary obstacle to estimating this model.

The linear CEM in this general form is rarely estimated due to data limitations. To see why, suppose, for example, we can obtain testing data on 3rd grade through 6th grade for each child. If we want to allow for the possibility that all previous teachers (in this case, the E_{it} vector may be composed of teacher dummy variables) or that all previous school, classroom, and teacher characteristics (in this case, E_{it} may be composed of so-called “program” variables) affect current outcomes, we need to have data relating to students and teachers in 2nd and 1st 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 Cumulative Effects Model

One way to solve the data limitations issue and conserve on parameters in the general linear CEM is to impose restrictions on the distributed lag coefficients. The most commonly applied restriction, and simplest to work with, is a *geometric* distributed lag (GDL), which imposes geometric decay on the parameters in (2) for some $0 \leq \lambda \leq 1$:

$$\beta_s = \lambda^s \beta_0, \gamma_s = \lambda^s \gamma_0, \quad s=1, \dots, T \quad (4)$$

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 in the past increases. With these restrictions, after subtracting $\lambda A_{i,t-1}$ from both sides of (2) and performing substitutions and simple algebra, we obtain a much simpler estimating equation:

$$A_{it} = \tau_i + \lambda A_{i,t-1} + E_{it} \beta_0 + X_{it} \gamma_0 + \pi_i c_i + e_{it} \quad (5)$$

where

$$e_{it} = u_{it} - \lambda u_{i,t-1} \quad (6)$$

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 decay 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 & Alexander, 1992; Downey, Hippel & Broh 2004). Although allowing rates of decay to vary by individuals or groups is possible in (5), this is rarely, if ever, done in the literature on teacher effects.

In discussing estimators derived from (5), as we do in the next section, we will need to consider exogeneity of the inputs. This includes possible correlation with the unobserved time-invariant individual-specific component 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

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

and likewise for $\text{Cov}(X_{it}, e_{it})$. 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 is unlikely). In this case, the weakest exogeneity condition is that E_{it} is uncorrelated with $u_{it} - \lambda u_{i,t-1}$. This assumption could be true even if we do not assume E_{it} is uncorrelated separately with $u_{i,t-1}$ 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 $\text{Cov}(E_{it}, e_{it}) = 0$. A straightforward way to state the strict exogeneity assumption is

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

The difference between assumptions (8) and (3) is that (8) includes the entire history of observed inputs, including *future* inputs (this is why the t in (3) is replaced with T in (8)).

In addition to possible correlation between the covariates and e_{it} , we must worry about correlation with c_i . If c_i is present in (5)—as is likely—it is virtually impossible for c_i to be uncorrelated with $A_{i,t-1}$. In addition, 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 represents complete decay. In this case, (5) reduces to what is often referred to as a “level-score” equation. Taken 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 (see, for example, Dee, 2004). In the case of estimating teacher value added, for example, 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 very widely used formulation sets $\lambda = 1$ (no decay) and subtracts $A_{i,t-1}$ from both sides of (5), thereby achieving a so-called “gain score” formulation:

$$\Delta A_{it} = \tau_i + E_{it}\beta_0 + X_{it}\gamma_0 + \pi_i c_i + e_{it} \tag{9}$$

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 discusses five commonly used estimation methods and the assumptions underlying their use. A summary of these assumptions is found in the appendix. One caveat to apply to our

discussion of underlying assumptions is that we appeal 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, then 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. Pooled OLS

We begin our discussion of estimators based on equation (9), where the gain score, ΔA_{it} , is used as the dependent variable and contemporaneous inputs are the explanatory variables. Not surprisingly, when λ is actually 1, the gain score approach has several advantages. If we can ignore the presence of c_i or successfully introduce proxies for it, pooled OLS (POLS) is a natural estimation method. Unfortunately, POLS estimation in (9) is generally inconsistent if the heterogeneity is correlated with E_{it} , which will be the case if students are assigned to educational inputs based on time-constant unobservables. Controlling for a rich set of family background variables can mitigate the problem, but proxies for c_i are hard to come by, and those easily available (for example, gender or race) are likely insufficient to proxy motivation or persistent correlates of ability.

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 (7): 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 then POLS is consistent provided the inputs are uncorrelated also with the unobserved heterogeneity. Inference for pooled OLS that allows arbitrary serial correlation and heteroskedasticity in the composite error $\pi_t c_i + e_{it}$ is straightforward.

3.2. Random Effects

A drawback to POLS – again assuming for the moment that $\lambda = 1$, the inputs are strictly exogenous, and the inputs are uncorrelated with student heterogeneity – is that it is generally inefficient because it ignores, in estimating β_0 , the serial correlation and heteroskedasticity in the composite error, $\pi_t c_i + e_{it}$. If we assume π_t is constant and assume that $\{e_{it}\}$ is serially uncorrelated and homoskedastic in equation (9), then random effects (RE) estimation can be used to improve over POLS. Like POLS, RE assumes the heterogeneity is uncorrelated with inputs. Thus, like POLS, consistency of RE relies on some strong assumptions, and it is guaranteed to be the efficient generalized least squares estimator only when $\{e_{it}\}$ satisfies ideal assumptions.

An attractive feature of RE estimation is that, when POLS and RE are both consistent, RE can improve upon POLS in terms of efficiency even if $\{e_{it}\}$ is serially correlated or contains heteroskedasticity.⁴ 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 . 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 in this paper, however, as it is rare in applications.

3.3. Fixed Effects

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 fixed effects (FE) (at least when π_t is constant). The FE estimator also requires a form of strict exogeneity of E_{it} and X_{it} because FE employs a time-demeaning transformation that requires that the e_{it} are uncorrelated with the time-demeaned inputs. 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, fixed effects is attractive. POLS and RE will suffer from systematic bias. Except in experimental studies, FE tends to be

⁴ Efficiency gains using RE in such settings is not guaranteed, but it is often more efficient than POLS because it accounts for serial correlation to some extent, even if not perfectly. This is the motivation behind the generalized estimating equations literature (see, for example, Zeger, Liang, & Albert 1988 or Wooldridge 2010, Chapter 10).

viewed as more reliable. 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.

3.4. Dynamic Ordinary Least Squares

If we do not wish to impose $\lambda = 0$ or $\lambda = 1$ (or λ some other known value), we must estimate the parameters in (5) – that is, those in the underlying CEM – using methods other than those based on the gain-score equation. If we write equation (5) with a composite error v_{it} , as

$$A_{it} = \tau_i + \lambda A_{i,t-1} + E_{it}\beta_0 + X_{it}\gamma_0 + v_{it}, \quad (10)$$

and ignore the properties of v_{it} – that it depends on $\pi_i c_i$ and (the possibly serially correlated) e_{it} – then we might take a seemingly naïve approach and simply estimate a dynamic regression. In other words, we might estimate λ , β_0 , and γ_0 using a pooled OLS regression. Of course, subtracting $A_{i,t-1}$ from both sides only changes the coefficient on $A_{i,t-1}$, so we can think of this as a gain-score regression but with the inclusion of the lagged achievement score. To distinguish this estimator from the static POLS gain-score regression, we will refer to it as “dynamic pooled ordinary least squares” (DOLS).

Consistency of the DOLS estimator for β_0 , and γ_0 (and λ) hinges on strict exogeneity of the inputs (with respect to $\{u_{it}\}$) and no serial correlation in $\{e_{it}\}$. In addition, the presence of $\pi_i c_i$ generally causes inconsistency because c_i is correlated with $A_{i,t-1}$ (and possibly the inputs, too). Nevertheless, it is possible that DOLS provides better estimates of β_0 than other methods in various scenarios. For example, if the $\pi_i c_i$ are sufficiently “small,” ignoring this component of the composite error term v_{it} might not be costly. Further, including $A_{i,t-1}$ means we are not assuming a known value of λ . Perhaps most importantly, controlling for $A_{i,t-1}$ explicitly allows for the kinds of dynamic assignment of students to inputs based on prior test scores.

3.5. First Differencing with Instrumental Variables

Rather than ignore the heterogeneity c_i as well as the possibility that the inputs fail strict exogeneity, a combination of first differencing and instrumental variables can be used to estimate all of the parameters (except the time-varying intercepts). To account for unobserved heterogeneity, again assuming that π_t is a constant, we can eliminate c_i by first differencing (5) to obtain:

$$\Delta A_{it} = \chi_t + \lambda \Delta A_{i,t-1} + \Delta E_{it} \beta_0 + \Delta X_{it} \gamma_0 + \Delta e_{it} \quad (11)$$

Generally, this differenced equation cannot be 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 natural 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, and we do not consider it in our simulations. An attractive feature of estimating (11) in this manner, however, is that it does not restrict serial correlation in the original errors, $\{u_{it}\}$, whereas the more popular instrumental variables estimator discussed next does impose restrictions.

3.6. The Arellano and Bond Approach

A second and much more commonly applied strategy is to choose instruments for the lagged gain score from the available achievement lags. One can then use the estimator outlined in Arellano and Bond (1991) (AB) or simpler IV versions using the same kinds of moment restrictions. (A simpler version of the estimator, which is transparent and allows one to easily study the first-stage regressions, is a system

two stage least squares (2SLS) estimator, which is the same as the AB estimator with an identity weighting matrix.) However, the AB approach requires that there be no serial correlation in the $\{e_{it}\}$.

To claim that the $\{e_{it}\}$ are serially uncorrelated, one must restrict the serial correlation in the original errors by asserting that they follow an $AR(1)$ process, namely $u_{it} = \rho u_{i,t-1} + r_{it}$ where $\{r_{it}\}$ is serially uncorrelated. In addition, 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 $e_{it} = u_{it} - \lambda u_{i,t-1}$ in (5) are the same as the serially uncorrelated r_{it} . If we add strict exogeneity of the inputs E_{it} and X_{it} as given by (8), we can assume that e_{it} is unpredictable given past achievement and the entire history of inputs:

$$E(e_{it} | A_{i,t-1}, A_{i,t-2}, \dots, A_{i0}, E_{iT}, E_{i,t-1}, \dots, E_{i0}, X_{iT}, X_{i,t-1}, \dots, X_{i0}, c_i) = 0 \quad (12)$$

The usefulness of assumption (12) is that it implies that $\{A_{i,t-2}, \dots, A_{i0}\}$ are uncorrelated with e_{it} and so these are instrumental variable candidates for $\Delta A_{i,t-1}$ in (11). Typically, $\{A_{i,t-2}, \dots, A_{i0}\}$ 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 (11) 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 “close” to 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 may not 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 (11), Δe_{it} depends on

$\{u_{it}, u_{i,t-1}, u_{i,t-2}\}$ and so, if we hope to relax strict exogeneity of the inputs in (2), we must choose our IVs from $\{A_{i,t-2}, \dots, A_{i0}, E_{i,t-2}, \dots, E_{i0}, X_{i,t-2}, \dots, X_{i0}\}$. This approach imposes substantial data requirements.

The fact that IV estimation of (11) relies on the CF restriction raises a subtle point when comparing (11) and methods based on the gain-score equation (9). At first glance it appears that (11) is more general because it does not impose $\lambda = 1$. But for the common estimation methods to be consistent, the CF restriction on the errors generally needs to hold. An important implication is that it that estimating λ when it is unity can be costly when using the first-differenced equation (11). In particular, if $\lambda = 1$ and the inputs are strictly exogenous, FE estimation of (9) consistently estimates the teacher VAMs without the CF restriction whereas IV estimation of (11) is generally inconsistent for the parameters in the CEM if the CF restriction fails. It is important to remember that when the goal is to estimate the parameters of the underlying CEM, the AB approach, in addition to imposing geometric decay on the effects of the inputs, relies on the CF restriction. Analyses that adopt the AB approach tend to ignore the restriction $\rho = \lambda$ (and the failure of the approach when applied to (5) with general serial correlation in the structural shocks $\{u_{it}\}$). Because of this, we must be careful not to claim superiority of the AB approach over methods that do not require the CF restriction.

3.7. Summary of Estimation Approaches

In summary, estimation of the parameters of the cumulative effects model, even when we impose the geometric distributed lag restriction to arrive at equation (5), requires numerous additional assumptions. POLS and RE on the gain score equation require strict exogeneity of inputs E_{it} and X_{it} and no correlation with c_i . FE allows for correlation between c_i and inputs E_{it} and X_{it} but maintains strict exogeneity. For either RE or FE to be an appropriate estimation method, however, λ must equal 1 (or a different known value). The benefit of FE estimation is that it allows input assignment to be correlated with c_i and does not restrict the serial correlation of the errors in the CEM. Pooled OLS estimation of the dynamic equation – what we have called DOLS – requires strict exogeneity of inputs E_{it} and X_{it} and effectively imposes the common factor restriction on an AR(1) model for $\{u_{it}\}$. In addition, the method is

not generally consistent if c_i is in the equation. The AB approach generally requires $\lambda < 1$, no serial correlation in the $\{e_{it}\}$, and strict exogeneity of the inputs E_{it} and X_{it} . Compared with DOLS, it explicitly recognizes the presence of c_i , which is why differencing is used followed by IV estimation. However, this does not guarantee that the AB approach dominates DOLS across various data generating mechanisms. If, say, the inputs are not strictly exogenous, both estimators are technically inconsistent; the interesting issue, as with the other estimation approaches, is which method does a better job recovering the coefficients on the inputs.

4. Situating Theory in Context

Until now, we have discussed the assumptions about value-added models upon which various estimators rely in relatively abstract terms. In this section, we discuss different types of educational scenarios and whether they might be expected to violate exogeneity assumptions, particularly when the goal is to estimate the effectiveness of individual teachers.

If schools engage in grouping students on the basis of their perceived ability—a practice commonly known as “tracking”—then random assignment to teachers within schools could potentially be compromised. Tracking can take a number of forms. Students may be grouped together on the basis of either their prior test score, $A_{i,t-1}$,⁵ their level of achievement or ability upon entering school, A_{i0} , or their potential for learning gains, c_i . The first type of tracking, a relatively common practice in educational settings, might be called “dynamic tracking.” The second and third types of tracking, both forms of “static tracking,” are less common. They might occur when, for example, schools either formally or informally assess the level of learning or the growth potential of children upon entering school, group the children accordingly, then keep the more or less the same groups of children together for several grades. The practice of keeping the same group of children together as they proceed through school—called, in some cases, “looping”—is occasionally used in the elementary grades.

⁵ Here for simplicity we refer to just one prior test score. However, principals might average over a series of prior test scores.

It is important to recognize that tracking does not, in and of itself, induce correlation between unobserved factors affecting student performance and teacher effects. 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 students to teachers.

Assignment of classrooms to teachers can take 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.

[Insert Table 1 Here]

It is important to recognize that a mixture of these grouping and assignment methods can be used in any given district or even in a given school. However, for the purpose of understanding and evaluating the performance of various estimators, we keep the scenarios distinct when we conduct our simulations.

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 classroom of 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—that is, when students are grouped in classrooms on the basis of prior test scores—coupled with nonrandom assignment

of classrooms to teachers—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 . As it turns out, dynamic grouping with nonrandom assignment need not cause a violation of sequential exogeneity when the lagged score is controlled for in the equation, as in DOLS. But DOLS in this scenario can still violate heterogeneity exogeneity if heterogeneity exists.

The two cases of static grouping need to be carefully considered, as they differ in important ways. For example, suppose students are grouped on a baseline score and then assigned to teachers nonrandomly. 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, if the teacher assignments in E_{it} are 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 essentially 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. Thus, assignment is correlated with a variable that should not be omitted from the equation, generally causing bias in any of the usual estimators applied to (9).

It is more obvious that static assignment based on c_i causes problems for estimating equations such as (9) unless $\pi_i c_i$ is removed from the equation. When π_i 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 (10). In this equation, both $A_{i,t-1}$ and E_{it} are correlated with c_i (which, remember, is part of the

composite error v_{it}). Our simulations help to identify the bias in estimating β_0 when this form of heterogeneity is ignored.

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 simulations scenarios, we therefore consider three distinct “school sorting” cases. In Case 1, 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 Case 2, 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. In Case 3, 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. One issue we shed light on is the interaction between nonrandom assignment of teachers to schools and nonrandom assignment of students to teachers.

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. Thus we consider the importance of how VAM-based measures of teacher effectiveness might 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. If, on the other hand, districts wish to know how effective particular teachers are compared with, say, the average, then the estimated teacher effects themselves are of primary importance, possibly standardized by a measure of dispersion such as the standard deviation. Our study investigates the performance of various estimators with respect to all three criteria. Our summary measures are described in detail in the methods section

5. Methods

Our empirical investigations consist of a series of simulations to evaluate the quality of various VAM estimation approaches. We use artificially generated data to investigate how well different estimators recover true effects under different scenarios. These scenarios correspond to data generating processes (DGPs) that vary the mechanisms used to assign students to teachers. To data generated from each DGP, we apply the set of estimators discussed in Section 3. We then compare the resulting estimates 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, 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 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 that the CF restriction holds (i.e., $\lambda=\rho$). Finally, there are no time effects embedded in our DGPs.

Our data are constructed to represent three upper elementary grades 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:

$$\begin{aligned}
 A_{i3} &= \lambda A_{i2} + \beta_{i3} + c_i + e_{i3} \\
 A_{i4} &= \lambda A_{i3} + \beta_{i4} + c_i + e_{i4} \\
 A_{i5} &= \lambda A_{i4} + \beta_{i5} + c_i + e_{i5}
 \end{aligned}
 \tag{13}$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of children entering third grade, λ is a time constant decay 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 that A_{i3} depends on A_{i2} using the same decay λ , it makes sense to think of A_{i2} as a second-grade test score or a pre-test score.) 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 uncorrelated with the baseline test score A_{i2} .

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 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 is 1, each representing approximately 5, 19, and 76 percent of the

total variance in gain scores, respectively, In the second, 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.

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 (thus 120 teacher overall)
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools

To create different scenarios, we vary certain key features: the sorting of students and teachers into schools, 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. Within each of the three school-sorting cases outlined in the previous section, we study the 10 different mechanisms for the assignment of students also outlined in that section. Note that we introduce a small amount of noise into each grouping process. Finally, we vary the decay parameter λ as follows: (1) $\lambda = 1$ (no decay or complete persistence) and (2) $\lambda = .5$ (fairly strong decay). The DGPs chosen for each simulation reproduce scenarios in which the assumptions mentioned above either hold or are violated. Thus, we explore $3 \times 10 \times 2 = 60$ different scenarios in this paper. We use 100 Monte Carlo replications per scenario in evaluating each estimator.

5.2. Methods for Estimating Teacher Effects

Section 3 described common approaches to estimating teacher value added. Here we provide a brief summary of the estimators that we study in the simulations. It is helpful to reproduce three modified versions of the estimating equations from Section 3. These 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 $\pi_t = 1$:

$$\Delta A_{it} = E_{it}\beta_0 + c_i + e_{it} \quad (14)$$

$$A_{it} = \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + e_{it} \quad (15)$$

$$\Delta A_{it} = \lambda \Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta e_{it} \quad (16)$$

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).

In real-life applications, there is the potentially important issue of what to use for the test score. A common practice is to standardize test scores by subtracting off within-grade means and dividing the result by within-grade standard deviations – producing a measure in which the sample average is zero and sample standard deviation is one. A priori, one might question whether standardizing test scores is harmless or whether it introduces distortions in the teacher effects estimates. Standardization can be important when different scales are used in different grades and one wishes to compare teachers not just within grade but also across grades. In our simulations, however, the test scores are vertically scaled and thus designed to allow for comparisons across grades. Because the standardization of test scores is a popular starting point for many value-added studies and teacher evaluation calculations,⁶ we therefore study whether standardizing test scores helps in estimating the teacher effects or causes distortions by reporting and discussing simulation results based on both actual and standardized test scores. For each of

⁶ See, for example, Buddin (2010), who uses standardized test scores to compute value-added performance measures for teachers in Los Angeles.

the 100 iterations pertaining to one DGP, we estimate effects for each teacher using one of five estimation methods: POLS, RE, and FE applied to (14), POLS applied to (15) (which we have called DOLS), and Arellano and Bond (AB) applied to (16) using both the unstandardized vertically scaled scores and standardized scores. We use the statistical software Stata for all data generation and estimation.

5.3. 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 of $\hat{\rho}$ across the 100 iterations.

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. 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. In addition to examining rank correlations and misclassification rates, it is also helpful to have a measure that quantifies some notion of the magnitude of bias in the estimates. Given that many teacher effects are estimated simultaneously, some high and some low, it is difficult to capture bias in a simple way. Our approach is to create a statistic (which we call $\hat{\theta}$) that captures how well the size of the deviation of the estimates from their mean tracks the size of the deviation of the true effects from the true mean. To create this measure, we regress the estimated 119 teacher effects (where one teacher is the base case) from each of the five estimators on the true effects generated from the simulation as follows:

$$\hat{\beta}_j = \alpha + \theta\beta_j + \varepsilon_j \tag{17}$$

where $\hat{\beta}_j$ is the estimated effect of teacher j (obtained using a particular estimation approach) and β_j is the true effect of teacher j . From this simple regression, we report the average coefficient $\bar{\hat{\theta}}$ and its standard deviation across the 500 simulations. Regressing the estimated teacher effects on the true teacher effects tells us whether the estimated teacher effects are correct when compared with the average teacher. To see this, recall that we can write a simple regression equation as

$$\hat{\beta}_j - \bar{\hat{\beta}} = \hat{\theta}(\beta_j - \bar{\beta}) + \text{residual}_j \quad (18)$$

where the overbars represent averages (across the 119 teachers), and $\hat{\theta}$ is the simple regression coefficient from (17). If $\hat{\theta} = 1$ then a movement of β_j away from its mean is tracked by the same movement of the estimate away from *its* mean.

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

Bias in the estimated teacher effects is not the most relevant issue for how teacher VAMs are used in practice, however. In the vast majority of cases, rankings are used. Nevertheless, along with the Spearman correlation we report the average value of the $\bar{\hat{\theta}}$ across the simulations. Doing so allows us to determine which methods, under which scenarios, produce estimated teacher effects whose magnitudes have meaning. Of equal importance, $\bar{\hat{\theta}}$ provides insight into why some methods are successful at ranking the teachers even though the estimated teacher effects are systematically biased.

6. Simulation Results

6.1. Case 1 (Random Sorting of Students and Teachers across Schools) and $\lambda = 1$ (No Decay) with Small Teacher Effects

We first discuss the findings for Case 1 – the case in which students and teachers are randomly sorted into schools – and $\lambda = 1$; these are given 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—a percentage frequently reported in literature. Each cell in Table 2 contains three numbers specific to the particular estimator-scenario combination. The first is the average rank correlation between the estimated and true teacher effects. The second is the average proportion of above average teachers who are misclassified as being below average. And the third is the average value of $\hat{\theta}$ from regression (17). The remaining tables have the same structure.

We expect all estimators to work well in the sense of little or no evidence of bias⁷ 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, POLS, and RE yield rank correlations above .8, with RE producing a rank correlation of about .885. FE and AB have rank correlations under .7. The FE and AB estimators are not systematically biased, but they yield notably lower correlations. The DOLS, POLS, and RE estimators are also better at classifying the teachers than the other two methods. More precisely, RE incorrectly classifies an above average teacher as being below average about 14% of the time; the misclassification rates for POLS and DOLS are similar. The misclassification rate of FE, on the other hand, is quite large at 50%. Clearly, the estimation error in the teacher effects using FE 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

⁷ Technically, not all the estimators are unbiased. Some are only consistent as the number of students taught by each teacher grows.

proportion of times at 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 20 to 100. Part a of Figure 1 shows the superiority of DOLS and RE over FE in the RG-RA scenario with λ equal to one. However, it should be noted that even for these estimators under these idealized conditions, identification of the “worst” teachers can be subject to a considerable amount of error.

As expected, the average values of $\hat{\theta}$ are all very close to one for all estimators in the RG-RA scenario, with AB very slightly worse than the others. But as the findings for the rank correlations and misclassification rates suggest, precision in the estimates plays an important role. These findings indicate that RE is the preferred estimation method under RG-RA with no decay, something that econometric theory leads us to expect because RE is the (asymptotically) efficient estimation method. However, POLS produces very similar results, and DOLS is fairly similar, as well, despite the fact that it is technically inconsistent because of the presence of the unobserved student effect and its correlation with lagged achievement.

Using standardized test scores in the RG-RA setting has minor effects on the ranking and classification statistics in this baseline scenario (see results in the first row of the left side of Table 3). However, because the teacher effects are effectively in gain score units – not changes in standardized test scores – the average of $\hat{\theta}$ is well below unity when the standardized scores are used.

Nonrandom grouping mechanisms for students have relatively minor consequences for RE, DOLS, and POLS when the actual test scores are used 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) – although heterogeneity grouping results in lower rank correlations for all estimators, ranging from .528 for AB to .749 for RE. (Note that all the random assignment scenarios are shown in shaded cells in the tables.) RE, POLS, and DOLS continue to yield relatively high correlations under dynamic and baseline grouping, ranging from .779 to .889. The methods that remove the student effect, FE and AB, do a much worse job in ranking and also classifying teachers. In addition,

for dynamic grouping, the FE and especially AB estimators appear to have some systematic bias, as is evidenced by $\bar{\theta}$ less than one.

Interestingly, using standardized test scores to estimate teacher effects results in worse performance for DOLS, POLS, and RE in the dynamic and baseline grouping cases (Table 3). For example, for DG-RA, the rank correlation for DOLS is .829 with misclassification rate .18 using the original scores, but these numbers change to .737 and .22 using the standardized scores. However, the deterioration is worse for POLS and RE for the BG-RA scenario. Thus, even when teachers are randomly assigned to classrooms within schools, the use of standardized test scores can lead to notably worse performance of several of the estimators.

Generally, nonrandom grouping of students causes all estimation methods 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. For example, in the HG-RA scenario, the rank correlation for RE falls to .749 from .885 in the RG-RA case. Nevertheless, the overall picture is fairly promising for the three estimators – especially for random effects – when teachers are randomly assigned to classrooms, $\lambda = 1$, and the original, vertically scaled test scores are used

When teachers are *nonrandomly assigned* to classrooms, however, the properties of the estimation procedures change markedly – and it depends critically on the nature of the nonrandom assignment. When dynamic assignment is used and better students are assigned to the better teachers (i.e., the DG-PA scenario), the estimators that remove the student-level heterogeneity – FE and AB – perform especially poorly. In particular, the FE estimator produces a negative rank correlation between the estimated and true teacher effects (–.34) and misclassifies 56% of the above-average teachers as below average. The poor performance of FE is highlighted in Figure 1, part b, which vividly illustrates how the worst teachers have a lower probability of being classified as underperforming than the best ones. AB gives a positive rank correlation between the estimated and true teacher effects, but it is low, indicating

that this procedure will not be very helpful in distinguishing among teachers. AB misclassifies more than 40% of the teachers.

Interestingly, POLS and RE – which both leave c_i in the error term – do well in ranking teachers in the DG-PA scenario (where FE and AB fare so poorly). In fact, the rank correlations for FE and RE are right around .90, even somewhat higher than DOLS, which is the indicated estimation approach for this scenario because it directly controls for the assignment mechanism. In addition, the misclassification rates are small: .13 for both POLS and RE. Interestingly, the average $\hat{\theta}$ in equation (20) is well above unity for POLS and RE. In other words, relative to the true teacher effects, POLS and RE amplify the estimated teacher effects: good teachers are estimated to be even better and below-average teachers are estimated to be even worse. Of course, this means that the magnitudes of the POLS and RE teacher effects cannot be used to determine how much better (as measured by average test scores) one teacher is compared with another. By contrast, for DOLS the average $\hat{\theta}$ is one to three decimal places, so that it would be valid to compare the magnitudes of the DOLS estimated of teacher effects across teachers. Furthermore, DOLS does a good job ranking the teachers, too (rank correlation = .841) and the misclassification is .18, slightly above that for POLS and RE.

The impact of using standardized test scores is startling when teachers are nonrandomly assigned to classrooms (see Table 3, left panel). Still looking at the DG-PA scenario, POLS and RE go from doing very well to performing abysmally. (FE goes from abysmal to something worse than that.) The rank correlation for both RE and POLS is actually negative, $-.271$, and the misclassification rate is approximately 60%. **Figure 2 part b shows how** RE joins FE in producing perverse classification outcomes. This is a potentially important finding because standardization is so common in empirical work.⁸ By contrast, the performance of DOLS is essentially unaffected by the measurement of the test score. The stability of DOLS almost surely derives from its estimating a coefficient on the lagged test score, which can adjust when the units of the dependent and lagged dependent variable change. (Recall

⁸ It appears in this situation that POLS and RE are very similar – maybe even identical – across all simulations. This happens whenever the variance of c_i is estimated to be negative, which may be caused if using the wrong test scores effectively introduces a negative serial correlation in the idiosyncratic errors.

that the POLS and RE estimators set the coefficient on the lag to one, which is the correct restriction only when the original test scores are used.)

When students are grouped on the basis of their prior test scores and the better teachers are assigned to the worse students (i.e., the DG-NA scenario), all estimators perform worse than DOLS for ranking the teachers when the original test scores are used (with AB producing a rank correlation less than .2). FE actually produces a fairly high rank correlation in this case, but this is because the estimated effects are amplified relative to the true effects as evidenced by a $\bar{\hat{\theta}}$ of 2.04, which indicates that the magnitudes of the effects are, on average, twice as large as they should be. In fact, FE does better than both POLS and RE, whose estimates are compressed relative to the true teacher effect.

Using the standardized test scores actually makes POLS and RE look considerably better in the DG-NA scenario. For example, for POLS the rank correlation is .637 using the unstandardized scores with a misclassification rate of .28. With the standardized scores, these change to .89 and .12, respectively. This appears to be a case of two wrongs making right: the dynamic assignment along with using the wrong test scores interact to make POLS perform relatively well for ranking and classification. This hardly seems like a sensible way to justify VAMs for use in policy analysis, however, as we will never know what combinations of all possible misspecifications will, by luck, lead to good performance.

So far, even though we have discussed only the case of no decay, we can summarize some useful insights. First, the choice to standardize test scores is hardly innocuous. Doing so can lead to very poor performance in situations where using the unstandardized test scores would produce acceptable results. Standardizing can help in some specific scenarios due to an artifact of the estimation process, but, at this point, it is hard to see how one could know that without knowing a lot about the student grouping and teacher assignment mechanisms. An important practical point is that DOLS appears to be the least sensitive, by far, to using the standardized or unstandardized scores.

We can understand the relatively good performance of DOLS under the various dynamic grouping scenarios by noting that if the DGP did not include a student effect the DOLS estimator would

be robust to the kind of dynamic assignment created by this scenario. Namely, the teacher dummies E_{it} are correlated with $A_{i,t-1}$ but the latter is controlled for in the DOLS regression. POLS, RE, and FE do not control for the lagged achievement score because they use the gain score as the dependent variable and omit the lag. POLS and RE – which both leave c_i in the error term – suffer from an omitted variable problem because assignment is based on the lagged test score and the lagged score is positively correlated with c_i . In the DG-PA case, the resulting bias in estimating the teacher effects by POLS or RE actually helps with ranking the students, but it hurts in the DG-NA case. DOLS, on the other hand, exhibits the same behavior whether assignment is RA, PA, or NA.

Interestingly, DOLS does well even for estimating $\beta_j - \bar{\beta}$ even though, technically, it is inconsistent due to the presence of the student effect. Evidently, the size of the student effect in our DGPs – despite the fact that it accounts for 4/21 of the total variance in the gain score – does not translate into much bias in estimating the teacher effects. This is likely due to the fact that putting in the lagged achievement has two positive effects: it absorbs most of the unobserved student effect while also making the dynamic assignment exogenous in equation (15). The average DOLS estimate of λ in the simulations is about 1.088 (not shown in the table), showing a slight upward bias probably due to the neglected heterogeneity

Still using the $\lambda = 1$ data generating mechanism with random sorting of students and teachers into schools, we also study the effects of nonrandom teacher assignment with static grouping of students. In one scenario students are grouped at the outset according to their baseline test scores. The better student groups are assigned to the better teachers in one case (BG-PA), and in the other case better students are assigned to worse teachers (BG-NA). Provided we use the original scores in constructing the gain score, we know at the outset that POLS, RE, and FE are all consistent for estimating the teacher effects because the variable determining teacher assignment (the second-grade test score, A_{i2}) does not appear in equation (14). In other words, assignment depends on a variable that has no direct effect on the dependent variable in (14) – at least when $\lambda = 1$. In fact, RE is still the efficient estimation method

because c_i is independent of all teacher assignments. If, however, c_i were correlated with the base score—arguably a more plausible situation than the independence assumption we simulate here for didactic purposes—then POLS and RE would suffer from omitted variable bias.

As we see from Table 2, POLS is similar to RE in these two scenarios. As predicted, the behavior of POLS, RE, and FE is very similar across the BG-PA and BG-NA (as well as BG-RA) scenarios, with FE performing worse because it unnecessarily removes c_i . AB performs even slightly worse than FE, presumably because, in addition to removing c_i , it estimates λ rather than setting it to the correct value of one.

The DOLS estimator works slightly worse than POLS and RE when students are grouped with respect to the base test score, particularly under “positive assignment.” It turns out that DOLS is systematically biased because it effectively controls for the wrong explanatory variable, $A_{i,t-1}$, when it is the base score, A_{i2} , that should be controlled for. This can be seen, with $\lambda = 1$, by writing A_{it} as a function of all past inputs, shocks, and the initial value. 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 A_{i2} . An extension of the DOLS estimator used here would be to add further lags of the test score, although doing so reduces the number of grades available for estimation.

Under BG-NA, the DOLS estimator amplifies the true teacher effects about their mean, and this leads to good ranking properties (though slightly worse than POLS and RE). Under BG-PA, the DOLS estimates are compressed, making DOLS substantially worse for ranking. As expected, POLS, RE, and FE produce $\bar{\theta}$ all near one (and AB, too, although it is slightly lower at .95).

Unfortunately, as in the dynamic assignment case, the decision to standardize the test scores can again prove to be very costly, depending on the estimation method used. In the BG-PA case, POLS and RE go from working very well when the unstandardized scores are used to not working at all. The rank correlation between the estimated and actual teacher effects is about zero (.013) and 50% of the above

average teachers are misclassified. On the other hand, as in the DG case, DOLS is essentially unaffected by the decision to standardize or not.

The other 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. That is, c_i is added, in each period, to the previous score. When the students with the highest growth potential are grouped with the best teachers (HG-PA), the bias in POLS and RE – both lead to estimated teacher effects that are severely amplified relative to the true effects as evidenced by $\bar{\hat{\theta}}$ – lead them to rank and classify the teachers very well. But negative assignment causes them to do much worse. In fact, in the HG-NA scenario no estimator does very well – the highest rank correlation is .626 (FE) and the lowest misclassification rate is .25 (FE). **Figure 1.c illustrates** the decline in performance of RE and DOLS relative to the scenario depicted in part a. The FE estimator, as expected from the theory, delivers $\bar{\hat{\theta}}$ close to one in all cases but its imprecision makes it only slightly better than RE for ranking and classifying teachers. The AB estimator behaves somewhat worse than the FE estimator.

Theoretically, the FE estimator is the most efficient estimator among those that place no restrictions on the relationship between c_i and the teacher dummies E_{it} . But the consistency of FE (and AB) for the deviated teacher effects is of small comfort, as they do not outperform the estimators that effectively treat c_i and the teacher dummies E_{it} as being uncorrelated along the dimensions that count most: ranking and classifying. The DOLS estimator has properties similar to POLS and RE although it performs a little worse than RE in the HG-NA scenario.

The previous findings show that when students and teachers are randomly sorted into schools and $\lambda = 1$ certain estimators perform very poorly under some of the assignment mechanisms – even some estimators that effectively use $\lambda = 1$ in estimation. Estimators that are intended to be robust to static assignment do poorly under dynamic assignment. One useful finding is that, looking across all assignment mechanisms and whether scores are standardized or not, DOLS does best: it is far superior for dynamic assignment and still has value for ranking teachers under static assignment.

6.2. Case 1 (Random Sorting of Students and Teachers across Schools) and $\lambda = .5$ (Strong Decay) with Small Teacher Effects

The performance of most estimators deteriorates substantially 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 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, an omitted variable, $A_{i,t-1}$ in equation (14) will have a coefficient of $-.5$ on it; this is important to remember in interpreting the findings.

Dynamic misspecification has only minor effects with respect to bias when teachers are randomly assigned to classrooms, even if students are nonrandomly grouped. In scenarios RG-RA, DG-RA, BG-RA, and HG-RA, $\bar{\theta}$ hovers around one for the most part. With λ now better identified via differencing and instrumenting, AB does somewhat better in the DG-RA scenario than it did in when $\lambda = 1$.

Where the misspecified dynamics have the biggest effect in these situations is on the precision of the estimates. This comes through when looking at the rank correlations and misclassification rates. Compared with the $\lambda = 1$ DGP, the rank correlations for POLS and RE are substantially worse when $\lambda = .5$ (Because of the negative correlation induced in the errors by incorrectly setting $\lambda = 1$, POLS and RE turn out to be identical across all of these simulations.) For example, even in the RG-RA scenario, the rank correlation is only .55, down from around .88. The misclassification rate is .36 compared with .15 when $\lambda = 1$. For some reason, using standardized test scores helps in the RG-RA situation (but not in others; see below). The impact on RE for misclassifying low-performing teachers is seen in [Figure 1.d](#), and the boost it gets from standardization is visible in [Figure 2.d](#).

When coupled with dynamic assignment, dynamic misspecification – see Table 2 under DG-PA and DG-NA – has very serious consequences for all estimators with the notable exception of DOLS. In fact, when the best students are matched with the best teachers, RE actually produces a negative rank

correlation and has a 56% misclassification rate. The striking effects for misclassification at the tails of the quality distribution are visible in [Figures 1.e and 2.e](#). Using standardized test scores makes things even worse – except for DOLS. In fact, DOLS is so superior to the other estimators that it seems clear that, among the group of estimators considered in this study, DOLS is easily preferred under dynamic grouping: looking across the different assignment mechanisms and both values of λ , no estimator is even a close second.

Dynamic misspecification also has consequences in the case of static assignment. RE (which is the same as POLS for these DGPs, as discussed earlier) does a very bad job of ranking and classifying students in the BG-PA case. The rank correlation is only .165. (Standardizing the test scores makes things far worse.) FE and AB do better in this case, but both are clearly inferior to DOLS. With negative assignment, RE does substantially better, but it is always inferior to DOLS. Yet again, DOLS is stable whether the original or standardized test scores are used.

Even when grouping is based on heterogeneity, DOLS now does better than all estimators compared with the $\lambda = 1$ case – at least when the unstandardized scores are used. Standardizing helps RE in the HG-NA case, but, again, it is hard to use this finding in practice. DOLS does not do particularly well in the HG-NA setting but at least it produces a nontrivial rank correlation (.4). No estimator works very well in the HG-NA case, and none achieves a classification rate better than guessing.

The simulations for Case 1, when both students and teachers are randomly assigned to schools, point to several conclusions. While DOLS is not uniformly better across all of the grouping, assignment, and decay assumptions, it is nearly so. And where a different estimator is preferred to DOLS, DOLS is always a pretty close second. The performance of DOLS is stable across values of λ , and it is insensitive to whether unstandardized or standardized test scores are used. The other estimators show much more sensitivity to the value of λ and to whether the test scores are standardized. However, we should again note that the potential for misclassification under even DOLS can approach levels that might be considered unacceptable for policy purposes.

6.3. Large Teacher Effects

In this section, we briefly discuss the simulation results that correspond to those discussed above when teacher effects represent a much larger relative share of the total variance in student gains. These are reported in Table 4, as well as Figure 3. As to be expected, when the size of the teacher effects is raised relative to the student effect and errors, rank correlations improve and misclassification rates decline somewhat. However, the same overall patterns discussed above hold. The relative superiority of DOLS over POLS and RE in the case with strong decay is still evident although somewhat less pronounced when teacher effects are large. The FE and AB estimators improve their rank correlations in most 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. First, we looked at the impact of nonrandom sorting of students and teachers across schools. These different sorting scenarios did little to affect the general patterns described above, suggesting that the primary threat to the estimation of teacher effects stems from within-school assignment to teachers.

We also ran a full set of simulations with $\lambda = .75$, 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 is essentially between the $\lambda = 1$ and $\lambda = .5$ cases but it is somewhat surprising that the performance is much closer to the $\lambda = 1$ case. A general conclusion is that the performance of the POLS and RE estimator seems to deteriorate slowly for modest amounts of decay but then deteriorates quickly as the amount of decay grows. The DOLS estimator is hardly affected by the value of λ .

We also added classical measurement error to the test scores in our DGPs. In addition, we ran simulations in which serial correlation was introduced in the errors (i.e., relaxing the common factor restriction). While these complications served to depress rank correlations slightly and to increase misclassification rates, they did not affect the overall patterns described. However, they provided evidence of the fact that as we introduce more real-world complexity into our DGPs, the performance of the estimators will be expected to worsen.

7. Conclusions and Future Directions

Simulated data with known properties and parameters permits the systematic exploration of the ability of various estimation methods to recover true teacher effects (parameters used to generate the data). This study has taken the first step in evaluating different value-added estimation strategies under conditions where 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.

Our main finding is that no one method is guaranteed to accurately capture true teacher effects in all contexts even under these 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 contexts. Both the teacher assignment mechanism and the nature of the dynamic relationship between current and past achievement (i.e., λ) 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.

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. 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 somewhat correlated with prior achievement.

These findings suggest that choosing estimators on the basis of structural modeling considerations may produce inferior results. 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. 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 (e.g., 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 findings in this paper, though special, 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.

Another important result is that standardizing vertically scaled test scores prior to estimating effects generally produces inferior and potentially untrustworthy results. Here again, however, we find that DOLS is less sensitive to this issue than other estimators, reinforcing our finding that using methods that do not directly control for lagged achievement on the right-hand side are risky.

Finally, despite the relatively robust performance of DOLS, we find that even in the best scenarios and under the highly and unrealistically 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 substantial, 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 five to more than 50 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 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. In future research, we will explore the extent to which various estimation methods, including more sophisticated dynamic treatment effects estimators, can handle further complexity in the DGPs. The addition of test measurement error, school effects, time-varying teacher effects, and different types of interactions among teachers and students are a few of many possible dimensions of complexity that must be studied. Finally, diagnostics are needed to identify the structure of decay and prevailing teacher assignment mechanisms. If contextual norms with regard to grouping and assignment mechanisms can be deduced from available data, then it may be possible to determine which estimators should be applied in a given context. For this purpose, structural modeling considerations may be helpful in that they yield tests that have the potential to identify

violations of particular assumptions. Investigations in these areas are the subject of current research in progress by the authors.

Clearly, although value-added measures of teacher performance hold some promise, more research is needed before they can confidently be implemented in policies. Our findings suggest that sets of 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, 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. 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 recently been the case in Los Angeles and may soon be the case in 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 stakes to these measures until their properties have been better understood.

References

- Arellano, M. & Bond, S. (1991) Some Tests of Specification of Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58, pp.277-298.
- Blundell, R. & Bond, S. (1998) Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87, 11-143.
- Briggs, D. & Weeks, J. (2009) The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale, *Education Finance and Policy*, 4(4), 384-414.
- Downey, D., Hippel, P., & Broh, B. (2004) Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year, *American Sociological Review*, 69(5), 613-635.
- Entwistle, D. & Alexander, K. (1992) Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School, *American Sociological Review* Vol. 57, No. 1 (Feb., 1992), pp. 72-84
- Hanushek, E. "The Economics of Schooling: Production and Efficiency in the Public Schools," *Journal of Economic Literature*, XXIV (3): 1141-78, 1986.
- Hanushek, E. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14(3): 351-388, 1979.
- Harris, D. & Sass, T. (2006) Value-Added Models and the Measurement of Teacher Quality, Unpublished Draft.
- Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, Working Paper 14607, National Bureau of Economic Research.
- Lechner, M. (2008), Matching Estimation of Dynamic Treatment Models: Some Practical Issues, in *Advances in Econometrics*, Volume 21 (Modeling and Evaluating Treatment Effects in

Econometrics). Daniel Millimet, Jeffrey Smith, and Edward Vytlacil (eds.), 289-333-117.
Amsterdam: Elsevier, 2008

Martineau, J. (2006) Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based, Value-Added Accountability, *Journal of Educational and Behavioral Statistics*, 31(1), pp. 35-62.

McCaffrey, D., Lockwood, J.R., Louis, T., & Hamilton, L. (2004) Models for Value-Added Models of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), pp. 67-101.

Raudenbush, S. (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.

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement*, 9, 401–412.

Reckase, M. D. & Li, T. (2007). Estimating gain in achievement when content specifications change: a multidimensional item response theory approach. In R. W. Lissitz (Ed.) *Assessing and modeling cognitive development in school*. Maple Grove, MN: JAM Press.

Rothstein, J. (2008) Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *NBER Working Paper Series*, Working Paper 14442, <http://www.nber.org/papers/w14442>.

Sanders, W. & Horn, S. (1994) The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel in Education*, 8, 299-311.

Sanders, W., Saxton, A., & Horn, B. (1997) The Tennessee Value-Added Assessment System: A Quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is student Achievement a Valid Evaluational Measure?* Thousand Oaks, CA: Corwin Press, Inc., 137-162.

Todd, P. & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485), 3-33.

US Department of Education (2009) Race to the Top Program: Executive Summary,
<http://www2.ed.gov/programs/racetothetop/executive-summary.pdf>, accessed on 9/8/10.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2e. MIT Press:
Cambridge, MA.

Zeger, S., Liang, K., & Albert, P. (1988) Models for Longitudinal Data: A Generalized
Estimating Equation Approach. *Biometrics*. 44(4), 1049-1060.

Appendix: Structural Value-Added Models and Assumptions

General Cumulative Effects Model (GCEM)

$$A_{it} = f_t(Z_{it}, Z_{it-1}, \dots, Z_{i0}, c_i, u_{it})$$

Linear Cumulative Effects Model (LCEM), allowing possibility of serial correlation

$$A_{it} = \alpha_t + Z_{it}\beta_0 + Z_{it-1}\beta_1 + \dots + Z_{i0}\beta_t + \eta_t c_i + u_{it}$$

$$u_{it} = \rho u_{i,t-1} + v_{it}$$

Geometric Distributed Lag (GDL)*

$$A_{it} = \tau_t + \lambda A_{i,t-1} + Z_{it}\beta_0 + \pi_t c_i + e_{it}$$

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

$$Cov(Z_{it}, e_{it}) = -\lambda Cov(Z_{it}, u_{i,t-1})$$

Assumptions used in value-added

- 1) Linear in parameters
- 2) Non-varying function over time
- 3) Family inputs prior to t_0 are captured in c_i
- 4) Individual heterogeneity c_i
 - a) $\eta_t = \text{constant} \rightarrow \pi_t = \text{constant}$
 - b) c_i uncorrelated with Z
 - c) $h_t = 0 \rightarrow \pi_t = 0$
- 5) Exogeneity
 - a) Contemporaneous (error term uncorrelated with current Z)
 - b) Sequential (error term uncorrelated with current and past Z)
 - c) Strict (error term uncorrelated with past, current, and future Z)
- 6) GDL : $\beta_s = \lambda^s \beta_0, s=1, \dots, T$
 - a) $0 < \lambda < 1$
 - b) $\lambda = 1$
 - c) $\lambda = 0$
- 7) Serial correlation in u_{it}
 - a) $\rho \neq 0, \rho \neq 1, \rho \neq \lambda$
 - b) $\rho = \lambda$ (common factor restriction)
 - c) $\rho = 1$
 - d) $\rho = 0$

Common Estimators that Impose the GDL and Other Assumptions

Estimator	Assumptions	Minimum Data Requirements	Estimating Equation
DOLS ¹	1-3, 4c, 5c, 6a, 7b	Current Z , 1 lag of scores	$A_{it} = \tau_t + \lambda A_{i,t-1} + Z_{it}\beta_0 + e_{it}$
POLS ²	1-3, 4c, 5c, 6b, 7b	Current Z , 1 lag of scores	$\Delta A_{it} = \tau_t + Z_{it}\beta_0 + e_{it}$
RE ³	1-3, 4b, 5c, 6b, 7b	Current Z , 1 lag Z , 1 lag of scores	$\Delta A_{it} = \tau_t + Z_{it}\beta_0 + v_{it} \quad v_{it} = \pi_t c_i + e_{it}$
FE ⁴	1-3, 4a, 5c, 6b, 7b	Current Z , 1 lag Z , 1 lag of scores	$\Delta A_{it} = \tau_t + Z_{it}\beta_0 + \pi_t c_i + e_{it}$
AB ⁵	1-3, 4a, 5c, 6a, 7b	Current Z , 1 lag Z , 2 lags of scores	$\Delta A_{it} = \tau_t + \lambda \Delta A_{i,t-1} + \Delta Z_{it}\beta_0 + \Delta e_{it}$

¹ Pooled OLS regression of gain score on lagged score and inputs. Referred to as “dynamic OLS” (DOLS). ² Pooled OLS regression of gain score on inputs. ³ Random effects regression of gain score on

inputs.⁴ Fixed effects regression of gain score on inputs.⁵ First difference with instrumental variables regression of gain score on lagged score and inputs.

***Derivation of the GDL equation from the LCEM**

Say there are three periods per student (0,1,2). The level equation for, say, period 2 is:

$$A_{i2} = \alpha_2 + Z_{i2}\beta_0 + Z_{i1}\beta_1 + Z_{i0}\beta_2 + \eta_2c_i + u_{i2}$$

The lagged equation is:

$$A_{i1} = \alpha_1 + Z_{i1}\beta_0 + Z_{i0}\beta_1 + \eta_1c_i + u_{i1}$$

Multiply the lagged equation by λ the GDL decay parameter

$$\lambda A_{i1} = \lambda\alpha_1 + \lambda Z_{i1}\beta_0 + \lambda Z_{i0}\beta_1 + \lambda\eta_1c_i + \lambda u_{i1}$$

Add and subtract this from the right-hand side of the period two equation:

$$A_{i2} = \alpha_2 + Z_{i2}\beta_0 + Z_{i1}\beta_1 + Z_{i0}\beta_2 + \eta_2c_i + u_{i2} + [\lambda A_{i1} - \lambda\alpha_1 - \lambda Z_{i1}\beta_0 - \lambda Z_{i0}\beta_1 - \lambda\eta_1c_i - \lambda u_{i1}]$$

Apply the geometric distributed lag assumption $\beta_1 = \lambda\beta_0$ and $\beta_2 = \lambda^2\beta_0$

$$A_{i2} = \alpha_2 + Z_{i2}\beta_0 + Z_{i1}\lambda\beta_0 + Z_{i0}\lambda^2\beta_0 + \eta_2c_i + u_{i2} + \lambda A_{i1} - \lambda\alpha_1 - \lambda Z_{i1}\beta_0 - \lambda Z_{i0}\lambda\beta_0 - \lambda\eta_1c_i - \lambda u_{i1}$$

Simplify:

$$A_{i2} = \alpha_2 - \lambda\alpha_1 + \lambda A_{i1} + Z_{i2}\beta_0 + \eta_2c_i - \lambda\eta_1c_i + u_{i2} - \lambda u_{i1}$$

This same equation can be expressed as:

$$A_{i2} = \tau_2 + \lambda A_{i1} + Z_{i2}\beta_0 + \pi_2c_i + e_{i2}$$

where: $\tau_2 = \alpha_2 - \lambda\alpha_1$ $\pi_2 = \eta_2 - \lambda\eta_1$ $e_{i2} = u_{i2} - \lambda u_{i1}$

Hence, the more general expression, using the t subscript:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + Z_{it}\beta_0 + \pi_t c_i + e_{it}$$

Note that if the decay parameter λ is equal to the autoregressive parameter ρ in the original unobservables, the unobservables in the GDL equation will have no serial correlation. This strong (and not very plausible) assumption is called the “common factor restriction.”

$$e_{it} = u_{it} - \lambda u_{i,t-1} \quad u_{it} = \rho u_{i,t-1} + v_{it} \quad \rightarrow \quad e_{it} = \rho u_{i,t-1} + v_{it} - \lambda u_{i,t-1}$$

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 of Case 1. Vertically scaled-test scores. Row 1: Average rank correlation
 Row 2: Fraction of above average teachers misclassified as below average Row 3: Average theta

Small Teacher Effects		$\lambda=1$					$\lambda=.5$				
Assignment Mechanism	Estimator	POLS	DOLS	RE	FE	AB	POLS	DOLS	RE	FE	AB
	RG-RA		0.881	0.840	0.885	0.619	0.555	0.550	0.836	0.550	0.479
		0.15	0.17	0.15	0.26	0.29	0.36	0.17	0.36	0.32	0.27
		1.004	1.002	1.004	1.027	1.013	0.998	1.003	0.998	1.193	1.028
DG-RA		0.779	0.829	0.811	0.369	0.103	0.438	0.816	0.438	0.252	0.282
		0.20	0.18	0.19	0.34	0.45	0.36	0.19	0.36	0.36	0.37
		0.993	0.996	0.996	0.937	0.600	1.002	0.997	1.002	1.037	0.876
DG-PA		0.899	0.841	0.904	-0.335	0.166	-0.091	0.824	-0.091	-0.47	-0.138
		0.13	0.16	0.13	0.56	0.43	0.56	0.16	0.56	0.56	0.49
		1.342	1.001	1.268	-0.457	1.316	-0.092	1.005	-0.092	-1.256	-0.26
DG-NA		0.637	0.824	0.699	0.773	0.157	0.78	0.819	0.78	0.714	0.576
		0.28	0.20	0.25	0.18	0.42	0.21	0.21	0.21	0.20	0.27
		0.636	0.986	0.693	2.168	0.671	1.928	0.991	1.928	3.087	2.095
BG-RA		0.883	0.791	0.887	0.619	0.564	0.517	0.801	0.517	0.448	0.592
		0.15	0.20	0.14	0.26	0.29	0.38	0.20	0.38	0.33	0.27
		1.002	1.003	1.003	1.030	1.008	1.004	1.003	1.004	1.189	1.02
BG-PA		0.880	0.699	0.885	0.630	0.552	0.165	0.723	0.165	0.527	0.581
		0.15	0.26	0.14	0.26	0.28	0.50	0.23	0.50	0.28	0.27
		0.999	0.728	1.002	1.036	1.017	0.273	0.776	0.273	1.218	1.029
BG-NA		0.881	0.867	0.885	0.622	0.533	0.714	0.857	0.714	0.374	0.573
		0.16	0.16	0.15	0.26	0.29	0.32	0.18	0.32	0.37	0.28
		1.001	1.237	1.000	1.009	1.007	1.489	1.162	1.489	1.162	1.007
HG-RA		0.689	0.697	0.749	0.625	0.528	0.534	0.699	0.534	0.47	0.609
		0.25	0.26	0.23	0.26	0.29	0.36	0.25	0.36	0.33	0.26
		0.992	0.992	0.994	1.026	1.003	0.991	0.992	0.991	1.197	1.025
HG-PA		0.917	0.903	0.92	0.624	0.537	0.673	0.902	0.673	0.419	0.597
		0.11	0.13	0.11	0.26	0.29	0.31	0.13	0.31	0.36	0.27
		1.661	1.586	1.527	1.031	0.993	1.339	1.576	1.339	1.192	1.012
HG-NA		0.36	0.393	0.53	0.626	0.487	0.333	0.4	0.333	0.515	0.598
		0.4	0.38	0.33	0.25	0.32	0.43	0.39	0.43	0.29	0.27
		0.337	0.377	0.471	1.012	0.951	0.557	0.383	0.557	1.182	0.996

Table 3: Results from 100 replications of Case 1. Standardized test scores. Row 1: Average rank correlation
 Row 2: Fraction of above average teachers misclassified as below average Row 3: Average theta

Small Teacher Effects		$\lambda=1$					$\lambda=.5$				
Assignment Mechanism	Estimator	POLS	DOLS	RE	FE	AB	POLS	DOLS	RE	FE	AB
	RG-RA		0.855	0.868	0.855	0.509	0.573	0.836	0.877	0.836	0.538
		0.16	0.15	0.16	0.31	0.27	0.18	0.15	0.18	0.29	0.28
		0.513	0.512	0.513	0.565	0.449	0.73	0.727	0.73	0.884	0.705
DG-RA		0.566	0.737	0.566	0.211	0.394	0.437	0.812	0.437	0.218	0.347
		0.29	0.22	0.29	0.4	0.32	0.35	0.19	0.35	0.4	0.34
		0.513	0.506	0.513	0.487	0.37	0.735	0.722	0.735	0.758	0.631
DG-PA		-0.217	0.841	-0.217	-0.581	-0.179	-0.444	0.845	-0.444	-0.583	-0.242
		0.59	0.17	0.59	0.66	0.52	0.67	0.17	0.67	0.66	0.54
		-0.085	0.565	-0.085	-0.786	-0.093	-0.379	0.731	-0.379	-1.179	-0.271
DG-NA		0.890	0.678	0.89	0.699	0.761	0.893	0.806	0.893	0.708	0.69
		0.12	0.25	0.12	0.21	0.19	0.12	0.19	0.12	0.21	0.21
		0.92	0.482	0.920	1.706	0.904	1.712	0.738	1.712	2.618	1.595
BG-RA		0.602	0.799	0.602	0.438	0.57	0.569	0.829	0.569	0.417	0.588
		0.27	0.18	0.27	0.33	0.27	0.27	0.17	0.27	0.34	0.26
		0.516	0.513	0.516	0.561	0.448	0.738	0.729	0.738	0.876	0.716
BG-PA		0.013	0.691	0.013	0.607	0.568	-0.016	0.740	-0.016	0.606	0.593
		0.5	0.24	0.5	0.26	0.28	0.5	0.23	0.5	0.26	0.27
		0.012	0.322	0.012	0.546	0.401	-0.019	0.511	-0.019	0.882	0.684
BG-NA		0.889	0.891	0.889	0.389	0.519	0.856	0.893	0.856	0.334	0.542
		0.12	0.13	0.12	0.35	0.30	0.14	0.13	0.14	0.37	0.29
		0.913	0.69	0.913	0.585	0.505	1.287	0.889	1.287	0.880	0.745
HG-RA		0.729	0.677	0.729	0.486	0.566	0.755	0.701	0.755	0.51	0.584
		0.22	0.26	0.22	0.31	0.28	0.21	0.25	0.21	0.3	0.27
		0.509	0.506	0.509	0.569	0.453	0.726	0.719	0.726	0.889	0.706
HG-PA		0.884	0.904	0.884	0.399	0.556	0.859	0.913	0.859	0.422	0.582
		0.12	0.11	0.12	0.35	0.29	0.14	0.11	0.14	0.34	0.27
		0.729	0.818	0.729	0.547	0.444	0.914	1.125	0.914	0.846	0.685
HG-NA		0.488	0.347	0.488	0.593	0.573	0.576	0.391	0.576	0.615	0.584
		0.35	0.40	0.35	0.26	0.28	0.31	0.38	0.31	0.25	0.27
		0.238	0.174	0.238	0.578	0.454	0.440	0.272	0.44	0.914	0.722

Table 4: Results from 100 replications of Case 1. Vertically scaled-test scores. Row 1: Average rank correlation
 Row 2: Fraction of above average teachers misclassified as below average Row 3: Average theta

Large Teacher Effects		$\lambda=1$					$\lambda=.5$				
Assignment Mechanism	Estimator	POLS	DOLS	RE	FE	AB	POLS	DOLS	RE	FE	AB
	RG-RA		0.972	0.956	0.974	0.69	0.677	0.835	0.955	0.835	0.61
		0.07	0.09	0.07	0.23	0.24	0.19	0.09	0.19	0.27	0.24
		1.002	1.002	1.002	1.022	1.008	1.003	1.002	1.003	1.199	1.008
DG-RA		0.925	0.951	0.948	0.605	0.264	0.747	0.947	0.747	0.461	0.501
		0.11	0.09	0.09	0.26	0.39	0.23	0.1	0.23	0.3	0.31
		0.996	0.998	1	0.934	0.655	1.004	0.999	1.004	1.04	0.877
DG-PA		0.965	0.958	0.972	0.318	0.25	0.484	0.953	0.484	-0.051	0.365
		0.07	0.08	0.06	0.37	0.39	0.36	0.08	0.36	0.46	0.36
		1.184	0.99	1.125	0.33	0.908	0.489	1.003	0.489	-0.027	0.423
DG-NA		0.901	0.945	0.932	0.782	0.515	0.895	0.947	0.895	0.732	0.647
		0.14	0.11	0.11	0.18	0.3	0.14	0.11	0.14	0.21	0.25
		0.79	0.995	0.84	1.411	0.893	1.366	0.997	1.366	1.888	1.423
BG-RA		0.973	0.936	0.974	0.689	0.657	0.797	0.94	0.797	0.595	0.677
		0.07	0.1	0.07	0.23	0.23	0.2	0.1	0.2	0.28	0.24
		1.001	1.003	1.002	1.024	0.986	1.005	1.003	1.005	1.197	1.004
BG-PA		0.972	0.925	0.974	0.695	0.64	0.627	0.927	0.627	0.646	0.65
		0.07	0.12	0.07	0.23	0.25	0.29	0.11	0.29	0.25	0.25
		1	0.853	1.001	1.026	1.033	0.634	0.87	0.634	1.208	1.026
BG-NA		0.972	0.95	0.974	0.693	0.641	0.862	0.949	0.862	0.527	0.65
		0.07	0.1	0.07	0.23	0.26	0.2	0.1	0.2	0.32	0.26
		1.001	1.103	1	1.015	1.016	1.142	1.058	1.142	1.185	1.023
HG-RA		0.869	0.879	0.908	0.693	0.653	0.808	0.88	0.808	0.604	0.683
		0.16	0.16	0.13	0.23	0.24	0.2	0.16	0.2	0.27	0.23
		0.996	0.997	0.997	1.022	0.986	1	0.997	1	1.203	1.015
HG-PA		0.964	0.959	0.969	0.694	0.646	0.871	0.959	0.871	0.543	0.685
		0.07	0.08	0.07	0.23	0.25	0.18	0.08	0.18	0.31	0.23
		1.379	1.316	1.293	1.025	1.006	1.165	1.309	1.165	1.198	1.025
HG-NA		0.814	0.82	0.885	0.694	0.631	0.69	0.82	0.69	0.652	0.688
		0.21	0.21	0.16	0.23	0.25	0.25	0.21	0.25	0.24	0.23
		0.62	0.637	0.706	1.015	0.942	0.718	0.639	0.718	1.194	1.006