This is a preprint of an article published in the *Journal of Policy Analysis and Management* available at http://onlinelibrary.wiley.com/doi/10.1002/pam.21781/abstract

How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added

Steven G. Dieterle
University of Edinburgh

Cassandra M. Guarino
University of Indiana

Mark D. Reckase
Michigan State University

Jeffrey M. Wooldridge
Michigan State University

May 16, 2014

Abstract

The federal government's Race to the Top competition has promoted the adoption of test-based value-added measures (VAM) of performance as a component of teacher evaluations throughout many states, but the validity of these measures has been controversial among researchers and widely contested by teachers' unions. A key concern is the extent to which nonrandom sorting of students to teachers may bias the results and lead to a misclassification of teachers as high or low performing. In light of potential for bias, it is important to assess the extent to which evidence of sorting can be found in the large administrative data sets used for VAM estimation. Using a large longitudinal data set from an anonymous state, we find evidence that a nontrivial amount of sorting exists—particularly sorting based on prior test scores—and that the extent of sorting varies considerably across schools, a fact obscured by the types of aggregate sorting indices developed in prior research. We also find that VAM estimation is sensitive to the presence of nonrandom sorting. There is less agreement across estimation approaches regarding a particular teacher's rank in the distribution of estimated effectiveness when schools engage in sorting.

The work here was supported by IES Statistical Research and Methodology grant #R305D10028 and in part by a Pre-Doctoral Training Grant from the IES, U.S. Department of Education (Award # R305B090011) to Michigan State University. The opinions expressed here are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. The authors would like to thank Doug Harris and session participants at the Association for Education Finance and Policy annual meeting for helpful comments.

INTRODUCTION

The federal government's Race to the Top competition has promoted the adoption of test-based performance measures as a component of teacher evaluations throughout many states. The validity of test-based measures of teacher performance has been the subject of ongoing debate among researchers (see, for example, Harris, 2009, and Hill, 2009) and has been widely contested by teachers' unions, however. A key concern is the extent to which nonrandom assignment of students to teachers may bias the results and lead to a misclassification of teachers as high or low performing (Aaronson, Barrow, & Sander, 2007; Kane & Staiger, 2008; Rothstein, 2010; Koedel & Betts, 2011; Guarino, Reckase, & Wooldridge, in press). While the potential for nonrandom assignment to bias teacher value-added measures (VAMs) has been well recognized, little research has investigated how principals assign students to teachers in practice and the direct consequences of their assignment behaviors for ongoing teacher evaluations.

It is important to assess the extent to which evidence of nonrandom assignment can be found in the large administrative data sets used for VAM estimation. A few studies have approached this issue by considering broad statistical measures of sorting behavior (Clotfelter, Ladd, & Vigdor, 2006; Aaronson, Barrow, & Sander, 2007). Our study makes several key contributions to the literature. First, we develop tests of sorting that are more useful and precise than those previously used in the literature. Our tests lead us to revise prior conclusions as to the prevalence of sorting. Using a large longitudinal data set from an anonymous state,[1] we find clear evidence that student grouping exists in a nontrivial number of schools—particularly grouping based on prior test scores—and that the extent of grouping varies considerably both within and across schools, a fact obscured by the approaches developed in prior research.

[1] As a condition of data use, it has been requested that we do not refer to the state explicitly.

Second, we investigate teacher-student matching, extending the research beyond the simple investigation of tracking patterns. We distinguish between two components of nonrandom assignment and examine evidence of both: the grouping of students together on the basis of some characteristic and the systematic assignment of these groups to teachers. We find evidence to suggest that in many cases teachers are nonrandomly assigned to classes. In particular, teachers with higher measured prior effectiveness tend to be assigned to classrooms with higher average prior achievement.

Third, we show the implications of sorting for value-added using our statewide administrative data. We demonstrate that statistical methods matter and that they react very differently to different sorting scenarios. To do so, we define subsamples of school-grade-years that exhibit different grouping and assignment behaviors and then examine correlations within subsamples among VAMs estimated in different ways. We find the sensitivity of value-added to particular estimators differs in potentially important ways by subsample and that these differences align with predictions based on the standard value-added framework (Guarino, Reckase, & Wooldridge, in press). These findings have important consequences for the proliferation of teacher evaluations systems—and particularly for sanction-based policies such as the *deselection* or involuntary transfer of low performers (Winters & Cowen, 2013; Grissom, Loeb, & Nakashima, 2014)—that are currently the subject of intense scrutiny and controversy.

This paper is organized as follows. The following section provides a framework for thinking about the process by which principals assign students to teachers and discusses the implications for VAMs. The next section discusses the data used, which is followed by a section on previous approaches to identifying nonrandom assignment in administrative data. This discussion leads to a section that outlines our approach to detecting nonrandom grouping and

assignment and presents the findings. The penultimate section shows how our results on the grouping and assignment decisions of schools can be used to inform value-added estimation, and the final section concludes.

FRAMEWORK AND BACKGROUND

The theoretical motivation for value-added models of teacher performance typically rests on the specification of an education production function, in which achievement is modeled as a function of all relevant past and present child, family, and schooling inputs. Here, we focus on two estimating equations derived from this model that serve as the basis for most value-added estimation (for a detailed discussion of the derivation of these equations from the general model and the underlying assumptions see Hanushek, 1979, 1986; Todd & Wolpin, 2003; Harris, Sass, & Semykina, 2010; Guarino, Reckase & Wooldridge, in press). We start with a *lag score* specification controlling for prior achievement on the right-hand side:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + T_{it}\gamma + X_{it}\beta + c_i + u_{it} \qquad (1)$$

where
A_{it} is student $i's$ test score in time t
$A_{i,t-1}$ is prior achivement
T_{it} is a vector of teacher indicators
X_{it} are student and family characteristics
c_i is an unobserved student heterogeneity term
u_i is an unobserved error term

Occasionally, researchers use the gain in test scores as the dependent variable, effectively assuming that λ is equal to 1. We will refer to this as the *gain score* specification:

$$A_{it} - A_{i,t-1} = \tau_t + T_{it}\gamma + X_{it}\beta + c_i + (\lambda - 1)A_{i,t-1} + \nu_{it} \qquad (2)$$

Note that we include the additional term, *(□□□□ $A_{i,t-1}$*, on the right-hand side of equation (2) in order to emphasize the fact that if □≠□□□the choice to use a gain score specification may lead

to an omitted variables bias. This potential omitted variables problem will be the key focus of the analysis to follow.

Generally speaking, our ability to consistently estimate the teacher value-added coefficients (γ) hinges on what our estimation method requires about the correlation between teacher assignments (captured by T_{it}) and the unobserved factors affecting achievement, u_{it}, c_i, and, in the case of the gain-score specification, $(\square\square\square\square A_{i,t-1}$. Here, our concern lies with understanding how different student sorting and teacher assignment mechanisms employed by schools may affect these correlations and, in turn, value-added estimates based on equations (1) and (2).

Throughout the paper, we distinguish how students are grouped together into classrooms from how teachers are assigned to those classrooms. This leads to three distinct types of assignment mechanisms that each has different value-added implications: random grouping of students into classes and random assignment of teachers to those classes, nonrandom grouping of students but with random assignment of teachers to the classes, and finally nonrandom grouping with nonrandom assignment.

In the simplest case, students may be randomly grouped into classrooms with no consideration given to the within-class composition of student ability or to the quality of the teacher assigned to the groups. In this case, given a sufficient number of observations per teacher, estimates of teacher value-added based on either equation (1) or (2) will tend to perform well since any omitted factors that contribute to achievement will be uncorrelated with teacher assignment.

Now consider the case in which schools actively group students of similar ability together based on, say, prior achievement, demographic characteristics related to ability, or markers of

ability unobserved by those outside the school. Further assume that teachers are assigned to these classrooms in a systematic way according to each teacher's ability to raise achievement. Grouping based on observable student demographic characteristics (captured in X_{it}) is of less concern for estimators that partial out this correlation as both equation (1) and (2) control for those factors. Note, however, that grouping based on prior test scores coupled with nonrandom assignment of teachers based on ability to those groups is problematic for estimates based on equation (2). Specifically, *(□□□□* $A_{i,t-1}$ is non-zero, correlated with teacher assignment, and omitted from the model in this case. In contrast, by not restricting λ=1, estimates based on equation (1) are not subject to the same omitted variables bias. Effectively the cost of assuming λ=1 is higher in these cases.[2]

To help illustrate the implications of the bias, we appeal to a simple stylized example of estimating value-added with a model of only two teachers. While this certainly abstracts from the general problem of estimating equations (1) and (2), the simplified model will provide clear insights into the nature of the biases and inconsistencies that apply to the final estimation problem. Consider the case with two teachers (denoted Teacher 0 and Teacher 1) where the true education production function is given by:

$$A_i = \lambda A_i^L + \gamma T_i + u_i \qquad (3)$$

where
A_i is current achievement
A_i^L is prior achivement
$T_i = 0, 1$ is an indicator for having Teacher 1
u_i is a random error term
$0 \leq \lambda \leq 1$

[2] Cases of explicit test score grouping and assignment will also be more sensitive to possible misspecification of the current-score-lag-score relationship, including possible nonlinearities. In the analyses presented in this paper, we focus on specifications that assume a linear relationship between current and prior test scores. However, we ran sensitivity analyses that used specifications that included various polynomials in prior achievement and found virtually identical results.

It will be helpful to subtract λA_i^L from both sides of the equation:

$$A_i - \lambda A_i^L = \gamma T_i + u_i.$$

This framework can be used to illustrate the direction of the bias associated with using the gain score equation when $\lambda \neq 1$. In this simple setup, the OLS estimate of γ using the lag score specification is simply the Wald Estimator comparing the mean outcomes, expressed net of the prior score ($A_i - \lambda A_i^L$), for students with Teacher 1 to those for Teacher 0:

$$\begin{aligned} \hat{\gamma} &= E[A_i - \lambda A_i^L | T_i = 1] - E[A_i - \lambda A_i^L | T_i = 0] \\ &= E[\gamma T + u_i | T_i = 1] - E[\gamma T + u_i | T_i = 0] = \gamma \end{aligned} \quad (4)$$

Many of the potential issues we encounter will stem from using the "wrong" λ. In the case of the gain score specification, we have assumed $\lambda = 1$, implying the following estimating equation:

$$A_i - A_i^L = (\lambda - 1)A_i^L + \gamma T_i + u_i$$

Now the Wald Estimate of γ can be expressed as:

$$\begin{aligned} \hat{\gamma} &= E[A_i - A_i^L | T_i = 1] - E[A_i - A_i^L | T_i = 0] \\ &= E[(\lambda - 1)A_i^L + \gamma T_i + u_i | T_i = 1] - E[(\lambda - 1)A_i^L + \gamma T_i + u_i | T_i = 0] \\ &= E[(\lambda - 1)A_i^L | T_i = 1] + \gamma - E[(\lambda - 1)A_i^L | T_i = 0] \\ &= \gamma + (\lambda - 1)\left[E[A_i^L | T_i = 1] - E[A_i^L | T_i = 0]\right] \end{aligned} \quad (5)$$

If $\lambda \neq 1$, then $\hat{\gamma} \neq \gamma$ whenever the average prior achievement for students assigned to Teacher 0 is not the same as for Teacher 1 (i.e., $[E[A_i^L | T_i = 1] - E[A_i^L | T_i = 0] \neq 0$). This formulation also illustrates how the type of non-random assignment will matter for our ability to rank the two teachers correctly. Assume that Teacher 1 is the *better* teacher ($\gamma > 0$) and the school engages in positive assignment with the best prior performing students matched to the better teacher so that $E[A_i^L | T_i = 1] > E[A_i^L | T_i = 0]$. In this case, $(\lambda - 1) < 0$ and due to the assignment process $E[A_i^L | T_i = 1] - E[A_i^L | T_i = 0] > 0$ implying a negative bias in $\hat{\gamma}$. The magnitude of the bias term is driven by two factors: how far off the $\lambda = 1$ assumption is and the degree of grouping into the

classes. Importantly, if the negative bias is large enough (i.e., larger than γ), our estimate of Teacher 1's value-added relative to Teacher 0 will be of the wrong sign. While the true ranking of the teachers would place Teacher 1 above Teacher 2, our estimates would reverse this ranking. Just as important is the fact that if the bias is relatively small (i.e., smaller than γ), we may have a biased estimate of Teacher 1's value-added, but we will still get the relative ranking right. If the school engages in negative assignment placing the lowest prior performing students with the better teacher, the sign of the bias term will be the product of two negative components since now $E[A_i^L|T_i = 1] - E[A_i^L|T_i = 0] < 0$. In this case, no matter the size of the bias we still get the appropriate ranking of the two teachers. This provides a clear implication for assessing value-added. When the assignment process is generally negative (lower performing students with higher performing teachers), we expect rankings of teachers to be less sensitive to the choice of specification than under a positive assignment mechanism.[3] Of course, any policy use depending on the magnitudes of the value-added point estimates will still be adversely affected by this bias.

Finally, consider the third case in which schools nonrandomly group students based on ability as before, however, now the teachers are randomly assigned to these classes. Such a grouping and assignment policy may be driven by the belief that teachers can better target their teaching with more homogeneous classrooms, coupled with an effort to fairly assign teachers to classes. While the random assignment of teachers to the classes may, at first glance, seem to alleviate concerns over value-added estimates, this scenario can still lead to biased gain score estimates. This problem stems from once again leaving $(\square\square\square\square A_{i,t-1}$ in the error term and having some teachers assigned classes with better prior performing students by chance.

[3] Assignment based on a potential *match effect* (i.e., Teacher 1 is good with low-preforming students) is more complicated. The simplified example, however, is sufficient to illustrate the potential for divergent results between gain and lag score estimating equations when grouping and assignment is based on prior performance.

To illustrate the nature of the bias, let us return to our stylized example. For now, assume that the two teachers are equally effective so that $\gamma = 0$, and we can express our gain score estimate as:

$$\hat{\gamma} = (\lambda - 1)\left[E[A_i^L|T_i = 1] - E[A_i^L|T_i = 0]\right].$$

A clear implication of the $\gamma = 0$ assumption is that assignment based on teacher ability is not possible and must effectively be random. Assume that, by luck, Teacher 1 is given the better prior performing class so that $E[A_i^L|T_i = 1] - E[A_i^L|T_i = 0] > 0$ and γ is underestimated. Instead of ranking the teachers the same (the true ranking), we will rank Teacher 1 lower. Even though nonrandom assignment of teachers based on ability is impossible in this case, we have biased value-added estimates due to a correlation in the sample between uncontrolled for student ability and teacher assignment. Arguments for consistent estimation with nonrandom grouping, but random assignment of teachers are based on the number of classes per teacher becoming large. With random assignment to heterogeneous groups, a teacher's luck in one year may be balanced out in the future. With many classes per teacher and random assignment of teachers to classes, this small sample bias becomes less important with teachers receiving a range of class types over time. A similar argument can be made in the presence of grouping on observable student characteristics for estimators that do not partial out this correlation.

Returning to equations (1) and (2), assignment based on unobserved factors found in c_i or u_{it} are more difficult to characterize. For instance, prior test scores and student characteristics may only capture some of the considerations involved in making assignment decisions, but may miss differences in parental involvement in the decision process. To be clear, such unobserved factors driving assignment decisions will only lead to an omitted variables problem if they also affect current test performance. For the parental involvement example, we might suspect that

parents who actively pursue a particular teacher assignment may also provide more educational investments leading to a non-zero correlation between these other unobserved investments and both the teacher dummy variables and current test score. For the time-invariant factors (captured in c_i), methods that aim to account for this, such as student fixed effects or dynamic instrumental variable approaches, may be useful. However, such methods typically involve strong additional assumptions (either that λ=1 or that the errors in Equation [1] are serially uncorrelated) and greatly reduce the identifying variation, leading to potentially poor performance (Guarino, Reckase, & Wooldridge, in press). Importantly, prior test scores may serve as a decent proxy in these cases as they are a function of c_i. That is, highly involved parents have likely been involved throughout their child's education, so that part of this investment will be captured in the coefficient on prior scores. When the grouping decision is based on time varying unobserved factors, there is little that can be done to directly control for this. Once more, prior test scores may serve as a decent proxy for these factors if, say, parents are responding to factors that affected prior performance.

While not ubiquitous in the literature, gain-score formulations of the achievement regression have been used in recent work (for example, Jackson, 2009, Kinsler, 2011; Subedi, Swan, & Hynes, 2011; Koedel, Leatherman, & Parson, 2012; Lefgren & Sims, 2012; Oketch et al., 2012). The motivation for using the gain score rather than the lag score varies. It may be done to address issues of serial correlation (Jackson, 2009) or measurement error (Koedel, Leatherman, & Parson, 2012) in test scores, or to take advantage of panel data estimators aimed at improving efficiency (Hierarchical Linear Models, Feasible GLS, empirical Bayes) or tackling identification issues (Fixed Effects) that are potentially inconsistent when lagged dependent variables are present.

Given concerns that test scores are noisy measures of achievement, it is worth considering the measurement error motivation for using the gain score in more detail. If the measurement error satisfies the classic errors in variables (CEV) assumptions, then it can lead to an attenuation bias in the estimate of λ in equation (1). Importantly, under the CEV assumptions, measurement error in the dependent variable does not lead to biased estimates. This fact helps motivate the use of the gain-score specification in the presence of measurement error, as it moves all of the error into the dependent variable.

The first thing to note here is that we are not evaluating the estimate of λ, but are instead concerned with the estimated teacher effects. The attenuation bias in the estimate of λ is propagated to the teacher effect estimates depending on the relationship between prior scores and teacher assignments. The role of attenuation bias in λ due to measurement error can also be explored in our stylized model. Recall that our lag score specification was represented by the Wald estimator for the following model:

$$A_i - \lambda A_i^L = \gamma T_i + u_i.$$

However, with classical measurement error in prior achievement, we estimate $\tilde{\lambda} < \lambda$ yielding the modified equation:

$$A_i - \tilde{\lambda} A_i^L = (\lambda - \tilde{\lambda}) A_i^L + \gamma T_i + u_i.$$

Now the appropriate Wald Estimate is:

$$\begin{aligned} \hat{\gamma} &= E\left[A_i - \tilde{\lambda} A_i^L \middle| T_i = 1\right] - E\left[A_i - \tilde{\lambda} A_i^L \middle| T_i = 0\right] \\ &= E\left[(\lambda - \tilde{\lambda}) A_i^L + \gamma T_i + u_i \middle| T_i = 1\right] - E\left[(\lambda - \tilde{\lambda}) A_i^L + \gamma T_i + u_i \middle| T_i = 0\right] \\ &= \gamma + (\lambda - \tilde{\lambda}) \left[E[A_i^L | T_i = 1] - E[A_i^L | T_i = 0]\right]. \end{aligned} \quad (6)$$

Due to attenuation,[4] $(\lambda - \tilde{\lambda}) > 0$, so the sign of the bias term depends on the sign of $\left[E[A_i^L|T_i = 1] - E[A_i^L|T_i = 0]\right]$. If it is positive (Teacher 1 has the better students), then $\hat{\gamma} > \gamma$, and we have overestimated Teacher 1's value-added relative to Teacher 0. If the assignment is negative (Teacher 0 has the better students), we will underestimate γ, which implies overestimating Teacher 0's ability relative to Teacher 1. More generally, teachers with the better prior performing classes will be made to look better due to the attenuation bias in λ. Intuitively, the attenuation bias will reduce the estimated effect of prior scores. This implies that when we see good prior performing students do better on current tests, too much of that achievement will be attributed to their teacher instead of their prior performance. On the other hand, teachers with poorer performing students will look worse since the part of their student's poor performance that is attributable to past achievement is underestimated.

While there will certainly be a trade-off between the attenuation bias of λ when estimating equation (1) and assuming $\lambda = 1$ in equation (2), when teacher assignments are based on prior scores, it is not clear which bias is more severe. Indeed, it seems quite plausible that an attenuated λ is less of a concern than assuming $\lambda = 1$. Indeed, Guarino, Reckase, and Wooldridge (in press) find simulation evidence that classical measurement error in test scores leads to only very small biases in the ranking of teachers in the data generating processes they consider. The sensitivity to measurement error may well be different when considering the magnitude of VAMs rather than the ranking.

[4] The extent of the attenuation depends on the variance of the measurement error and the underlying knowledge being measured once the other covariates have been partialed out. When more of the variation in true knowledge is explained by the teacher indicators (as well as other covariates) the attenuation is stronger. Therefore, the strength of the relationship between the true knowledge and teacher assignment will influence the attenuation. The equations presented here are helpful for considering the extent to which a given level of attenuation is propagated to the teacher effect estimates.

Second, it is likely that the measurement error in test scores does not meet the CEV assumptions, as it is derived from the aggregation of the error on separate item responses by students. As such, the attenuation bias result does not necessarily hold, and there may be problems with mismeasured dependent variables leaving specification (2.2) susceptible to bias as well.[5] Ultimately, the analysis that follows will help identify scenarios in which the distinction between using equation (1) or (2) may lead to empirically important differences in the ranking of teachers. While the issues underlying the motivation for the gain score specification may certainly be important, it is equally important to weigh these considerations next to the cost outlined above of assuming λ=1, particularly if grouping based on prior test scores is common.

Finally, it is important to emphasize that the lagged test score serves two functions: one is to correctly partial out prior test scores and the other is to proxy for factors related to the assignment mechanism. If observed prior scores are the basis for assignment, then they are important, and properly measured, controls.

The focus of the first analytic section of this paper is to develop ways to best identify different grouping and assignment mechanisms in the types of administrative data sets commonly used for value-added in order to inform VAM estimation decisions. While it is fundamentally impossible to identify perfectly the scenarios outlined above, it *is* possible to systematically characterize situations in which some estimators and models are likely to deviate from each other. Once detection strategies for grouping and assignment have been developed, we demonstrate their importance in influencing the results of value-added estimation.

DATA

[5] In fact, gain-score specifications can perform especially poorly under the measurement error induced by Item Response Theory scaling procedures—i.e., nonclassical measurement error (Guarino et al., 2013).

The data used for this study come from the administrative records of a large and diverse state in the southeastern region of the US. The data tracks students and teachers in grades 1 through 6 in the state's public school system from the 2000–01 to the 2007–08 school year. With individual student test scores and course indicators linking students to their teachers, the data are ideal for the estimation of teacher value-added. Importantly, the presence of course-level linkages (as opposed to the school grade or exam-proctor linkages found in some similar data sets) allows us to identify the set of teachers a student could have potentially been assigned to in a given year. Throughout the paper, we use student test scores in mathematics from the statewide standardized year-end exams. Typical of such large administrative data sets, there is limited student information—primarily demographics and information on school attendance/absences. The student demographics are race/ethnicity, gender, disability status,[6] limited English proficiency (LEP), free-or-reduced-price lunch eligibility (FRL), and country of birth. In addition, the data include demographic (race/ethnicity and gender) and professional (certification status, degree level, and experience[7]) variables for teachers. The set of student and teacher characteristics will allow us to examine the extent of sorting and matching on observables. Finally, we limit our analysis to teachers teaching a regular mathematics course (typically in middle school) or a comprehensive general education class (typically in elementary school). Most of the analysis also focuses on school-grade-years with at least two teachers (i.e., situations in which nonrandom grouping and assignment is possible), leaving 26,177 school-grade-years covering 2,533 schools.

[6] We distinguish between students with common *high incidence* disabilities and those with less common *low incidence* disabilities. The disability categories coded as high incidence are Educable Mentally Handicapped, Trainable Mentally Handicapped, Orthopedically Impaired, Speech Impaired, Language Impaired, Emotional/Behavioral Disability, Specific Learning Disability, Autistic Spectrum Disorder, Other Health Impaired. The disability categories coded as low incidence are Deaf or Hard of Hearing, Visually Impaired, Hospital/Homebound, Profoundly Mentally Handicapped, Dual Sensory Impaired, Severely Emotionally Disturbed, Traumatic Brain Injured, Developmentally Delayed, and Established Conditions.

[7] Experience is measured as the sum of prior years spent in public and private schools both within and outside the state studied.

Table 1 displays descriptive statistics for our main analysis sample. In addition to means and standard deviations for the student and teacher variables, we will use throughout the paper, we also provide counts of the number of students, teachers, and school-grade-year cells. In each grade, we have roughly 900,000 students. In fourth and fifth grade, there are over 46,000 teachers, while in sixth grade there are just under 15,000, with many math teachers teaching multiple sections.

[Insert Table 1 Approximately Here]

PREVIOUS APPROACHES TO IDENTIFYING NONRANDOM GROUPING

Given the difficulty of detecting nonrandom assignment to teachers, most researchers approach the problem by investigating evidence of some form of tracking or grouping of students into classrooms. While many papers have considered the teacher assignment decision quite generally from both quantitative and qualitative perspectives (Conger, 2005; Feng, 2010; Kalogrides, Loeb, & Beteille, 2011), we are concerned with approaches that allow researchers to distinguish between different assignment processes and categorize schools accordingly. Here we review two particularly influential approaches that have been applied to large administrative data sets from the Chicago Public Schools (Aaronson, Barrow, & Sander, 2007) and North Carolina (Clotfelter, Ladd, & Vigdor, 2006). Both approaches have been used in a variety of papers to evaluate and justify the estimation of education production functions (Ammermuler & Pischke, 2009; Gao 2012; Goldhaber & Hansen, 2010, 2012; Goldhaber, Cowan, & Walch, 2012; Koedel, 2009; Koedel & Betts, 2010, 2011; Koedel, Leatherman, & Parsons, 2012; Lavy, 2011; Lugo, 2011; Whitmore, 2005).

Aaronson, Barrow, and Sander (2007) (ABS) calculate the average within-class standard deviation of prior test scores for separate grade and year groupings. This average *Actual* standard

deviation is then compared with two counterfactual standard deviations. The first counterfactual, referred to as *Perfect Sorting*," is obtained by ordering students based on their prior test score and creating counterfactual classrooms based on this hierarchy. A second, *Random Sorting*, counterfactual is created in a similar way by ordering students randomly. The goal of this exercise is to see if the average Actual standard deviation is closer to the Perfect or Random sorting counterfactuals. In their study of data from Chicago Public high schools, ABS found that the Actual was much closer to the Random Sorting outcome. Applying this approach to our data yields similar results.[8]

Clotfelter, Ladd, and Vigdor (2006) (CLV) look for evidence of student grouping in North Carolina by conducting a series of six chi-squared tests of whether students' classroom assignments were independent of the following characteristics: gender, race, FRL, attended same school in the prior year, had an above average prior test score, and the prior year's report of parental education. The chi-squared tests are performed by school on data from a single year and are pooled over third, fourth, and fifth grade. CLV then categorize the 44.9 percent of schools that do not reject the null of random assignment in all six cases as non-tracking. Once more, applying this approach to our data gives similar results, with 54 percent of schools classified as non-tracking.

Both the ABS and CLV approaches have been used as evidence of random student grouping in order to justify the validity of education production function estimation. However, by pooling data together and observing an aggregate measure in the ABS approach, the method misses important heterogeneity in the sorting behavior of schools. Also, the test focuses on a single student characteristic while not exploring other observable characteristics that may drive

[8] Results available upon request from the authors.

the student grouping decision. While the CLV approach considers other characteristics, each is tested independently without considering the potential relationships among different characteristics. Thus, the CLV approach is susceptible to mischaracterizing the basis for sorting. For example, it can easily identify a school as failing the test of independence for both prior test scores and free-and-reduced-price lunch status, when in fact the perceived grouping based on FRL status is driven entirely by poorer test performance of FRL students.

In this paper, we implement methods that allow us to uncover the heterogeneity in sorting behavior and take into consideration the relationship among several student characteristics. Further, we move beyond measures of student grouping and tackle the more difficult problem of detecting nonrandom teacher assignment to groups of students. Finally, we demonstrate how grouping and assignment affect the results of value-added teacher performance estimation using different specification and estimation choices. This discussion is particularly important for policy applications in which it is not possible to isolate random grouping subsamples of schools in implementing policies.

INVESTIGATION OF STUDENT GROUPING AND TEACHER ASSIGNMENT

Nonrandom Grouping of Students into Classrooms

The student grouping and teacher assignment decision is a complex choice problem facing the school administration with potential input from others including teachers and parents. Considerations in such decisions are varied, including: achievement goals, noncognitive outcomes, peer interactions, and class size constraints, among many others. Our interest lies in detecting observable differences across classroom groups that result from the student-teacher assignment decision and that may impact value-added estimators. We therefore estimate a series of Multinomial Logit (MNL) models of student assignment to classrooms separately for each

school-grade-year combination, modeling the probability a student is assigned to a particular teacher given the student's characteristics: [9]

$$P(T = j|x) = \frac{\exp(x\delta_j)}{1+\sum_{h=1}^{J}\exp(x\delta_h)}, \quad \text{where } j = 1, 2, \ldots, J \text{ indexes teachers in the school} - \text{grade} - \text{year} \tag{7}$$

The student characteristics in **x** include the student's lagged math score, indicators for race/ethnicity, gender, disability status, free or reduced price lunch status, limited English proficiency, whether a student was foreign born, new to the school, and the number of schools the student attended in the prior year.[10] We are primarily interested in whether each of the characteristics is a statistically significant predictor of which teacher a student is assigned and less interested in the magnitude of the estimated partial effects, denoted $\partial P(T = j \mid x)/\partial x_k$. Therefore, for each MNL, we test that null that the partial effect for a given characteristic, x_k, is zero for all teachers:

$$H_0: \frac{\partial P(T=1|x)}{\partial x_k} = \frac{\partial P(T=2|x)}{\partial x_k} = \cdots = \frac{\partial P(T=J|x)}{\partial x_k} = 0 \tag{8}$$

We limit our analysis to cases in which the MNL likelihood function maximization converged within 300 iterations, covering over 99 percent of the possible cases.[11]

This procedure gives a large number of results (up to 26,177) to be summarized. In Table 2, we show the percentage of school-grade-years for which a particular characteristic was found to be statistically significant at the 5-percent level (rejecting the null in equation [7]). The table

[9] Although essentially a reduced form approach, the properties of the MNL as a good approximation in modeling choice probabilities are well known (see Cramer, 2007, for the binary case and McFadden, 1974).
[10] The potentially time-varying student characteristics are recorded in the fall of the school year and are therefore based on prior evaluations, rather than responding to current teacher or class assignments.
[11] In order to improve the convergence rate, we use three maximization algorithms: Newton-Raphson for the first 100 iterations, Davison-Fletcher-Powell for the next 100, and Broyden-Fletcher-Goldfarb-Shanno for the final 100.

also displays the number of times the hypothesis in equation (7) was tested for a given variable.[12] By looking at these rejection rates, we gain insight into the observable characteristics of students that tend to be related to classroom assignment across the state.[13] We begin with MNL estimates from models that only included the lagged test score. This set of results ties directly to the prior literature that looks for grouping based on prior achievement in isolation from other characteristics. The significance rates for these MNL estimates are found in the first row of Table 2. We see that roughly 25 percent of the school-grade-year cells show evidence of grouping based on prior achievement in both fourth and fifth grade. In sixth grade, this percentage is much higher at 67 percent. This is perhaps not surprising, as in the state studied here many students make a promotional school change in grade 6. More specialization in courses occurs as students move to middle school. Moreover, if administrators in the new school have less private information on the student's ability, we might expect them to use observed prior achievement to engage in ability grouping. Furthermore, these new middle schools tend to be larger, drawing from several feeder elementary schools, allowing the schools more opportunity to create differentiated sections of courses.

[Insert Table 2 Approximately Here]

Moving down the table, we present rejection rates from MNL estimates including the student covariates. These results directly allow for relationships between prior test scores and student characteristics that had been ignored in previous approaches. Among the characteristics, only the lagged test score shows evidence of being predictive of teacher assignment with a

[12] Note that the number of times a particular hypothesis test was run may be less than total number of estimates; for example, if there were no Asian students in the school, then that particular hypothesis test could not be run.
[13] By looking at statistical significance, our approach is easy to apply uniformly across a large number of estimates and, as we show later, is effective at identifying cases where value-added estimation is sensitive to non-random grouping. While potentially interesting, a comparison of the magnitudes of partial effects becomes much less tractable with more than two teachers.

substantial degree of frequency. While the rejection rates for prior scores in this specification fall slightly compared to those in the first row, suggesting that some of the perceived ability grouping may be driven by other characteristics, the general pattern across grades remains the same.

Characteristics of Schools that Engage in Nonrandom Achievement Grouping

We next examine which characteristics of schools are associated with being more likely to reject the null in equation (7) for the student's prior test score. To do so, we further disaggregate the rejection rates in Table 2 across quartiles of school-level student characteristics. Table 3 presents these results using the 5-percent rejection rates for the prior test score from the estimates of MNL models that included other student covariates. Note the u-shaped pattern across the distribution of black student populations in G4 and G5, with higher rejection rates in the low and high proportion black schools. This may relate to the extent of racial heterogeneity there is within schools (i.e., in more mixed schools, race becomes a characteristic to sort on in lieu of or in addition to using test scores, limiting the role test score sorting may play). A similar pattern holds for the FRL populations as well. Moving on, we see higher rejection rates for larger schools, those with a larger proportion of Hispanic and LEP students, and lower proportion disabled (G6 only). On the surface, the higher rejection rates for larger schools fits nicely with the idea that larger schools are afforded more opportunities to create specialized classes. However, in this context we cannot separate this effect from the fact that larger schools may have more precise estimates due to having more observations in the MNL.[14]

[Insert Table 3 Approximately Here]

[14] In simulations with students randomly grouped into classes, the rejection rate for the MNL test are 0.05, 0.02, 0.01, and 0.01 for school-grade-years with 40, 80, 160, and 240 students, respectively. These school-grade-year sizes were chosen to roughly reflect the actual distribution of size in our data. Simulation details are available upon request and are similar in nature to those found in Guarino, Reckase, and Wooldridge (in press).

The above evidence points to three key improvements over the prior approaches to identifying grouping. First, the across school variation in grouping patterns explored in Table 3 would be missed entirely by the ABS approach. Second, the low rejection rates for other student characteristics suggest that, conditional on prior test scores, there is little scope for these characteristics to explain student grouping. This result differs from what would be concluded by the CLV approach, which considers each characteristic independently. Finally, to highlight the heterogeneity that might be missed by previous approaches to identifying grouping, we can explore the stability of the grouping category for the same school-grades over time or across grades within the same school-year cell. We see the school-grades fall in different categories in consecutive years between 35 and 38 percent of the time and, of the school-years with multiple grades, 20 percent are categorized differently across the grades. This within school variation in grouping would be missed by the CLV and ABS approaches.

Nonrandom Assignment of Teachers to Classrooms

The previous estimates attempt to uncover evidence of nonrandom grouping of students together into the same classrooms. As discussed in Framework and Background, such nonrandom grouping may lead to issues for value-added estimation even in the presence of the random assignment of groups to teachers. However, the systematic assignment of teacher to these groups of students raises additional concerns. Of particular concern for value-added estimation is whether high or low ability students are assigned teachers who are better or worse at improving achievement. The following approach is aimed at identifying cases of explicit matching of students to particular teachers based on the ability (or characteristics) of both the students and teachers.

In order to explore the potential matching of students to teachers in this manner, we modify the previous MNL approach to include match-specific variables describing some aspect of a potential student-teacher match. We will refer to these new estimates as the *matching logit* estimates.[15] The estimates of δ_j from the previous MNLs varied by teacher (i.e., a different δ for each teacher) to give an indication of the likelihood that a student with particular characteristics is assigned to a particular teacher (indexed by j) relative to a comparison teacher. The matching logits do the same for student characteristics, but estimate a single coefficient across all teachers for the match variables discussed below for each school-grade-year cell, giving us an indication of whether the assignment process seemed to favor that particular type of match.

In practice we estimate four separate models each with a different match-specific variable aimed at capturing some aspect of the student-teacher match that is related either directly or indirectly to ability matching. The four match variables (1) pair student-teacher ability measures, (2) examine the consistent placement of high ability students with particular teachers, (3) pair more experienced teachers with high-performing students, and (4) match teachers and students on the basis of race.

The first *MATCH* variable relies on an OLS estimate of prior teacher value-added based on the lag score specification as a measure of teacher ability. We use value-added estimated using all the prior years of data we have for the teachers. We then create a variable indicating whether a given teacher is above average in prior value-added (High Value-Added) compared with all other teachers in that school-grade-year cell, denoted $VAM_j^H = 1(VAM_j > \overline{VAM})$.[16] We

[15] Such a model can be estimated in Stata using the *asclogit* command. Again, we apply a reduced form approach to obtain information about the realized student-teacher assignments.
[16] For this match variable, we necessarily must drop teachers without prior value-added.

also define an indicator for a student with above average prior achievement in that cell (High Achieving) $ACH_i^H = 1(A_{i,t-1} > \bar{A}_{t-1})$. The *MATCH* variable is then defined by[17]

$$MATCH_{ij}^1 = \begin{cases} 1 \; if \; VAM_j^H = 1 \text{ and } ACH_i^H = 1 \; OR \; VAM_j^H = 0 \text{ and } ACH_i^H = 0 \\ 0 \; if \; VAM_j^H = 0 \text{ and } ACH_i^H = 1 \; OR \; VAM_j^H = 1 \text{ and } ACH_i^H = 0 \end{cases}.$$

Here, a positive estimate of γ suggests the school prefers to have high (low) ability students matched with high (low) ability teachers, while a negative estimate suggests that it prefers having high (low) ability students paired with low (high) ability teachers.

While the approach based on estimated value-added is certainly informative and interesting, it rests on having a reliable estimate of value-added. As a major part of the motivation for this exercise is to determine conditions under which informative value-added estimation may be plausible, it is difficult to make this assumption ex ante. In order to address this, we create a second match variable that does not rely on a potentially inconsistent value-added estimate. We view observing the *consistent placement* of teachers with high or low performing students as a potential marker of ability matching. To be clear, this does not presume anything about the ability of the teachers assigned to the classes, but, in the presence of the sort of ability assignment we are concerned with, we would expect the same teachers to have similar classes year-to-year. Finding evidence of consistent placement does not necessarily indicate ability matching is taking place; however, the absence of consistent placement is certainly suggestive that ability assignment is not likely, or at least not persistent.

To operationalize this concept, the second match variable is created in a similar manner using the teacher's prior incoming class average of student scores, rather than value-added. We define an indicator for a teacher having an above average incoming class the previous year as

[17] Note that by using prior value-added, these estimates are based on different cohorts of students than those we are using to make the match variables avoiding any mechanical relationship between prior VAM and prior test scores.

$CLASS_j^H$. Importantly, this measure is based on the performance of those students the year before they had that teacher. Therefore, the second *MATCH* variable is defined as

$$MATCH_{ij}^2 = \begin{cases} 1 \; if \; CLASS_j^H = 1 \text{ and } ACH_i^H = 1 \; OR \; CLASS_j^H = 0 \text{ and } ACH_i^H = 0 \\ 0 \; if \; CLASS_j^H = 0 \text{ and } ACH_i^H = 1 \; OR \; CLASS_j^H = 1 \text{ and } ACH_i^H = 0 \end{cases}.$$

Third, we consider whether more experienced teachers receive higher performing students, given the finding in some prior research that more experienced teachers may be more effective at raising test scores (Goldhaber, 2008). First we define an indicator function for a teacher with above average experience in that school-grade-year-cell (High Experience) by $EXP_j^H = 1(EXP_j > \overline{EXP})$. The third *MATCH* variable is then defined as

$$MATCH_{ij}^3 = \begin{cases} 1 \; if \; EXP_j^H = 1 \text{ and } ACH_i^H = 1 \; OR \; EXP_j^H = 0 \text{ and } ACH_i^H = 0 \\ 0 \; if \; EXP_j^H = 0 \text{ and } ACH_i^H = 1 \; OR \; EXP_j^H = 1 \text{ and } ACH_i^H = 0 \end{cases}.$$

Finally, we create a racial match variable. Schools may choose to match students to teachers based on race for a variety of reasons (see Dee, 2004, for evidence that racial matches improve student achievement). Given potential differences in student or teacher ability by race, this may indirectly lead to ability matching. The indicator for whether a potential student-teacher match represents a racial match is constructed as follows:

$$MATCH_{ij}^4 = \begin{cases} 1 \; if \; RACE_i = RACE_j \\ 0 \; if \; RACE_i \neq RACE_j \end{cases}.$$

Two matching logits are estimated separately for each *MATCH* variable, one with and one without a set of student specific variables.[18] In specifications that include the student covariates, we exclude those student-level variables that were used to create the applicable *MATCH* variable. For instance, we exclude the child race indicators for the race match variable

[18] The included student covariates are the number of absences the prior year, race indicators, the student's prior achievement, indicators for gender, FRL status, and whether a student is new to a school. We utilize the same maximization scheme as for the MNL, allowing for 300 iterations alternating between three maximization algorithms.

and the student's prior test score for the other three match variables. As before, we present rejection rates for the null that $\gamma=0$. We also present rejection rates for one-tail tests to look for evidence that $\gamma>0$ or $\gamma<0$, as unlike in the MNL case, the sign of γ provides information on the sorting behavior. We also display the total number of hypothesis tests.

Beginning with the student-score-teacher-value-added match variable, we see that with no additional covariates we reject the null that schools do not match students to teachers based on the prior performance of both students and teachers 15 percent and 16 percent of the time in fourth and fifth grade, respectively. We find the evidence of this sort of matching is much stronger in sixth grade with a rejection rate of 42 percent. We find statistically significant negative assignment between 7 percent and 16 percent of cases. There is evidence that positive assignment is much more common among the school-grade-year cells tested. When including the set of student covariates, we see the rejection rates fall slightly in all grades, suggesting that some of the perceived matching of high (low) prior performing students with high (low) prior value-added teachers uncovered in the first three columns is being driven by the grouping of students with similar observed characteristics into classrooms.

[Insert Table 4 Approximately Here]

The evidence here suggests that ability matching, while not the prevailing assignment mechanism, influences principals' decisions to assign students to teachers in a nontrivial number of schools—as we reject the null that the coefficient on the match variable is zero in 10 to 15 percent of 4th-grade school-year cells, 11 to 16 percent of 5th-, and 33 to 42 percent of 6th-grade school-year cells. Of course, it should be noted that with this many estimates one might expect a rejection about 5 percent of the time, so some of these lower percentages may not be indicative of a noticeable amount of nonrandom assignment. On the other hand, however, if other criteria

related to student and teacher ability are being used to make decisions, to the extent that our ability measures are only proxies, we may understate the extent of ability matching.

The match variable based on the incoming ability of the teacher's previous class is found to be statistically significant more frequently than the value-added based indicator for all but the negative one-tail tests (bottom panel of Table 4). This is perhaps not surprising, as we have noted that this measure will likely capture any sort of persistent assignment of teachers to high or low performing students. The rejection rates follow a similar pattern to the VAM-based matching case as we add covariates. However, these results are stronger than those for matching on the teacher's prior value-added—in some cases, quite a bit stronger. These findings suggest that regardless of whether principals are matching students to teachers based on ability, many are consistently assigning certain teachers high or low ability classes. In particular, in 51 to 64 percent of the school-years in the sample, 6th-grade teachers who had high ability classes in the past year were likely to get high ability students again, which, as we have shown above, may cause problems for value-added.

From the teacher experience/student test score match, we see that in 14 percent and 15 percent of 4th- and 5th-grade cells there is evidence of matching based on this characterization. However, in sixth grade, nearly half of all cells reject the null. This would seem to suggest that many middle schools assign more experienced teachers to classrooms of better prior performing students. Adding other student characteristics reduces the rejection rate to 36 percent. Here, we also see that some schools show evidence of negative matching (high experience with low performers).

Finally, for the racial match variable, we see that when excluding other covariates, nearly10 percent of cases show some evidence of matching based on this characteristic for fourth

and fifth grade and nearly 18 percent for sixth grade. The inclusion of the student covariates does little to change the overall rejection rates in the two earliest grades; however, it does reduce the rejection rate for sixth grade to roughly 9 percent. Importantly, none of the school-grade-years tested provide evidence of explicit racial *mismatch* (a preference for assigning students to teachers of a different race) as shown by the second row displaying 0 percent for each grade and specification.

It is worth noting the lower convergence rates for the matching logit than for the MNL estimation. For instance, in fourth grade there were 11,116 school-grade-year cells in which the MNL estimation converged when including our full set of covariates while only 3,993 did so in the racial matching logit estimation with student covariates.[19] This represents a nontrivial drop in the number of results and serves as a limitation of this approach. However, for the school-grade-cells in which estimation was possible, this approach provides useful information on the underlying preferences driving student-teacher assignment decisions. Furthermore, in more localized settings with only a handful of schools, it may be possible to appropriately troubleshoot in order to find specifications and maximization algorithms that perform better.

COMPARING THE PERFORMANCE OF COMMON VALUE-ADDED ESTIMATORS UNDER DIFFERENT ASSIGNMENT CONDITIONS

Our preceding analyses have established the fact that schools can differ widely in the observed use of student tracking and teacher assignment mechanisms. Given the importance of

[19] The lower rates of convergence can be the result of several factors. In many cases, multicollinearity creates flat regions of the likelihood function. For instance, when all the teachers in a school-grade-year are of the same race, say White, there is no within-student-across-teacher variation in the racial match variable. Effectively the matching logit becomes a MNL with a single White/Other Race indicator rather than the set of race indicators in the MNL we estimate. With little variation across students in this variable (i.e., if most students are White), the match variable becomes highly collinear with the constant in the model. Since the other match variables rely on an above-average below-average distinction within school-grade-years, this leads to more variation and better convergence. Generally, the matching logit requires a more complicated likelihood function that can be more difficult to estimate. See Gould (1996) for a discussion of the convergence of MLE estimation in Stata.

understanding the context driving such decisions for the estimation of teacher value-added, we now consider how to use the information gathered so far to inform VAM estimation.

We first describe a set of four value-added estimators in fairly common use and discuss how they should be expected to perform in random versus nonrandom grouping and assignment scenarios. Under random grouping and assignment, the estimators can be expected to show more agreement in their rank ordering of teachers than under nonrandom grouping and assignment (Guarino, Reckase, & Wooldridge, in press). To test our predictions, we estimate teacher value-added in mathematics and reading[20] using subsets of our data based on the degree of nonrandom grouping and assignment, and we display rank correlations within each subsample among the estimates produced by the different estimators.

Using the MNL results that included all student covariates, we distinguish between two types of school-grade-year cells, those that exhibited evidence of grouping students based on rejecting the null that prior test scores were related to classroom grouping at the 5-percent level (the *Grouping* subsample) and those that did not (the *Non-Grouping* subsample).[21] The labels Grouping and Non-Grouping were chosen to emphasize that the MNL results tell us about the grouping of students into classes, but nothing about the subsequent assignment of teachers to these classes.

To address the potential teacher assignment decisions, we similarly divide our sample of school-grade-years into *Positive Matching*, *Negative Matching*, and *Non-Matching* subsamples based on the teacher VAM/student score matching logits that included additional student

[20] To save space throughout, we have reported only mathematics results up to this point. However, as the value-added implications are the key focus of the study, we provide the reading results here for comparison. The grouping and assignment categories described below are based on analogous reading analyses, the results of which are available upon request.

[21] While we could use other student characteristics to define groups, the fact that we found little evidence of grouping on the other characteristics, conditional on prior scores, implies that the prior score results are the most empirically interesting. The results are robust to using a 10-percent significance level cutoff.

covariates. While this distinction explores the grouping and assignment decision in more detail, there are advantages to using the MNL results as well. Namely, with higher rates of convergence and not requiring prior value-added, the MNL based subsamples give better empirical coverage while still reflecting grouping scenarios that may lead to problems in identification. In the end, both can be thought of as providing markers of potentially problematic grouping/assignment mechanisms.

Estimation Approaches

We estimate teacher value-added using separate grade-year cross sections of studentlevel observations and employ four separate estimation approaches involving the two estimating equations discussed in the Framework and Background section.[22] The main features of estimation that we vary are the lag score versus the gain score specifications and the treatment of the teacher effects as fixed or random. The specifications with fixed teacher effects (equations [1] and [2]) are estimated by Ordinary Least Squares (OLS), include teacher indicator variables, and retain their coefficients as our teacher effects, yielding our OLS Lag and OLS Gain estimators.

Teacher effectiveness estimates derived from the lag-score and gain-score specifications would be expected to differ under nonrandom student grouping and nonrandom teacher assignment based on prior test scores. As such, we expect the two approaches to yield similar

[22] We also estimate teacher value-added using student-level panel data (i.e., with several years of data for each student) to estimate value-added for teachers in across multiple grades and years. Those results—which do not yield qualitatively different conclusions—are presented in the appendix. All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://www3.interscience.wiley.com/cgi-bin/jhome/34787. Panel data includes more information on teachers who have been teaching for longer periods of time, because we see the performance of multiple cohorts of students. As such, it can be helpful to address issues of noise, small sample biases (of the type discussed in section II), or unobserved student heterogeneity. However, collection of sufficient panel data for every teacher can be costly and delay feedback to teachers. Further, some of the estimator/model combinations we consider are not appropriate for use with panel data. Therefore, value-added based on cross-sectional data can be appealing for some policy uses.

value-added estimates in cases where there is little evidence of grouping and assignment based on prior achievement, but to diverge in cases where such evidence exists.

When teacher effects are treated as random, we use a mixed effects modeling approach estimated by Maximum Likelihood[23] to obtain empirical Bayes shrinkage estimates of teacher effects. These are labeled *EB Lag* and *EB Gain*; they are estimates of the Best Linear Unbiased Predictors (BLUP) of the teacher effects under appropriate assumptions (See Guarino, Reckase, & Wooldridge, in press; Ballou, Sanders, & Wright, 2004; and Guarino et al., 2014, for detailed discussions).

The EB approach used here is based on the following mixed effects model:

$$A_{ij} = \lambda A_{it-1} + X_{it}\beta + \mu_j + \varepsilon_{ij} \qquad (9)$$
$$\xi_{ij} = \mu_j + \varepsilon_{ij}$$
where i indexes students and j indexes teachers

In this set-up, the coefficients on the prior score (λ) and the student covariates (β) are treated as fixed, while the teacher effects (μ_j) are treated as random. Importantly, this loosely implies that teacher effects are assumed to be uncorrelated with the prior test scores and student covariates. In the mixed effects set up, the EB teacher effects estimates can be obtained by appropriately scaling an initial teacher effect estimate by a measure of reliability, specifically, $VA_{EB} = \bar{\xi}_J[\frac{\sigma_\mu^2}{\sigma_\mu^2+\frac{\sigma_\varepsilon^2}{n_j}}]$. Here, ($\bar{\xi}_J$) is the within-teacher mean student residual (inclusive of the teacher random effects), σ_μ^2 is the variance of teacher effects, σ_ε^2 is the student variance, and n_j is the number of student-level observations for teacher j. From here it is easy to see that the EB estimate shrinks an estimated teacher effect toward the mean (normalized to be zero) with noisier estimates based on fewer student observations shrunk more.

[23] In this setup, the coefficients in equation (9) can be estimated by Feasible Generalized Lest Squares (FGLS) or MLE. We opt for MLE using the *xtmixed* command in Stata with the BLUP random effect estimates easily obtained postestimation by the *predict, reffects* command.

A key difference between the OLS and EB approaches is that the OLS approach employed here includes indicators for each teacher, treating the teacher effects as fixed, rather than random as in the EB case. By leaving the teacher effects in the error term, EB approaches do not partial out the relationship between teacher assignment and the other included covariates, effectively assuming that this covariance is zero. The OLS approach adopted here does take this covariance into account when estimating both the teacher effects and the coefficients on the student covariates. In cases where teacher assignment is related to student covariates, we might expect this distinction between OLS and EB to become more important than when there is little evidence of such a relationship. However, the extent of these differences is an empirical matter.

Results Comparing Value-Added Estimation Approaches on Different Subsamples

Table 5 displays the VAM rank correlations across estimators within each sample using math test scores, while Table 6 shows the same for reading. For ease of reporting, the rank correlations are calculated pooling together all cross sectional value-added results (i.e., each teacher-grade-year accounts for one observation).[24] Starting in Panel A, the OLS Lag and OLS Gain estimates show a rank correlation in the Non-Grouping sample of 0.858 in math and 0.813 in reading. The rank correlation for the two OLS estimators drops noticeably to 0.754 when applied to the grouping sample in math. The difference for reading is even more stark, with a Grouping rank correlation of only 0.591. This closely matches our prediction that fixing $\lambda=1$ will be more important in cases where student grouping is related to prior student performance.

[Insert Table 5 Approximately Here]

Moving to the comparison between the OLS Lag and EB Lag estimates for the Non-Grouping sample, we see a very strong rank correlation of 0.982 in math and 0.963 in reading.

[24] Separate analysis by grade-year estimation sample yields very similar results and is available upon request.

The Grouping samples also show strong, albeit slightly smaller, rank correlations of 0.976 and 0.955. The small difference across samples and the overall strength of the rank correlations suggest, at least in this setting, that the decision to estimate by OLS or EB makes relatively little difference for ranking teachers when lagged test scores are included on the right-hand side. If instead of ranking teachers, we were interested in the relative magnitude of teacher effects, this distinction would become more pronounced. The other rank correlations across Panel A follow similarly, with the lag/gain distinction leading to empirically relevant differences in our teacher rankings.

In Panel B, we see a very similar story across our matching samples, however, the differences are less pronounced. While the comparison between the positive and negative matching subsamples generally goes in the direction predicted, the difference in rank correlations is rather small. The lack of a result here may be due to several reasons, including the fact that this is a small and select subsample of teachers with prior value-added or that there may be an orthogonal source of bias affecting both the prior value-added and the subsequent predictions by subsample. Regardless, the fact that the rank correlations are uniformly smaller in either matching scenario than the Non-Matching, indicates that we have uncovered a difference in assignment patterns that impacts value-added estimation.

Another way to check the robustness of teacher value-added estimates to nonrandom grouping and assignment is to consider how teachers would be classified into performance categories by different estimators. We thus divide teachers into quintiles based on their estimated math value-added. We then look to see how robust this grouping of teachers is to the use of alternative estimators across our samples. Figure 1 displays histograms that show how a teacher's designated quintile may differ across estimation approaches. For example, the first

histogram in the top panel of Figure 1 shows the distribution of teacher value-added quintiles using the OLS Gain estimates for all teachers who were in the 1st (lowest) quintile using the OLS Lag estimates for the grouping sample. The next histogram in the panel shows the distribution of quintiles based on the OLS Gain estimates for those in the 2nd quintile of the OLS Lag estimates for the same sample. The remaining panels follow similarly.

[Insert Figure 1 Approximately Here]

The histograms in Figure 1 tell a similar story to the rank correlations in Table 5 with stronger agreement among gain-score and lag-score estimates in the Non-Grouping sample than in the Grouping sample. For instance, nearly 74 percent of teachers placed in the highest quintile by the OLS Lag estimator are also in the top quintile by the OLS Gain estimator for the Non-Grouping sample. However, less than 64 percent in the top quintile by OLS Lag are also placed in the top quintile by OLS Gain when looking at the grouping sample. We also see that the probability of placing teachers in the same quintile by OLS Lag and EB Lag is slightly lower in the grouping than in the Non-Grouping sample (between 3 to 5 percentage points). This suggests that while the rank correlations presented above are relatively weakly affected by the choice of OLS versus EB estimation methods, there is some scope for this choice to affect the grouping of teachers into relative performance categories, a practice that is often suggested as a component of teacher evaluation and one that is sometimes used to allot rewards and sanctions.

CONCLUSION

In this paper, we have demonstrated the importance of methodological choices in estimating teacher performance using value-added models, uncovering a set of phenomena of high policy relevance in the current climate of educational reform aimed at accountability. We have developed and applied a careful approach to identifying evidence in large administrative data

sets of nonrandom assignment of students to teachers, documenting considerable differences across schools in the extent of this behavior and showing how to use this information to inform value-added estimation.

We find clear evidence that many schools do engage in student grouping based on prior academic performance, a fact that has been obscured by the more aggregated statistics used in the prior literature to identify such sorting. We find less evidence that schools commonly group students in classrooms based on other characteristics, conditional on prior achievement. Importantly, we see large variation in the extent of grouping when looking across school-grade-years. Further, we see some variation in the extent of this grouping across schools serving different student populations. For instance, schools with higher Limited English Proficiency student populations are more likely to be found to engage in test score grouping.

We also find evidence to suggest that some explicit student-teacher ability matching takes place, particularly for certain school-grade-years. The presence of matching represents a greater threat to the ability of VAMs to recover reliable effect estimates. Although we are limited in our ability to accurately pinpoint these instances and capture the full extent of ability matching, our results provide suggestive evidence that such matching does occur. Overall, our use of multinomial logit techniques represents a significant contribution to the effort to diagnose nonrandom grouping and assignment in nonexperimental contexts—an issue that must be grappled with in policy as well as research applications due to increased pressures to evaluate teachers according to their performance.

Importantly, we find that categorizing schools based on observed patterns of grouping and assignment can lead to substantial differences in the sensitivity of value-added estimates of teacher effectiveness to different estimation procedures. Namely, the manner in which the chosen

model takes prior student achievement into account, through a gain score or lag score specification, becomes more important in cases of student achievement grouping and assignment. In prior work using simulations (Guarino, Reckase, & Wooldridge, in press), OLS applied to a lag score specification that treats teacher effects as fixed was shown to be more adept at recovering true teacher effects across a number of different assignment scenarios. Here, our investigations have borne out predictions that this specification will produce estimates that diverge from those from the gain-score specification under circumstances in which nonrandom grouping and assignment based on prior scores is detectable. That the OLS Lag estimator controls for this potential confounder directly, reinforces the evidence that in many cases this estimator may be preferable to other popular estimators currently in use. At the very least, the motivation for using a gain-score formulation should be balanced against this clear and identifiable threat to validity in cases of explicit test-score grouping. This is particularly true in cases in which a single estimator of teacher effectiveness is relied upon (e.g., in many policy scenarios).[25]

Our results suggest caution when settling upon an estimation strategy that is to be universally applied across schools, and, in particular, in applying estimation strategies that rely on assumptions of persistent decay. Methods will matter and are of particular relevance in policy applications that assign rewards and sanctions to teachers based on value-added.

[25] Note that researchers comparing alternative estimators of education production functions as part of robustness checks should also consider our results in weighing the validity of each estimate.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. Journal of Labor Economics, 25, 95-135.

Ammermueller, A., & Pischke, J. (2009). Peer effects in European primary schools: Evidence from the Progress in International Reading Literacy Study. Journal of Labor Economics, 27, 315-348.

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

Cramer, J. (2007). Robustness of logit analysis: Unobserved heterogeneity and mis-specified disturbances. Oxford Bulletin of Economics and Statistics, 69, 545-555.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L.(2006). Teacher-student matching and the sssessment of teacher effectiveness. The Journal of Human Resources, 41, 778-820.

Conger, D. (2005). Within-school segregation in an urban school district. Educational Evaluation and Policy Analysis, 27, 225-244.

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

Feng, L. (2010). Hire today, gone tomorrow: New teacher classroom assignments and teacher mobility. Education Finance and Policy, 5, 278-316.

Gao, N. (2012). School incentives, principal characteristics and teacher assignment. Unpublished manuscript.

Goldhaber, D. (2008). Teachers matter, but effective teacher quality policies are elusive. In H.F. Ladd, E. B. & Fiske (Ed.), Handbook of research in education finance and policy (pp. 146-165). New York, NY: Routledge.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. Center for Education Data & Research,Working Paper 2012-7.2. Seattle, WA: Center for Education Data & Research.

Goldhaber, D., & Hansen, M. (2010). Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Center for Analysis of Longitudinal Data in Education Research, Working Paper 31. Washington, DC: Center for Analysis of Longitudinal Data in Education Research.

Goldhaber, D., & Hansen, M. (2012). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. Center for Analysis of Longitudinal Data in Education Research, Working Paper 73. Washington, DC: Center for Analysis of Longitudinal Data in Education Research.

Gould, W. (1996). Why does my mlogit take so long to converge? Retrieved August 1, 2013, from http://www.stata.com/support/faqs/statistics/convergence-of-maximum-likelihood-estimators/.

Grisson, J. A., Loeb, S., & Nakashima, N. A. (2014). Strategic involuntary teacher transfers and teacher performance: Examining equity and efficiency. Journal of Policy Analysis and Management, 33, 221-140.

Guarino, C. M., Ham, E. H., Reckase, M. D., Stacy, B., & Wooldridge, J. M. (2013). Sending value-added measures of teacher performance into tailspin: A simulation study of measurement error and nonrandom sorting. Draft presented at the American Economic Association, January, 2013, Philadelphia, PA.

Guarino, C. M., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2014). An evaluation of empirical Bayes' estimation of value-added teacher performance measures.

Working Paper #31. East Lansing, MI: The Education Policy Center at Michigan State University.

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

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

Hanushek, E. (1986). The economics of schooling: Production and efficiency in the public schools. Journal of Economic Literature, XXIV, 1141-1178.

Harris, D. (2009). Teacher value-added: Don't end the search before it starts. Journal of Policy Analysis and Management, 28, 693-699.

Harris, D., Sass, T., & Semykina, A. (2010). Value-added models and the measurement of teacher productivity. Unpublished manuscript.

Hill, H. (2009). Evaluating Value-added models: A validity argument approach. Journal of Policy Analysis and Management, 28, 700-709.

Jackson, C. K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. Journal of Labor Economics, 27, 213-256.

Kalogrides, D., Loeb, S., & Beteille, T. (2011). Power play? Teacher characteristics and class assignments. CALDER Working Paper No. 59. Washington, DC: Center for Analysis of Longitudinal Data in Education Research.

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

Kinsler, J. (2011). Beyond levels and growth: Estimating teacher value-added and its persistence. Unpublished manuscript.

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. Economics of Education Review, 28, 682-692.

Koedel, C., & Betts, J. (2010). Value added to what? How a ceiling in the testing instrument influences value-added estimation. Education Finance and Policy, 5, 54-81.

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

Koedel, C., Leatherman, R., & Parson, E. (2012). Test measurement error and inference from value-added models. The BE Journal of Economic Analysis & Policy, 12, 1-37.

Lavy, V. (2011). What makes an effective teacher? Quasi-experimental evidence. NBER Working Paper 16885. Cambridge, MA: National Bureau of Economic Research.

Lefgren, L., & Sims, D. (2012). Using subject test scores to efficiently predict teacher value-added. Educational Evaluation and Policy Analysis, 34, 109-121.

Lugo, M. (2011). Heterogeneous peer effects, segregation and academic achievement. Policy Research Working Paper 5718. The World Bank.

Mcfadden, D. (1974). The measurement of urban travel demand. Journal of Public Economics, 3, 303-328.

Oketch, M., Mutisya, M., Sagwe, J., Musyoka, P., & Ngware, M. (2012). The effect of active teaching and subject content coverage on student's achievement: Evidence from primary schools in Kenya. London Review of Education, 10, 19-33.

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

Subedi, B. R., Swan, B., & Hynes, M. (2011). Are school factors important for measuring teacher effectiveness? A multilevel technique to predict student gains through a value-added approach. Education Research International, 20111-10.

Todd, P., & Wolpin, K. (2003). On the specification and estimation of the production function for cognitive achievement. Economic Journal, 113, 3-33.

Whitmore, D. (2005). Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment. American Economic Review, 95, 199-203.

Winters, M. A., & Cowen, J. M. (2013). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. Journal of Policy Analysis and Management, 32, 634-654

Appendix A: Performance of Panel Data Value-Added Estimates

In the panel data context, we use four different model/estimator combinations. As in the cross-section case, we estimate value-added by OLS using both the Lag Score and Gain Score specifications (OLS Lag and OLS Gain). The panel context presents additional challenges and opportunities for estimating value-added. Namely, both OLS estimators ignore the presence of unobserved student heterogeneity. To address this possibility, the gain score specification can be easily estimated allowing for student fixed effects, yielding our *Fixed Effects* (FE Gain) estimator. The appeal of the FE Gain estimator comes at the cost of using the gain score specification. This is due to the strict exogeneity assumption needed for the consistency of FE that is violated when a lagged dependent variable is included on the right-hand side. Thus, like OLS Gain, it may lead to an omitted variables problem if teacher assignment is based on prior scores, here conditional on the student heterogeneity.

A final panel data estimator considered is the Empirical Bayes (EB) shrinkage estimate of teacher effects applied to the Gain score equation (EB Gain). Importantly, in the panel data context, the EB estimator requires a similar strict exogeneity assumption to FE Gain, once again precluding estimation of the lag score specification. Like the OLS Gain and Lag estimators, EB Gain does not allow for unobserved student heterogeneity to be correlated with inputs.

Many of the predictions outlined in the main text for the cross-sectional estimates apply here to the panel case. However, the introduction of the FE Gain estimates provides a distinct set of predictions. Differences in estimated value-added between OLS Lag and FE Gain will result from the appropriateness of the gain score specification, the importance of time-invariant unobserved student heterogeneity in the teacher assignment decision, potential violation of the strict exogeneity assumption, and increased noise due to the within-student demeaning. As such,

we might expect larger divergence between estimates for this comparison than others, regardless of the grouping and assignment scenario. In contrast, comparisons between OLS Gain and FE Gain will not depend on the appropriateness of the gain score specification as both estimators rely on the gain score assumptions. However, due to the other differences in assumptions, we expect ranking of teachers to generally diverge the most when comparing FE Gain to any of our other estimators.

Appendix Table A1 displays rank correlations between the panel data estimators within the different samples defined in the main text. As in the cross-sectional case, we see that the Gain/Lag decision holds more weight than the OLS/EB decision, with rank correlations diverging more when comparing an estimate from the gain score specification to one from the lag score specification. As predicted, the rank correlations with the FE Gain estimator tend to be relatively low overall yet slightly higher for OLS Gain than OLS Lag. Interestingly, the rank correlations are noticeably larger in the Non-Grouping and Non-Matching samples with particularly striking differences between matching and Non-Matching samples. The ranking of teachers in our matching samples are highly sensitive to the choice of estimating by OLS Lag or FE Gain with rank correlations under 0.25. Given the many reasons for these two estimators to diverge (outlined above), it is difficult derive simple recommendations other than to urge cautious interpretation of results and a careful choice of preferred estimator.

Table 1. Main analysis sample: School-grade-years with more than one class.

	Grade							
	G4		*G5*		*G6*		*All*	
Observations								
Students	950,915		949,849		883,423		2,784,187	
Teachers	48,947		46,800		14,718		110,465	
School-grade-years	11,139		10,984		4,054		26,177	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Student variables								
Prior test score	1400	276	1508	247	1651	222	1516	270
Asian	0.02	0.14	0.02	0.14	0.02	0.15	0.02	0.14
Black	0.23	0.42	0.22	0.42	0.23	0.42	0.23	0.42
Hispanic	0.23	0.42	0.23	0.42	0.23	0.42	0.23	0.42
Other race	0.04	0.19	0.03	0.18	0.03	0.17	0.03	0.18
Female	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Disability: High incidence	0.14	0.35	0.13	0.34	0.09	0.28	0.12	0.33
Disability: Low incidence	0.00	0.04	0.00	0.04	0.00	0.05	0.00	0.04
Free or reduced lunch	0.51	0.50	0.50	0.50	0.49	0.50	0.50	0.50
Limited English	0.05	0.21	0.04	0.19	0.04	0.18	0.04	0.20
Foreign born	0.08	0.27	0.09	0.28	0.10	0.30	0.09	0.28
Days absent prior year	7.17	6.91	7.19	6.95	7.42	7.39	7.25	7.08
New to school	0.95	0.21	0.34	0.48	0.94	0.24	0.74	0.44
Number of schools attended prior year	1.09	0.30	1.07	0.28	1.08	0.29	1.08	0.29
Teacher variables								
Asian	0.01	0.09	0.01	0.07	0.01	0.10	0.01	0.08
Black	0.13	0.33	0.15	0.36	0.19	0.39	0.15	0.35
Hispanic	0.10	0.30	0.08	0.28	0.08	0.28	0.09	0.29
Other race	0.00	0.05	0.00	0.06	0.00	0.05	0.00	0.06
Female	0.91	0.29	0.82	0.38	0.73	0.44	0.85	0.36
Years experience	9.77	9.82	11.29	10.66	10.24	10.34	10.48	10.28
Prior value-added	7.96	54.68	10.83	56.55	-10.43	55.88	6.98	56.04

Table 2. Predictors of classroom grouping: percentage of separate school-grade-year MNLs in which the predictor was significant at the 5-percent level.

	Grade					
	G4		*G5*		*G6*	
	Percentage	*# tests*	*Percentage*	*# tests*	*Percentage*	*# tests*
Specification 1	*Prior Score Only*					
Prior math score	24.52	11,137	25.35	10,981	67.34	4,054
Specification 2	*Prior Score with Other Covariates*					
Prior math score	20.62	11,110	21.41	10,927	63.23	4,030
Asian	0.29	5,828	0.31	6,200	1.38	3,041
Black	1.39	10,252	1.46	10,174	5.16	3,894
Hispanic	1.10	9,826	1.45	9,808	3.66	3,827
Other race	0.37	8,393	0.35	8,336	1.37	3,569
Female	1.54	11,095	1.63	10,914	7.65	4,025
Disabled—High incidence	4.66	10,915	5.30	10,734	17.91	3,965
Disabled—Low incidence	0.19	1,068	0.16	1,267	0.83	1,079
FRL	4.20	10,870	4.30	10,722	8.67	3,990
LEP	1.12	6,331	0.95	6,288	8.31	2,827
Foreign born	0.75	8,824	0.86	8,990	3.45	3,623
Prior year absences	3.98	11,103	3.99	10,919	7.03	4,028
Student in new school	6.18	8,156	3.76	9,172	11.89	2,574
Number of schools in year	2.94	10,717	7.03	10,378	6.44	3,976
Pseudo R^2 distribution						
5th percentile	0.11		0.11		0.05	
Median	0.21		0.21		0.16	
95th percentile	0.44		0.46		0.47	

Table 3. MNL rejection rates (percentage) for prior scores in specification 2 broken out by quartiles of school-level student characteristics.

School characteristic	Grade	Q1	Q2	Q3	Q4	All
Black	*4*	26.12	20.06	15.43	21.07	20.62
	5	25.58	20.50	17.15	22.67	21.41
	6	62.35	63.87	60.06	67.28	63.23
Hispanic	*4*	14.44	15.72	21.69	29.96	20.62
	5	14.92	15.83	23.39	30.76	21.41
	6	53.75	55.30	68.42	74.21	63.23
Disabled	*4*	22.20	21.82	21.38	17.95	20.62
	5	22.18	23.94	21.74	18.43	21.41
	6	67.79	67.04	56.56	49.20	63.23
FRL	*4*	22.69	17.32	17.21	25.30	20.62
	5	21.76	19.91	18.58	25.48	21.41
	6	60.94	63.54	63.03	66.35	63.23
LEP	*4*	13.31	19.48	24.04	25.31	20.62
	5	12.75	20.20	25.39	26.76	21.41
	6	51.18	60.04	67.46	73.30	63.23
Enrollment	*4*	9.89	14.00	22.69	37.15	20.62
	5	10.61	14.87	23.46	37.21	21.41
	6	37.60	48.31	61.94	71.46	63.23

Table 4. Five-percent significance-level rejection rates (percentage) of match variables from matching logit estimates.

		***MATCH* variable only**			**MATCH and other covariates**		
***MATCH* variable**	**Test**	**4th grade**	**5th grade**	**6th grade**	**4th grade**	**5th grade**	**6th grade**
VAM-score match	$\gamma \neq 0$	14.65	15.67	42.16	10.92	11.87	33.03
	$\gamma < 0$	7.52	7.44	15.92	6.41	6.79	13.16
	$\gamma > 0$	12.88	14.00	32.64	9.93	10.55	25.60
	#	5372	5915	1639	4743	5745	1535
Class-score match	$\gamma \neq 0$	33.00	32.84	63.78	21.71	22.99	50.99
	$\gamma < 0$	0.01	0.01	0.07	0.38	0.48	0.28
	$\gamma > 0$	41.71	41.74	68.89	29.42	29.90	56.04
	#	8269	8836	2681	7291	8544	2516
Exp-score match	$\gamma \neq 0$	13.94	15.26	44.47	10.23	11.93	36.15
	$\gamma < 0$	8.63	9.20	21.32	7.01	7.78	17.52
	$\gamma > 0$	11.59	12.10	29.33	8.82	10.06	24.68
	#	8049	8643	2584	7086	8390	2415
Racial match	$\gamma \neq 0$	9.83	9.39	17.92	9.69	9.48	9.37
	$\gamma < 0$	0.00	0.00	0.00	0.00	0.00	0.00
	$\gamma > 0$	10.03	10.37	17.34	9.17	9.57	9.53
	#	4507	4930	1384	3993	4798	1291

Note: The test column indicates the alternative hypothesis; therefore the first row indicates the percentage of times the test rejects the null of a zero coefficient, the second row of each panel indicates the percentage of times our results provide evidence of negative assignment, while the third row does so for positive assignment

Table 5. Cross-sectional math VAM rank correlations by grouping and matching samples.

Panel A: Grouping and Non-Grouping samples

Estimator/Model		*OLS Lag*		*OLS Gain*		*EB Lag*	
	Sample	*G*	*NG*	*G*	*NG*	*G*	*NG*
OLS Gain	*G*	0.754					
	NG		0.858				
EB Lag	*G*	0.976		0.752			
	NG		0.982		0.854		
EB Gain	*G*	0.737		0.969		0.776	
	NG		0.851		0.979		0.874

Panel B: Positive, Negative, and Non-matching samples.

Estimator/Model		*OLS Lag*			*OLS Gain*			*EB Lag*		
	Sample	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*
OLS Gain	*+M*	0.807								
	-M		0.813							
	NM			0.845						
EB Lag	*+M*	0.985			0.807					
	-M		0.984			0.812				
	NM			0.988			0.844			
EB Gain	*+M*	0.791			0.979			0.820		
	-M		0.801			0.982			0.825	
	NM			0.841			0.985			0.859

Sample sizes: G=50,812; NG=91,533; +M=8,483; -M=4,382; NM=44,614

Table 6. Cross-sectional reading VAM rank correlations by grouping and matching samples.

Panel A: Grouping and Non-Grouping samples

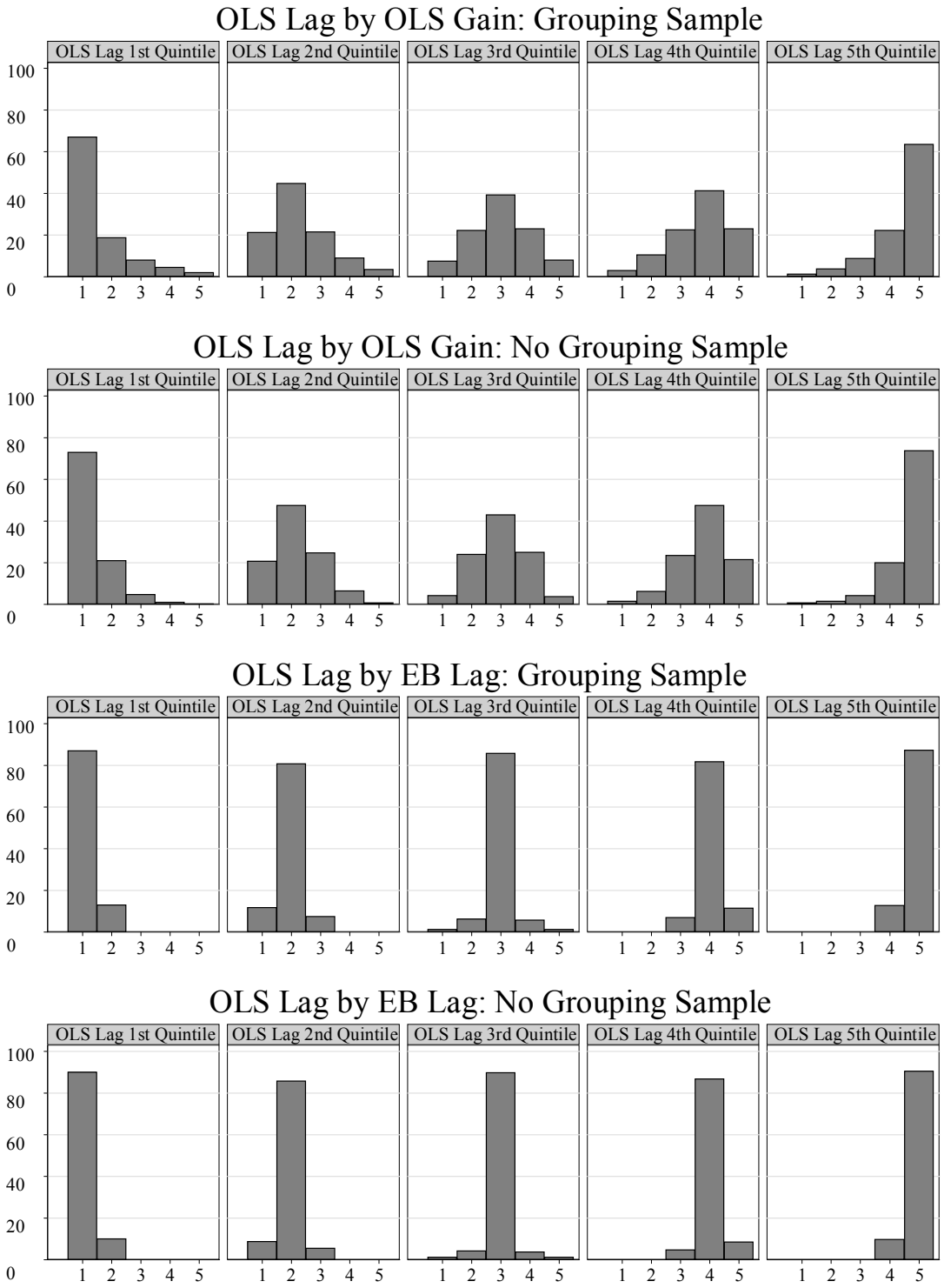
Estimator/Model		*OLS Lag*		*OLS Gain*		*EB Lag*	
	Sample	*G*	*NG*	*G*	*NG*	*G*	*NG*
OLS Gain	*G*	0.591					
	NG		0.813				
EB Lag	*G*	0.955		0.598			
	NG		0.963		0.803		
EB Gain	*G*	0.558		0.941		0.623	
	NG		0.792		0.957		0.834

Panel B: Positive, Negative, and Non-Matching samples

Estimator/Model		*OLS Lag*			*OLS Gain*			*EB Lag*		
	Sample	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*
OLS Gain	*+M*	0.702								
	-M		0.724							
	NM			0.811						
EB Lag	*+M*	0.974			0.720					
	-M		0.970			0.727				
	NM			0.969			0.806			
EB Gain	*+M*	0.677			0.969			0.727		
	-M		0.699			0.965			0.742	
	NM			0.790			0.968			0.828

Sample sizes: G=38,976; NG=78,573; +M=6,640; -M=3,954; NM=50,836

Figure 1, OLS Lag Quintile by OLS Gain and EB Lag Quintiles.



This is a preprint of an article published in the *Journal of Policy Analysis and Management* available at http://onlinelibrary.wiley.com/doi/10.1002/pam.21781/abstract

Appendix Tables

Appendix Table A1: Panel VAM rank correlations by grouping and matching samples.

Panel A: Grouping and Non-Grouping Samples from MNL results

Estimator/Model		*OLS Lag*		*OLS Gain*		*EB Gain*	
	Sample	*G*	*NG*	*G*	*NG*	*G*	*NG*
OLS Gain	*G*	0.805					
	NG		0.852				
EB Gain	*G*	0.777		0.966			
	NG		0.829		0.960		
FE Gain	*G*	0.517		0.573		0.578	
	NG		0.635		0.661		0.647

Panel B: Matching and Non-Matching samples from CL results.

Estimator/Model		*OLS Lag*			*OLS Gain*			*EB Gain*		
	Sample	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*	*+M*	*-M*	*NM*
OLS Gain	*+M*	0.858								
	-M		0.849							
	NM			0.854						
EB Gain	*+M*	0.766			0.913					
	-M		0.798			0.951				
	NM			0.831			0.968			
FE Gain	*+M*	0.232			0.271			0.266		
	-M		0.147			0.227			0.276	
	NM			0.567			0.596			0.580

Sample sizes: G=26,887; NG=36,421; +M=8,049; -M=4,300; NM=32,368