DISCUSSION PAPER SERIES

IZA DP No. 7112

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

Steven G. Dieterle
Cassandra M. Guarino
Mark D. Reckase
Jeffrey M. Wooldridge

December 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

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
Indiana University and IZA

Mark D. Reckase
Michigan State University

Jeffrey M. Wooldridge
Michigan State University and IZA

Discussion Paper No. 7112
December 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 7112
December 2012

ABSTRACT

How Do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-Added*

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, 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 this, 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.

JEL Classification: I0, I20, I21, I28, J01, J08, J24, J44, J45

Keywords: value added, teacher quality, teacher labor markets, education

Corresponding author:

Cassandra M. Guarino
Indiana University
School of Education
201 N. Rose Avenue
Bloomington, IN 47405
USA
E-mail: guarino@indiana.edu

* The authors would like to thank Doug Harris and AEFP session participants for helpful comments. 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.

I. 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 and has been widely contested by teachers' unions, however. A key concern in the debate 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 (Rothstein 2010; Kane & Staiger 2008; Aaronson, Barrow, & Sanders 2007, Guarino, Reckase, & Wooldridge 2012). Guarino et al. showed through simulations that the optimal method of computing teacher performance measures differed for different student-teacher assignment scenarios. Moreover, certain types of violations of random assignment are more amenable to statistical corrections than others. Knowing how students are grouped and assigned to teachers is therefore key to establishing confidence in the measures produced. Unfortunately, while the potential for nonrandom assignment to bias teacher VAMs has been well recognized, little research has investigated how principals assign students to teachers.

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 (Aaronson, Barrow, & Sander, 2007; Clotfelter, Ladd, & Vigdor 2006). Our study addresses the question more directly and in greater depth. Importantly, our approach acknowledges the fact that sorting behavior may differ across schools or even within schools over time and across grades. In addition, our study goes beyond the simple investigation of tracking patterns to investigate the matching of student and teacher ability and other characteristics. We distinguish between two

components of nonrandom assignment of students to teachers and examine evidence of both of these: students must be grouped together on the basis of some characteristic, and the groups must then be systematically assigned to teachers on the basis of some type of student-teacher match. Thus we provide a more comprehensive view of how students are assigned to teachers than can be gleaned by the broad tests found in the literature to date. These results can be useful to categorize schools based on the degree of sorting for states looking to fine tune the achievement portion of teacher evaluation or for researchers working with large administrative datasets. The approach adopted here is also informative for education production function studies of the effect of tracking on student achievement, by providing a measure of tracking to compare with or use in the absence of principal survey responses (Argys et al. 1996, Betts & Shkolnik 2000, Rees et al. 2000)

Using a large longitudinal data set from an anonymous state, 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 types of aggregate sorting indices developed in prior research. We also find evidence of nonrandom assignment of teachers to classes. In particular, teachers with higher measured effectiveness tend to be assigned to classrooms with higher average prior achievement. Finally, we investigate the sensitivity of different methods of estimating teacher value-added to different grouping and assignment scenarios. To do so, we combine the results of the above analyses to define subsamples of school-grade-years that exhibit different grouping and assignment behaviors and then examine correlations within subsamples among value-added teacher performance measures estimated in different ways. We find the sensitivity of teacher value-added to the particular estimator used differs by subsample. Importantly, these differences

align with predictions based on the standard value-added framework laid out in Guarino, Reckase, and Wooldridge (2012). For instance, value-added in subsets of schools that show evidence of test score grouping is more sensitive to the choice of model specification and estimator than in those that do not. This is because some approaches effectively account for the assignment mechanism in the estimation and others are subject to an omitted variables problem.

This paper is organized as follows. Section II provides a conceptual framework for thinking about the process by which principals assign students to teachers and discusses the implications of different assignment scenarios for value-added measures of teacher performance. Section III discusses the data used. Section IV discusses and replicates previous approaches to identifying nonrandom assignment in administrative data, highlighting the limitations of these approaches. Section V outlines our approach to detecting nonrandom grouping and assignment and presents the findings. Section VI shows how our results on the grouping and assignment decisions of schools can be used to inform value-added estimation. Section VII concludes.

II. Framework and Background

We begin with a basic conceptualization of value-added models and discuss how various forms of student sorting and teacher assignment mechanisms may alter measures of teacher performance based on these models. The theoretical motivation for value-added models of teacher performance typically rests on the specification of a structural "education production function," in which achievement at any grade 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 the education production function 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, 2012). The first estimating equation will be referred to here as a "lag score" specification due to the presence of prior achievement on the right hand side:

$$(2.1) \quad A_{it} = \tau_t + \lambda A_{i,t-1} + E_{it}\beta_0 + X_{it}\gamma_0 + c_i + u_{it}$$

where

A_{it} is student i's test score in time t
τ_t are year fixed effects
E_{it} are educational inputs (here teacher indicators)
X_{it} are student and family characteristics
c_i is an unobserved student heterogeneity term

Often, researchers use the gain in test scores as the dependent variable and omit prior test scores from the right-hand side of the model, effectively assuming that λ is equal to 1. We will refer to this as the "gain score" specification:

$$(2.2) \quad A_{it} - A_{i,t-1} = \tau_t + E_{it}\beta_0 + X_{it}\gamma_0 + c_i + (\lambda - 1) A_{i,t-1} + \upsilon_{it}$$

Note that we include the additional term, $(\lambda - 1) A_{i,t-1}$, on the right hand side of equation (2.2) in order to emphasize the fact that if $\lambda \neq 1$ the choice to use a gain score specification may lead to an omitted variables problem.

Generally speaking, our ability to consistently estimate the teacher value added coefficients (β_0) hinges on what our estimation method requires about the correlation between teacher assignments (captured by E_{it}) and the unobserved factors affecting achievement, u_{it} and c_i. Further, the gain-score specification in (2.2) shows that if we ignore the presensence of $(\lambda - 1) A_{i,t-1}$, estimators will suffer if teacher assignment is correlated with lagged achievement. Here, our concern lies with understanding how different student sorting and teacher assignment mechanisms employed by schools may affect the correlation between teacher assignment and

unobserved or omitted determinants of student achievement and, in turn, value-added estimates based on equations (2.1) and (2.2).

Throughout the paper, we distinguish how students are grouped together into classrooms from how teachers are assigned to those classrooms. In the simplest case, students may be randomly grouped into classrooms with no consideration given to the within classroom 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 (2.1) or (2.2) 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 based on, say, prior achievement, demographic characteristics related to ability, or markers of ability unobserved to those outside the school. But assume that the the schools still randomly assign teachers to these classrooms, i.e., teachers are assigned regardless of their ability to improve achievement. 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 classrooms. Grouping based on observable student demographic characteristics (captured in X_{it}) are of little concern for estimators that partial out this correlation as both equation (2.1) and (2.2) control for those factors. As a special case, grouping captured by prior test scores will not generally be problematic for the lag score specification (2.1). However, with very few classes per teacher such a grouping mechanism may lead to problems for estimates based on the gain score specification (2.2) if achievement gains in one year do not carry through completely to the next (i.e., $\lambda \neq 1$). This problem stems from not fully controlling for prior test scores (i.e. leaving $(\lambda - 1) A_{i,t-1}$) in the error term) and having some teachers assigned

the classes with better prior performing students creating a correlation in the sample between teacher assignment and student ability. With many classes per teacher and random assignment of teachers to classes, this small sample bias that arises from using the gain score specification becomes less important with teachers receiving a range of class types over time. Finally, grouping on unobservable determinants of achievement will also lead to small sample biases using either estimating equation that will similarly be alleviated by observing multiple classes for each teacher. This example highlights the importance of distinguishing between the student grouping and teacher assignment decisions. Importantly, not all deviations from a pure random grouping policy will necessarily lead to poor value-added estimates.

Finally, consider a case in which schools nonrandomly group students based on ability as before, however, now teachers are assigned to those classes in a systematic way according to each teacher's ability to raise achievement. Once more, grouping based on observable student demographic characteristics will not be problematic for estimators that 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 (2.2) regardless of the number of classes we observe for each teacher. Specifically, $(\lambda-1)A_{i,t-1}$ is non-zero, correlated with teacher assignment, and omitted from the model in this case. In contrast, by flexibly controlling for prior achievement (not restricting λ=1), estimates based on equation 2.1 are not subject to the same omitted variables bias. Therefore, the distinction between the two estimating equations becomes most important when students are grouped by prior achievement and then teachers are assigned to those classrooms based on ability. Effectively the cost of assuming λ=1 is higher in these cases.

Issues that arise from the use of unobserved determinants of achievement to group students in this scenario are, obviously, more difficult to characterize. In the case of unobserved time-invariant factors (captured in c_i), methods that aim to account for this, such as student fixed effects or instrumental variables, may be useful. However, such methods typically involve strong additional assumptions (either that λ=1 or that the errors in 2.1 are serially uncorrelated) and greatly reduce the identifying variation, leading to potentially poor performance when the underlying assumptions are violated (see Guarino et al. 2012 for simulation evidence). Importantly, prior test scores may serve as a decent proxy in these cases as it is a function of c_i, while still being robust to other assignment mechanisms particularly when basing estimation on the lag score specification. When the grouping decision is based on time varying unobserved factors, there is little that can be done to control for this. Once more, prior test scores may serve as a decent proxy for these factors.

While not ubiquitous in the literature, gain score formulations of the achievement regression are still used in recent work (for example, Jackson 2009, Koedel et al. 2012, Kinsler 2011, Lefgren & Sims 2012, Oketch et al. 2012, Subedi et al, 2011). The motivation for using the gain score rather than the lag score specification often varies. In may be done, in part, to address issues of serial correlation in the lag score equation (Jackson 2009), to help in addressing measurement issues with test scores (Koedel et al. 2012), or to take advantage of panel data estimators aimed at addressing efficiency (Hierarchical Linear Models, Feasible GLS, Empirical Bayes) or identification issues (Fixed Effects) that are potentially inconsistent with the presence of lagged dependent variables. While these issues 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. Guarino, Reckase, and

Wooldridge (2012) (GRW) demonstrate via simulations that the cost can be severe. Further, GRW find that the concerns of using estimators that include lagged achievement when such an approach is not theoretically justified are overblown. For example, controlling for a lagged test score is often effective even if unobserved student heterogeneity is present in the cumulative effects model, and even in some cases where teacher assignment is based on the heterogeneity.

Moving forward, the focus of this paper is to explore 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 value-added 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 perform poorly and others have a better chance of providing useful teacher value-added estimates. In these investigations, we also uncover descriptive information on how schools try to match students to teachers that may help inform research on organizational and power relations in schools (Kalogrides et al. 2011).

III. Data

The data used for this study come from the administrative records of a large and diverse state. The data tracks students and teachers in grades one through six in the state's public school system over an eight year period. 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 for our analyses. Typical of such large administrative data

sets, there is limited student information—primarily demographics (race/ethnicity, gender, disability status,[1] limited English proficiency, free or reduced lunch, country of birth), as well as information on school attendance/absences. In addition, the data include demographic (race/ethnicity and gender) and professional (certification status, degree level, and experience) variables for teachers. The set of student and teacher characteristics will allow us to examine the extent of sorting on observables in the state school system. Given the nature of the current study, additional data information will be provided as needed.

IV. 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. Here we review two such 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).

Aaronson, Barrow, and Sander (2007) (ABS) calculate the average within-class standard deviation of prior test score levels and gains 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. The highest scoring students are placed in the largest class followed by the next highest scoring students in the next largest class until each school-year-grade combination has the same number of

[1] 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/Beahvioral 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.

classrooms of the same size as in the actual data. The average of the within-class standard deviations for these counterfactual classrooms is then calculated within each grade and year. A second, "Random Sorting," counterfactual is created in a similar way by ordering students randomly before dividing them into classrooms. 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.

Table 1 displays the results of a replication of the ABS approach using our data to look for evidence of nonrandom sorting of students based on previous math test scores. For each grade and year combination, the average within-teacher standard deviation of previous test scores (both in levels and gains) are presented in the "Actual" column. Throughout we see that the actual standard deviations are closer to the random than the perfect, a result that accords with findings in ABS and others who have applied this exploratory measure to their data. This is particularly true for the lagged level scores. While this generally holds for the lagged gain scores as well, the range from perfect to random is much smaller, which makes for a less drastic comparison.

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 student's 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 (the expected random assignment distribution of

students is determined based on grade specific counts). CLV then categorize the 44.9% of schools that do not reject the null of random assignment in all six cases as non-tracking.

Table 2 summarizes a replication of the CLV approach to uncovering evidence of nonrandom sorting of students within schools using our data. The administrative data allow us to run the chi-squared tests for five of the six characteristics considered by CLV (all except parental education). With access to several years of data, we modify the CLV approach by pooling across all grades and years for each school, rather than simply pooling across grades in a single year. In presenting the results, we limit the sample to those schools for which all five tests were possible. We find that 53.69% of the included schools do not reject the null of independence across classrooms for all five characteristics. In the language of CLV, these schools are said to fail none of the five tests. This is of similar magnitude to the 44.9% of schools in North Carolina that failed none of the tests in CLV's analysis. Importantly, this test suggests that there may be substantial across school variation in the extent of student tracking on observables.

The above approaches to identifying evidence of the nonrandom sorting of students into classrooms provide either aggregate statistics (ABS) or school-level analysis (CLV). Also, in the case of ABS, the test focuses on a single student characteristic, prior test performance, while not exploring other observable characteristics that may drive the student grouping decision. While the CLV approach considers other characteristics, each is tested independently without considering the potential relationships between different characteristics. For instance, the CLV approach may 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.

V. Investigation of Student Grouping and Teacher Assignment

We now outline our approach to assessing the extent of nonrandom student grouping and teacher assignment. First we investigate how students are grouped into classrooms. Next we investigate the characteristics of schools that engage in nonrandom grouping. Following that, we investigate whether teachers are nonrandomly assigned to classrooms.

Nonrandom grouping of students into classrooms

We begin by estimating a series of multinomial logit (MNL) models of student assignment to classrooms separately for each school-grade-year combination. We are effectively modeling the probability a student is assigned to a particular teacher given their characteristics, P(T=j|**x**), where j=1, 2,..,J indexes the teachers in that school-grade-year cell. 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. Here, 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 of the student characteristics on the probability a student has a particular teacher, denoted $\frac{\partial \mathrm{P}(T = j \mid \mathbf{x})}{\partial x_k}$. Therefore, for each MNL we estimate, we test that null that the partial effect for a given characteristic, x_k, is zero for all teachers:

$$(5.1) \qquad \mathrm{H}_0 : \frac{\partial \mathrm{P}(T = 1 \mid \mathbf{x})}{\partial x_k} = \frac{\partial \mathrm{P}(T = 2 \mid \mathbf{x})}{\partial x_k} = \ldots = \frac{\partial \mathrm{P}(T = J \mid \mathbf{x})}{\partial x_k} = 0$$

A MNL is estimated for every possible school-grade-year combination with a few restrictions. First, cases in which a school had only one teacher in a grade in a year are obviously dropped. As seen in Table 3, this drops 2,143 of the 28,320 possible school-grade-

year cells. Also, we limit our analysis to cases in which the MNL likelihood function maximization converged within 300 iterations.[2] Table 3 shows the number of potential MNL estimates (school-grade-year cells with more than one teacher), the number that converged when only the student's lagged test score was included as an explanatory variable, and the number that converged when all our student level covariates were included.

This procedure gives a large number of results (up to 26,177) that need to be summarized. We opt to show the percentage of times a particular characteristic was found to be statistically significant (rejecting the null in (5.1)).[3] 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.

Table 4 shows the percentage of times each student characteristic was found to be statistically significant at the 5% level in the MNL estimates separately by grade. The table also displays the number of times the hypothesis in (5.1) was tested for a given variable.[4]

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 (ABS and CLV). The significance rates for these MNL estimates are found in the first row of Table 4. Here we see that roughly 25% of the school-grade-year cells show evidence of grouping based on prior achievement in both fourth

[2] 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.
[3] While a measure of the relative magnitude of partial effects across schools would certainly be interesting, operationalizing this would be difficult in this setting. If for instance all school-grade-years had only two teachers, the absolute value of the estimated partial effect could be compared across cases as a measure of the relative strength of grouping behavior. However, with more than two teachers (and varying number of teachers across cells) there are multiple partial effects to compare both within and across school-grade-years. 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 more sensitive to model and estimator assumptions.
[4] Note that although the number of school-grade-years for which convergence was achieved for particular models was presented in Table 3, the number of times a particular hypothesis test was run may be less than what was represented in Table 3; for example, if there were no Asian students in the school, then that particular hypothesis test could not be run.

and fifth grade. In sixth grade, this percentage is much higher at 67%. This is perhaps not surprising, as in the state studied here many students make a promotional school change in grade six. If the administration in the new school has less private information on the student's ability, we might expect them to use prior achievement (something readily available on transcripts) 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 heterogeneous classes. Recall that some grouping based on time-constant unobserved student heterogeneity may be captured here as the prior achievement is a proxy for this unobserved component. This will be particularly true when the unobserved student component is relatively large or the year-to-year persistence of measured learning is stronger.

Moving down the table, we present rejection rates from MNL estimates including all our student level covariates. Among these student characteristics, only the lagged test score shows evidence of being predictive of teacher assignment with a substantial degree of frequency, although some variables such as high incidence disability show non-negligible frequencies. While the rejection rates for prior scores fall slightly compared to 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 (5.1) for the student's prior test score. To do so, we further disaggregate the rejection rates in Table 4 across quartiles of school-level student characteristics. Table 5 presents these results using the 5% rejection rates for the prior test score from the estimates of MNL models that included other student covariates. The table also reports the overall rejection

rates for each grade (identical to the values in Table 4 for the prior test score). Here 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 a story that decision makers in larger schools have less specific knowledge of each student and must base grouping decisions on easily observable predictors of performance. 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. Moving on, 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.

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 Section II, such nonrandom grouping does not, in and of itself, lead to problems with value-added estimation. Therefore, it is important to explore whether teachers are nonrandomly assigned to these groups of students. 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. To begin to explore this question, we estimate a series of regressions of a particular teacher characteristic on the average characteristics of the students in that teacher's classroom. This approach is similar to that of Kalogrides et al. (2011), however, those authors regress the characteristics of the classrooms on a set of teacher characteristics (effectively "flipping" the dependent and independent variables).

Our approach looks to answer the question of whether classrooms with observably different groups of students are more or less likely to be assigned teachers exhibiting a particular characteristic conditional on the other observables of the class, rather than whether a class exhibiting a particular average student characteristic is assigned a teacher with particular observables conditional on the other characteristics of that teacher. These regressions take the following form:

$$Y_{jgst} = X_{jgst}\beta_1 + \alpha_t + \delta_s + \eta_g + \varepsilon_{jgst} \quad (5.2)$$

where Y_{jgst} is one of 8 teacher characteristics

X_{jgst} is a vector of classroom average student characteristics

α_t are year fixed effects

δ_s are school fixed effects

η_g are grade fixed effects

t indexes year, j indexes teacher, s indexes school, and g indexes grade

The separate teacher characteristics that we consider include teacher value-added estimated by pooled OLS on the lag score specification (2.1) (see section 6 for more on this estimator) using data from all prior years for each teacher, the teachers experience, and indicators for whether the teacher is female, Asian, Black, Hispanic, fully certified, or has an advanced degree, Note that, with the exception of the value-added and experience regressions, the estimates are therefore from linear probability models.

Table 6 displays the results from these regressions. Starting with column 1, we see that there is a statistically significant relationship between the average prior score of a class and the prior value-added of the teacher assigned to the class. To interpret, the point estimate of 0.063 suggests that classes with average prior student performance one standard deviation (within that school-grade-year cell) better are assigned, on average, to teachers with value-added that is 0.063

test score standard deviations (within the state-grade-year cell) better. This is over a quarter of the value-added standard deviation (0.236). In unreported estimates that flip the student and teacher characteristics to reflect the Kalogrides et al. approach, we find statistically significant relationships between teacher value-added and class characteristics for the following characteristics: Asian, Black, Hispanic, Disabled- High Incidence, Disabled- Low Incidence, FRL, LEP, Prior Absences, and New to School. Many of these estimated relationships were quite small, however. In the approach presented in Table 6, we only see statistically significant relationships for FRL, prior absences, and new to school suggesting that some of the perceived assignment based on several classroom characteristics when regressing classroom characteristics on a set of teacher variables may actually be driven by other, related, characteristics of the classes.

Moving on to the other teacher characteristics, we see some evidence of Black and Hispanic student-teacher racial matching and a tendency for gender matches. We also see classrooms with a one standard deviation increase in average prior scores being assigned to teachers with nearly one more year of experience on average. Among other results, classrooms with students who, on average, had more absences in the prior year and have more Hispanic, Other Race, FRL, and LEP students also receive teachers with less experience. Again, these regressions include school fixed effects, so this reflects within-school experience differences rather than differences across schools serving more or less able students.

This descriptive approach, while informative, is done at a high level of aggregation. Consistent with the evidence above that schools (and even grades within schools) differ on the extent of student grouping; it is also plausible that different assignment mechanisms may be used in different school-grade-year cells. We therefore conduct a more direct and fine-grained

analysis. 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. Recall that the grouping of students into classrooms based on either observable, as detected by the previous MNL-based analysis, or unobservable predictors of achievement does not, in and of itself, lead to inconsistent value-added estimates. When this sort of grouping is accompanied by the systematic matching of teachers of different ability to these students, however, the consistency of value-added is threatened. It is important to recall, however, that with a small number of classes per teacher, even grouping alone could potentially cause problems for credible value-added estimates.

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 and estimate what is sometimes referred to as a conditional logit[5] for each school-grade-year cell. Following McFadden (1974), this can be derived from an underlying maximization problem across the different choices. Here, we can think of the school or principal choosing a teacher, j, in order to maximize the unobserved y_{ij}^{*} for each student, i:[6]

[5] In some cases this is also called a multinomial logit, with the understanding that the MNL described earlier is a special case of the conditional logit. In order to distinguish between the two approaches, we will refer to the current model as a conditional logit. This nomenclature follows from Wooldridge (2010).

[6] Note that this formulation effectively treats each student teacher match decision as independent and relying solely on characteristics of the student and teacher that make up a potential match. A much more complicated model may allow for a comparison of student-teacher-peer matches that would typically not be tractable in practice. As an extreme example, consider a case with 40 students split evenly between two teachers. Each potential match consists of matching one student with a teacher and 19 potential classmates, giving a unique choice set for each student consisting of $2\times\binom{39}{19}=137,846,528,820$ possible student-teacher-peer matches.

$$y_{ij}^{*} = MATCH_{ij}\gamma + STU_i\delta_j + u_{ij}$$

where

(5.3) $MATCH$ is one of four student-teacher match specific variables

STU is a vector of student characteristics

u_{ij} has a Type I Exterme Value distribution

The resulting estimate of γ gives an indication of the preferences for that particular match characteristic. For instance, a positive estimate suggests that the school values that match characteristic when assigning students to teachers, a negative estimate suggests the school looks to avoid such matches, and an estimate close to zero suggests it is not concerned with that particular match characteristic. The estimates of δ are analogous to those from the MNL.[7]

In practice we estimate four separate models each with a different match specific variable aimed at capturing some aspect of the student-teacher match. The first is an indicator for whether a potential student-teacher match represents a racial match. Next we consider whether more experienced teachers receive higher performing students by using a match variable that equals one if a potential student teacher match consists of a teacher who has above average experience among all teachers in that school-grade-year cell and a student with above average prior performance in the cell or both are below average, and equal to zero otherwise. Finally we look at two indicators of ability matching. The first uses the same OLS estimate of prior teacher value-added based on the lag score specification used above as a measure of teacher ability. We use value-added estimated using all prior years of data we have for the teachers. For example, for a conditional logit estimated using teacher assignments in 2005 we use any available data for a teacher from 2001-2004 to first estimate value-added. Then we create a variable indicating whether a given teacher is above average in prior value-added compared to all other teachers in

[7] Note the *j* subscript on δ indicating a separate estimate for each potential teacher in a school-grade-year, as in the MNL case, whereas, in the case of the matching variable, a single γ is estimated in each school-grade-year case.

that school-grade-year cell. We then create an indicator for whether the student was above average among all students in that school-grade-year combination. The *MATCH* variable is then set equal to one if the student and teacher are both above average or if they are both below average and zero otherwise. 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 this 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. This second ability match variable is created in a similar manner using the teacher's prior incoming class average of student scores, rather than value-added. For example, a teacher teaching fifth grade in 2005 will be coded as having an above average prior incoming class if in 2004 the fourth grade score of their incoming fifth grade students (exams taken in 2003) is above average among all fifth grade teachers in that school in 2004. Therefore, the fourth *MATCH* variable is equal to one if the student's prior score is above average among his or her peers and the teacher's prior incoming class was above average or if both were below average.

Two conditional logits are estimated separately for each *MATCH* variable, one with only the *MATCH* variable and one with a set of student specific variables.[8] In each case, we exclude

[8] 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

the 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 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. As with the MNL results, we also display the total number of hypothesis tests.

The first panel of Table 7 displays the results for the racial match variable. We see that when including the racial match variable only, nearly10% of cases show some evidence of matching based on this characteristic for fourth and fifth grade and nearly 18% for sixth grade. The inclusion of the student covariates (again excluding student race indicators) 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%. 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% for each grade and specification.

From the teacher experience/student test score match, we see that in 14% and 15% of fourth and fifth grade cells there is evidence of matching based on this characterization. However, in sixth grade, nearly half of all cells do reject the null. This would seem to suggest that many middle schools assign more experienced teachers to classrooms of better prior performing students, at least initially (recall that for many schools sixth grade is the youngest grade in the school). Importantly, adding other student characteristics (excluding prior test scores) reduces the rejection rate to 36%. Here, we also see that some schools show evidence of negative matching (high experience with low performers).

---

maximization scheme as for the MNL, allowing for 300 iterations alternating between three maximization algorithms.

Moving down the table, 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% and 16% 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%. We find statistically significant negative assignment between 7% and 16% of cases, with the highest rejection rate in grade six. There is evidence that positive assignment is much more common among the school-grade-year cells tested with rejection rates of roughly 13% in fourth and fifth grade and 33% in sixth grade. When including the set of student covariates (here excluding prior test scores), 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.

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 fourth grade school-year cells, 11 to 16 percent of fifth, and 33 to 42 percent of sixth grade school-year cells. Of course, it should be noted that in this many runs 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. However, it is also possible that these findings are understated if principals know more about teachers' true ability than is captured in our value-added measure. Or, principals could be relying on a less robust estimate of teacher value-added to make the matching decision, thus their intention to engage in ability matching may not be fully captured here.

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 7). This is perhaps not surprising, as we have noted that this measure will likely capture both cases in which there is explicit ability matching between students and teachers and any sort of persistent assignment of particular teachers to high or low performing students. Importantly, the rejection rates follow a similar pattern to the value-added 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, sixth grade teachers who had high ability classes in the past year were likely to get high ability students again.

It is worth noting the lower convergence rates for the CL estimation than for the analogous MNL runs. 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 and 4,743 did so in the racial match and VAM-Score conditional logit estimation with student covariates. This represents a dramatic drop in the number of results estimated and serves as a limitation of this approach when applied uniformly to a large number of schools. However, for the school-grade-cells in which estimation was possible, this approach does provide useful and interesting information related to 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.

VI. 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 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 and discuss how they should be expected to perform in random versus nonrandom grouping and assignment scenarios. The set of estimators was chosen to represent approaches in fairly common use, while maintaining a manageable number of comparisons. Therefore, we do not replicate every approach found in policy and research, but focus on a select few that are in use and allow us to highlight violations of key assumptions related to the tracking and assignment scenarios studied here (See, for example, Wright et al. 2010, Value-added Research Center 2010, Buddin 2011 for policy applications).

Under random grouping and assignment, these estimators can be expected to show more agreement in their rank ordering of teachers by effectiveness than under nonrandom grouping and assignment—a prediction based on the simulation findings in Guarino, Reckase, and Wooldridge (2012). To test our predictions, we estimate teacher value-added in mathematics using subsets of our administrative data that are determined by the degree to which nonrandom grouping and assignment are present, and we display rank correlations within each subsample among the estimates produced by the different estimators.

The subsamples are defined using the results of our MNL and CL analyses. 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% level (the "Grouping" subsample) and those that did not ("Non-Grouping").[9] The labels Grouping and Non-Grouping were chosen to emphasize that the MNL results tell us nothing about the subsequent assignment of teachers to these classrooms. To address the potential assignment decisions, we similarly divide our sample of school-grade-years into "Matching" and "Non-Matching" subsamples based on the teacher VAM/student score match CLs that included additional student covariates. While the Matching/Non-Matching distinction more closely reflects the type of grouping and assignment mechanism we are concerned with, there are advantages to using the MNL results as well. Namely, with higher rates of convergence, the MNL based subsamples give better empirical coverage while still reflecting grouping scenarios that may lead to problems in identification. In addition, the MNL results do not rely on prior VAM estimates. 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 student level observations and employ four separate estimation approaches involving the two estimating equations discussed in section II. We also estimate teacher value-added using panel data, and those results—which do not yield qualitatively different conclusions—are presented in the appendix.[10] The main features of estimation that we vary are the lag score versus the gain score

[9] 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.

[10] Both cross-sections and panels may be applied in evaluation policies. 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

specifications and the treatment of the teacher effects as fixed or random. The specifications with fixed teacher effects are estimated by Ordinary Least Squares (OLS) and include teacher indicator variables and retain their coefficients as our teacher effects—directly estimating the teacher effects from equation (2.1) for the lag score specification and from equation (2.2) for the gain score equation- yielding our OLS Lag and OLS Gain estimators. When teacher effects are treated as random, we use a mixed effects modeling approach estimated by Maximum Likelihood[11] 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 et al. 2012 and Ballou et al. 2004 for detailed discussions).

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

$$(6.1) \quad A_{ij} = \delta A_{it-1} + X_{it}\theta + \mu_j + \varepsilon_{ij}$$
$$\xi_{ij} = \mu_j + \varepsilon_{ij}$$

where i and j indexes teachers.

In this set-up, the coefficients on the prior score (δ) and the student covariates (θ) 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, $VA_{EB} = \bar{\xi}_j[\frac{\hat{\sigma}_\mu^2}{\hat{\sigma}_\mu^2+\frac{\hat{\sigma}_\varepsilon^2}{n_j}}]$. Here, ($\bar{\xi}_j$) is the within teacher mean student residual (inclusive of the teacher random effects), $\hat{\sigma}_\mu^2$ is an estimate of the variance of teacher effects, $\hat{\sigma}_\varepsilon^2$ is an estimate of the student

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.

[11] In this setup, the coefficients in equation (6.1) below 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..

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 teacher effect (normalized to be zero) with noisier estimates, those based on fewer student observations, shrunk more.

A key difference between these OLS and EB approaches is that the OLS approach employed here explicitly 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 such a relationship. For instance, in the lag-score specification when prior test scores are predictive of classroom grouping, we may see differences in how our EB and OLS estimators rank teachers. Since we found little consistent evidence of student grouping based on other student characteristics, a priori, we do not expect to see large differences between the rankings produced by the OLS Gain and EB Gain estimates. However, the extent of these differences is an empirical matter.

Perhaps more important will be the distinction between the lag-score and gain-score specifications given the fact that we uncovered evidence that student grouping, and in some cases explicit student teacher ability matching, based on prior scores occurs in our sample. As discussed previously, student grouping and nonrandom teacher assignment based on prior test scores will tend to create problems in the OLS Gain estimates when $\lambda \neq 1$ (due to omitting a

portion of prior achievement correlated with teacher assignment) but not the OLS Lag estimates. As such, we expect the two approaches to yield similar value-added estimates in cases when there is little evidence of grouping and assignment based on prior achievement. The main difference between OLS Gain and OLS Lag is the choice of specification. In contrast, the EB estimator that uses the gain score equation imposes additional assumptions. The comparison between OLS Gain and OLS Lag, therefore, allows a simpler analysis of the importance of the assumption that $\lambda = 1$ in particular contexts. [12]

Results comparing value-added estimation approaches on different subsamples

Table 8 displays the value-added rank correlations across estimators within each sample for both our Grouping/Nongrouping (Panel A) and Matching /Nonmatching (Panel B) samples. 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).[13] Separate analysis by grade-year estimation sample yields very similar results and is available upon request.

Starting in Panel A, we see a very strong rank correlation between the OLS Lag and EB Lag estimates for the nongrouping sample of 0.982. The grouping sample also shows a strong, albeit slightly smaller, rank correlation of 0.976. That the rank correlation is smaller in the grouping sample accords with our prediction that treating teacher effects as random versus fixed will matter more in the grouping case. However, 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 little difference for ranking teachers.[14]

[12] In Appendix A, we consider several panel data estimators. Note that with the cross-sectional data we cannot address the possibility of unobserved student heterogeneity (c_i in equations 2.1 and 2.2) and we limit ourselves to one classroom of students per teacher leading to noisier estimates of performance.
[13] See Appendix A for the panel data results.
[14] If instead of ranking teachers, we were interested in the relative magnitude of teacher effects, this distinction would become more pronounced.

Moving to the comparison between the OLS Lag and OLS Gain estimates, we see a weaker relationship between these two estimators in the nongrouping sample than for the two lag-score estimators with a rank correlation of 0.858. The rank correlation for the two OLS estimators drops noticeably to 0.754 when applied to the grouping sample. This closely matches our prediction that fixing $\lambda=1$ will be more costly in cases where teacher assignment is related to prior student performance, assuming $\lambda\neq1$. The other rank correlations across Panel A follow similarly, with the lag/gain distinction seeming to matter more than OLS/EB one. In Panel B, we see a very similar story across our matching and nonmatching samples.

Another way to check the robustness of teacher value-added estimates to the different estimators on different samples is to consider how teachers would be grouped into performance categories under the different grouping and assignment regimes. Here, we divide teachers into quintiles based on their estimated 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 (labeled OLS Lag 1st Quintile under OLS Lag by OLS Gain: Grouping Sample) 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.

The histograms in Figure 1 tell a similar story to the rank correlations in Table 8 with stronger agreement among gain-score and lag-score estimates in the nongrouping sample than in the grouping sample. For instance, nearly 80% 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 nongrouping sample. However, closer to 60% 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 nongrouping sample. This suggests that while the rank correlations presented above are only 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.

VII. Conclusion

In this paper, 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. An important, yet subtle, distinction made throughout is between the nonrandom grouping of students to classrooms and the nonrandom assignment of teachers to these groups.

We find evidence that many schools do engage in student grouping based on prior academic performance. 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, a fact that has been obscured by the more aggregated statistics used in the prior literature to identify such sorting in the context of value-added estimation. 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 such tracking.

We also find evidence of explicit student-teacher ability matching for some school-grade-years. The presence of matching represents a greater threat to the ability of value-added measures to recover true teacher effects than grouping alone. Although we are limited in our ability to accurately pinpoint these instances and capture the full extent of ability matching, our conditional logits 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 the observed patterns of grouping and assignment leads 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 controls for prior student achievement, through a gain score or lag score specification, becomes more important in cases of student achievement grouping and assignment. In our prior work using simulations (Guarino, Reckase, and Wooldridge, 2012), OLS applied to the 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 using actual data have borne out predictions that approaches that do not use lag score specifications or treat teacher effects as fixed will diverge from those of the OLS-Lag estimator under circumstances in which nonrandom grouping and assignment based on prior scores is detectable. That the OLS Lag estimator controls for this potential confounder flexibly

(unconstrained λ) and directly (treating teacher effects as fixed), reinforces the evidence that in many cases this estimator is preferable to other popular estimators currently in use. While further "selection on unobservables" is obviously still possible in any nonexperimental setting, it seems particularly problematic to choose specifications and estimators that fail to fully control for teacher assignment based on a readily observable characteristic that is shown in this paper to be related to student grouping and teacher assignment decisions for a nontrivial number of schools. This is particularly true in cases in which a single estimator of teacher effectiveness is required (i.e., in many policy scenarios),[15] where there is little to be gained by adopting a different strategy, or there is little additional evidence to suggest that other factors may be more important for assignment decisions than prior test scores thereby justifying an alternative approach (i.e., in a small scale study with particular information on the assignment decisions gathered for a particular school or district). 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 and random teacher effects.

[15] 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*(1), 95-135.

Argys, L. M., Rees, D. I., & Brewer, D. J. (1996). Detracking America's Schools: Equity at Zero Cost? *Journal of Policy Analysis and Management, 15*(4), 623-645.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-added Assessment of Teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Betts, J. R. & Shkolnik, J. L. (2000). The Effects of Ability Grouping on Achievement and Resource Allocation in Secondary Schools. *Economics of Education Review, 19*, 1-15.

Buddin, R. (2011). Measuring Teacher and School Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools. Unpublished Draft.

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L.(2006). Teacher-Student Matching and the Assessment of Teacher Effectiveness. *The Journal of Human Resources, 41*(4), 778-820.

Conger, D.(2005). Within-School Segregation in an Urban School District. *Educational Evaluation and Policy Analysis, 27*(3), 225-244.

Guarino, C., Maxfield, M., Reckase, M., Thompson, P., & Wooldridge, J. (2012). An Evaluation of Empirical Bayes' Estimation of Value-added Teacher Performance Measures under Nonrandom Teacher Assignment. Unpublished Draft.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2011). Evaluating Value-added Methods for Estimating Teacher Effects. Unpublished Draft.

Hanushek, E. (1979). Conceptual and Empirical Issues in the Estimation of Educational Production Functions. *The Journal of Human Resources, 14*(3), 351-388.

Hanushek, E. (1986). The Economics of Schooling: Production and Efficiency in the Public Schools. *Journal of Economic Literature, XXIV* (3), 1141-78.

Harris, D., Sass, T., & Semykina (2010). Value-Added Models and the Measurement of Teacher Productivity. Unpublished Draft.

Jackson, C. K. (2009). Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation. *Journal of Labor Economics, 27*(2), 213-256.

Kalogrides, D., Loeb, S., & Beteille, T. (2011). Power Play? Teacher Characteristics and Class Assignments. CALDER Working Paper No. 59.

Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. Working Paper 14607, National Bureau of Economic Research.

Kinsler, J. (2011). Beyond Levels and Growth: Estimating Teacher Value-added and its Persistence. Unpublished Draft.

Koedel, C., Leatherman, R. & Parsons, E. (2012). Test Measurement Error and Inference from Value-added Models. Unpublished Draft.

Lefgren, L. & Sims, D. (2012). Using Subject Test Scores to Efficiently Predict Teacher Value-added. *Educational Evaluation and Policy Analysis,34*(1), 109-121.

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*(1), 19-33.

Rees, D.I., Brewer, D. J., & Argys, L. M. (2000). How Should We Measure the Effect of Ability Grouping on Student Performance? *Economics of Education Review, 19*, 17-20.

Rothstein. J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics, 125*(1), 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,* 1-10.

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

Value-aded Research Center (2010). NYC Teacher Data Initiative: Technical Report on the NYC Value-added Model. Wisconsin Center for Education Research.

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

Wright S.P., White, J. T., Sanders, W. L., & Rivers, J. C. (2010). SAS EVASS Statistical Models. SAS White Paper.

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 nongrouping and nonmatching samples with particularly striking differences between matching and nonmatching samples. The ranking of teachers in our matching sample is highly sensitive to the choice of estimating by OLS Lag or FE Gain with a rank correlation 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.

Tables

Table 1: Average Within-Class Prior Math Test Score Standard Deviations Across Sorting Mechanisms (ABS Replication)

		Sorting Mechanism					
		Actual		Perfect		Random	
Year	**Grade**	*Lag SD*	*Lagged Gain SD*	*Lag SD*	*Lagged Gain SD*	*Lag SD*	*Lagged Gain SD*
2001	*4*	232.69		85.11		243.22	
	5	223.29		85.27		232.91	
	6	215.16		111.94		233.39	
2002	*4*	246.55	177.36	89.74	165.95	256.62	179.21
	5	217.61	157.91	85.75	153.57	226.49	158.26
	6	194.31	151.47	101.81	148.76	213.20	152.28
2003	*4*	231.75	196.01	90.20	187.45	240.52	199.17
	5	215.26	166.92	85.11	164.04	223.99	167.81
	6	191.47	146.77	102.40	143.46	212.12	147.63
2004	*4*	224.79	190.92	79.85	185.30	236.04	191.52
	5	203.07	157.61	82.70	155.65	212.65	158.51
	6	188.01	145.90	96.98	142.08	209.71	147.22
2005	*4*	232.59	185.15	82.51	174.22	245.57	185.97
	5	196.93	152.87	73.71	150.76	207.85	153.38
	6	170.28	136.04	89.75	134.73	190.79	137.02
2006	*4*	227.66	181.47	78.32	176.78	241.05	181.76
	5	205.82	160.07	77.97	157.65	218.36	161.02
	6	170.58	131.69	89.58	129.62	194.22	132.69
2007	*4*	236.72	196.91	85.31	186.28	251.20	193.99
	5	199.85	158.18	78.53	155.10	214.15	158.94
	6	165.33	138.52	88.23	137.55	192.81	139.76

Table 2: Summary of Chi Squared Tests Pooled Across All Years (CLV Replication)

Number of Tests Failed	Number of Schools	Percent of Schools
0 of 5	1,288	53.69
1 of 5	684	28.51
2 of 5	248	10.34
3 of 5	176	7.34
4 of 5	3	0.13
5 of 5	0	0.00
Total	2,399	100.00

Table 3: MNL Convergence By Grade: School-Grade-Years with More Than One Teacher				
Grade	*4*	*5*	*6*	*All*
Total School-Grade-Years	11,673	11,617	5,030	28,320
Potential MNL	11,139	10,984	4,054	26,177
Converge with All Student Characteristics	11,116	10,946	4,040	26,102
Converge with Lag Score Only	11,137	10,981	4,054	26,172

Table 4: Predictors of Classroom Grouping: Percentage of Separate School-Grade-Year MNLs in which the Predictor was Significant at the 5% Level

	Grade		
	G4	*G5*	*G6*
Specification 1	*Prior Score Only*		
Prior Math Score	24.52%	25.35%	67.34%
	11,137[1]	10,981	4,054
Specification 2	*Prior Score with Other Covariates*		
Prior Math Score	20.62%	21.41%	63.23%
	11,110	10,927	4,030
Asian	0.29%	0.31%	1.38%
	5,828	6,200	3,041
Black	1.39%	1.46%	5.16%
	10,252	10,174	3,894
Hispanic	1.10%	1.45%	3.66%
	9,826	9,808	3,827
Other Race	0.37%	0.35%	1.37%
	8,393	8,336	3,569
Female	1.54%	1.63%	7.65%
	11,095	10,914	4,025
Disabled- High Incidence	4.66%	5.30%	17.91%
	10,915	10,734	3,965
Disabled- Low Incidence	0.19%	0.16%	0.83%
	1,068	1,267	1,079
FRL	4.20%	4.30%	8.67%
	10,870	10,722	3,990
LEP	1.12%	0.95%	8.31%
	6,331	6,288	2,827
Foreign Born	0.75%	0.86%	3.45%
	8,824	8,990	3,623
Prior Year Absences	3.98%	3.99%	7.03%
	11,103	10,919	4,028
Student in New School	6.18%	3.76%	11.89%
	8,156	9,172	2,574
Number of Schools in Year	2.94%	7.03%	6.44%
	10,717	10,378	3,976

[1] The number below the percentage indicates the number of school-grade-year MNL regressions that the significance of a given variables was testable.

Table 5: MNL Rejection Rates for Prior Scores in Specfication 2: Broken Out by Quartiles of School-level Student Characteristics

School Characteristic	Grade	Q1	Q2	Q3	Q4	All
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%
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%
Female	*4*	19.31%	21.96%	22.27%	18.55%	20.62%
	5	20.51%	23.76%	21.71%	19.20%	21.41%
	6	62.68%	64.51%	63.51%	61.78%	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%

Table 6: OLS Estimates from Regressions of Teacher Characteristics on Classroom Average Student Characteristics

	Dependent Variables: Teacher Characteristics							
	Prior VAM	*Female*	*Asian*	*Black*	*Hispanic*	*Fully Certified*	*Advanced Degree*	*Experience*
Class Characteristics								
Prior Score	0.063***	0.019*	0.000	-0.031***	-0.004	0.004	0.019**	0.831***
	(0.007)	(0.010)	(0.001)	(0.006)	(0.003)	(0.004)	(0.008)	(0.131)
Asian	-0.000	-0.001	0.001	0.003	-0.001	0.002*	-0.001	-0.055
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.050)
Black	-0.001	-0.004**	0.001*	0.030***	-0.006**	0.000	0.004*	0.070
	(0.002)	(0.002)	(0.000)	(0.003)	(0.002)	(0.001)	(0.002)	(0.061)
Hispanic	0.001	-0.001	0.000	0.002	0.006***	-0.001	-0.004	-0.111***
	(0.001)	(0.002)	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.037)
Other Race	-0.001	-0.005**	-0.001	0.007***	-0.002	-0.001*	-0.001	-0.127**
	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)	(0.051)
Female	0.002	0.022***	0.000	-0.001	0.001	0.001	-0.002	0.008
	(0.002)	(0.003)	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.064)
Disabled- HI	0.000	0.008***	-0.000	-0.006**	-0.004***	0.001	-0.001	-0.049
	(0.002)	(0.002)	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.055)
Disabled- LI	-0.002	0.000	0.001	-0.002	-0.002	0.006***	0.004	0.207**
	(0.002)	(0.006)	(0.001)	(0.002)	(0.002)	(0.002)	(0.006)	(0.081)
FRL	-0.008***	-0.002	-0.001***	0.007***	0.002**	-0.002	-0.005**	-0.180***
	(0.002)	(0.003)	(0.000)	(0.002)	(0.001)	(0.001)	(0.002)	(0.063)
LEP	0.001	0.008***	-0.000	-0.005*	0.010**	-0.005*	-0.002	0.156**
	(0.002)	(0.003)	(0.000)	(0.003)	(0.005)	(0.003)	(0.003)	(0.070)
Foreign Born	-0.001	0.001	-0.000	-0.004*	0.005*	-0.000	0.002	-0.013
	(0.002)	(0.002)	(0.000)	(0.002)	(0.003)	(0.001)	(0.003)	(0.055)
Prior Absences	-0.005***	-0.005***	0.000	0.001	-0.002	-0.001	-0.002	-0.120**
	(0.001)	(0.002)	(0.000)	(0.002)	(0.002)	(0.001)	(0.002)	(0.051)
New to School	-0.006***	0.004*	-0.001**	0.005**	-0.001	-0.001	-0.006**	-0.049
	(0.001)	(0.002)	(0.000)	(0.002)	(0.001)	(0.001)	(0.003)	(0.051)
Observations	41,987	51,628	51,628	51,628	51,628	51,628	51,628	51,628
R-squared	0.254	0.142	0.100	0.291	0.328	0.834	0.138	0.200

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 7: 5% Significance-level Rejection Rates of Match Variables from Conditional Logit Estimates

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

Note: For the one tail tests, the test column indicates the direction of the alternative; therefore 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 8: Cross-Sectional VAM Rank Correlations by Grouping and Matching Samples

Panel A: Grouping and Nongrouping Samples from MNL results

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

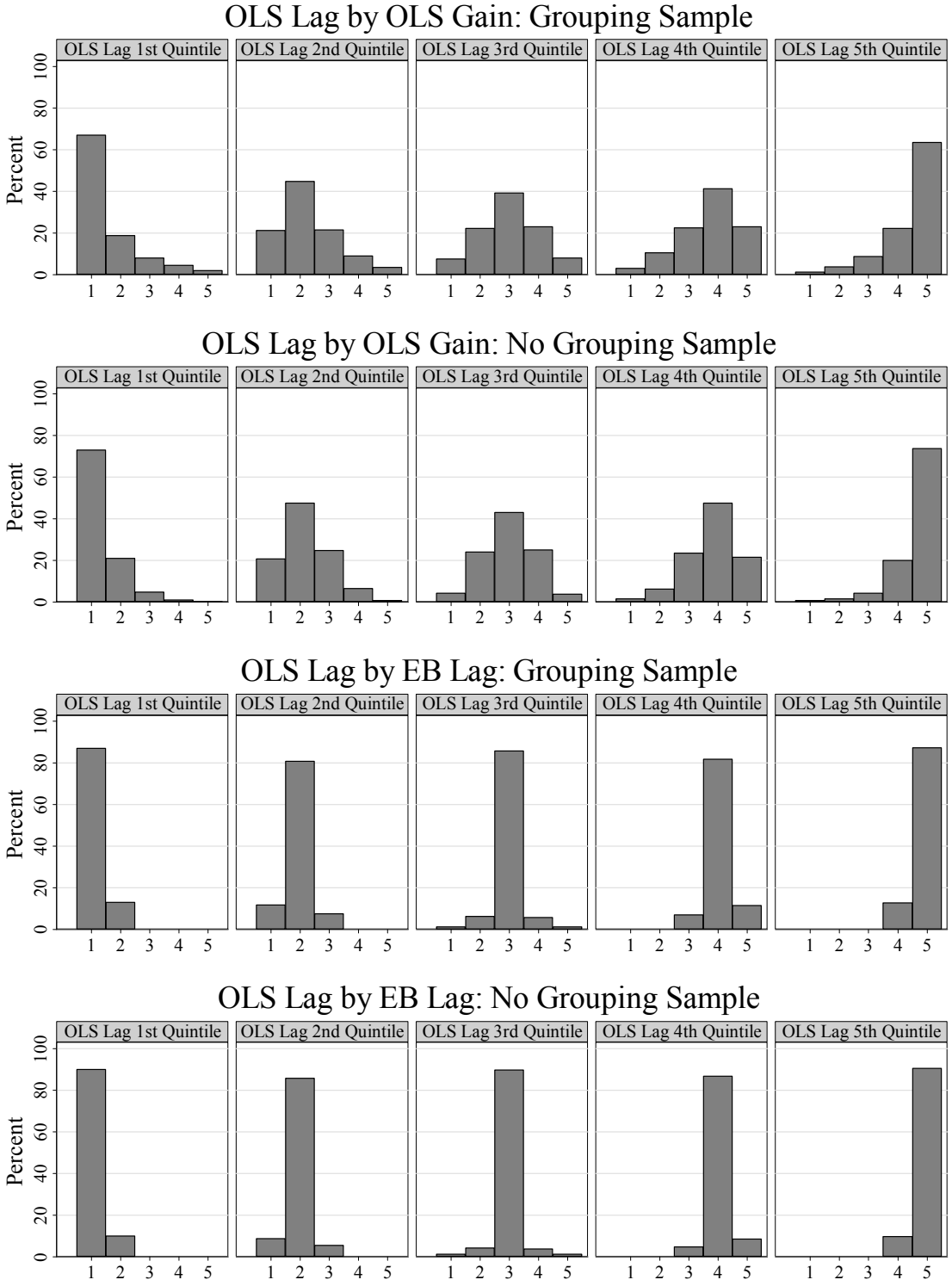
Panel B: Matching and Nonmatching Samples from CL results

Estimator/Model		*OLS Lag*		*EB Lag*		*OLS Gain*	
	Sample	*M*	*NM*	*M*	*NM*	*M*	*NM*
EB Lag	*M*	0.982					
	NM		0.988				
OLS Gain	*M*	0.797		0.7961			
	NM		0.845		0.844		
EB Gain	*M*	0.781		0.812		0.977	
	NM		0.841		0.859		0.986

Sample sizes: G=50,812; NG=91,533; M=9,463; NM=48,036

Figures

Figure 1: OLS Lag Quintile by OLS Gain and EB Lag Quintiles



Appendix Tables

Appendix Table A1: Panel VAM Rank Correlations by Grouping and Matching Samples

Panel A: Grouping and NonGrouping 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 Nonmatching Samples from CL results

Estimator/Model		*OLS Lag*		*OLS Gain*		*EB Gain*	
	Sample	*M*	*NM*	*M*	*NM*	*M*	*NM*
OLS Gain	*M*	0.854					
	NM		0.851				
EB Gain	*M*	0.793		0.925			
	NM		0.825		0.964		
FE Gain	*M*	0.243		0.289		0.284	
	NM		0.561		0.592		0.577

Sample sizes: G=26,887; NG=36,421; M=7,879; NM=25,453