Value-Added Estimation in the Presence of Missing Data

Niu Gao
Public Policy Institute of California

Anastasia Semykina
Department of Economics
Florida State University

Abstract

Ignoring missing data may introduce biases into value-added estimation. We consider a model where both correlated student effects and idiosyncratic factors may cause selection biases. Moreover, we model selection as a function of school-level factors (e.g. charter proximity) that may have distinct effects on teachers in different traditional public schools. We discuss a correction procedure and study its performance using simulations. We find that correction tends to produce noisier estimates of teacher productivity, but can help to substantially reduce the bias. Using observational data from North Carolina we find that corrected estimates of teacher productivity are very similar to those produced by OLS, which appear to be due to small partial effects of school-level covariates on the probability of selection.

Keywords: value-added estimation, missing data, selection correction

1. Introduction

Given that teachers are the single most important school-related resource in improving student achievement, major education reforms in the past two decades have focused on the recruitment, retention, and evaluation of teachers. The *No Child Left Behind Act of 2001 (NCLB)* required states to set standards for designating all teachers as highly qualified, which was generally based on observable characteristics such as education, certification, and subject-matter competency (U.S. Department of Education, 2001, 2009). The relatively low correlation between teacher qualification measures and teacher effectiveness, coupled with the increasing availability of statewide longitudinal database, have promoted policymakers to reconsider teacher evaluation by tying it to student performance. The Obama administration, through its Race to the Top competitive grants and its waivers of the NCLB requirements, encouraged states to incorporate student test scores as a significant component of any new teacher evaluation system (U.S. Department of Education, 2009; 2011). As of today, 43 states require objective measures of student achievement to be included in teacher evaluations and 17 states have further required student growth –calculated using value-added models (VAM) – to be the preponderant criterion (National Center on Teacher Quality, 2015).

The stakes associated with VAM estimates are high because they could be tied to teachers' salaries, promotions, and even employment. In empirical VAM estimation, however, missing data is a pervasive problem. In the state of North Carolina, which was among the first to develop a statewide teacher evaluation system using value-added models, 21 percent of students in middle schools are missing their test scores between 2007 and 2012 school years and 35 percent of these students are missing their test scores at least once.[1] A common practice in empirical VAM is to drop missing cases and proceed using complete cases (McCaffrey and

Lockwood, 2011, Karl et.al., 2013). However, ignoring missing data can introduce biases into VAM estimation when data are missing not at random. Distortions may arise if missingness is determined by factors that also influence student academic achievement (e.g., student ability, study habits, behavioral issues, and family involvement). In such cases, certain types of students (e.g. those who tend to perform worse) are more likely to have missing tests scores and, therefore, may be systematically underrepresented in the data.[2] Consequently, the data sample becomes distorted, which is referred to in the literature as a sample selection problem (Heckman 1976, 1979; Little, 1995). In what follows, we adopt this terminology and use the term 'selection' to describe the missing data problem.

Despite its potential implications for VAM estimation, the literature on addressing the missing data problem in VAM models is surprisingly thin. Several authors highlighted potential inadequacies in the VAM assessment of teacher effectiveness when test scores may be missing for some students (Kupermintz, 2003; McCaffrey et al., 2003, among others). However, to the best of our knowledge, only two papers proposed ways to account for the missing data problem when evaluating teacher productivity. McCaffrey and Lockwood (2011) considered a random effects model, where the unobserved student effects influence both student achievement and probability of observing student test scores. Karl et al. (2013) also estimated a correlated random effects model, but in addition to student random effects they also included teacher random effects in the attendance equation. Both papers have found that the non-randomness in missing data has little impact on the estimated productivity of elementary school teachers, which may be due to a few factors including the downweighting of scores from students with incomplete data (McCaffrey and Lockwood, 2011), and relatively small proportion of missing data in the sample (Karl et. Al., 2013).

In the present paper we extend the existing literature in three ways. Our first contribution is in allowing for additional sources of missing data biases. Along with the correlated student random effects in the achievement and selection equations, our model accounts for a possibility of nonzero correlation in time-varying unobserved factors (idiosyncratic errors) that affect student attendance and academic achievement. Such a correlation can occur, for example, if a sudden change in student's performance motivates their parents to transition them into a private or charter school. Correlated time-specific shocks are expected to generate additional biases in the value-added estimates of teacher productivity. In the present paper we consider an estimation approach that helps to address this problem.

Secondly, we contribute to the literature by including school-level variables (such as proximity to charter and private schools) as additional determinants of the probability that student test scores will be observed. Because a higher concentration of private and charter schools in an area is expected to boost school competition, student mobility is expected to increase (Hoxby, 2000). This would unequally impact the teachers in different traditional public schools (depending on the number of nearby charter and private schools), exacerbating the missing data problem in VAM. We address this issue in our selection correction procedures.

Finally, to gain a better understanding about the role of different factors in the non-random missing data problem, we perform simulations. We evaluate the performance of several estimators under different scenarios and identify the cases where correcting for sample selection is necessary and effective. Next, we draw on matched student-teacher data from the North Carolina Department of Public Instruction (NCDPI) to investigate the extent to which student test score data are missing and study its implications for the empirical VAM estimation. We compare the uncorrected estimates of teacher value-added to the ones obtained using the

selection correction procedures and use our simulation results to disentangle the real data estimates of teacher productivity.

The rest of the paper proceeds as follows. Section 2 introduces the value-added model of student achievement and missing data (selection) model. In Section 3 we discuss estimation and testing for selection bias in models where student test scores may be missing not at random. Section 4 presents simulation results, while Section 5 describes the North Carolina data and presents our empirical findings. Section 6 concludes.

2. Value-Added and Selection Models

We formulate a model of student achievement that is similar to the one used in previous studies of missing data in VAM estimation (McCaffrey and Lockwood, 2011; Karl et al., 2013). The main equation of interest is

$$y_{ig} = \tau_g + \mathbf{x}_{ig}\boldsymbol{\gamma} + \mathbf{T}_{ig}\boldsymbol{\beta} + c_i + u_{ig}, \qquad g = 1, \ldots, G, \tag{1}$$

where y_{ig} is the achievement of student i in grade g, τ_g is a grade-specific intercept, $\mathbf{x}_{ig}$ is a vector of observed student characteristics (such as race and gender), c_i is the student random effect that includes unobserved time-invariant factors (e.g. innate ability), and u_{ig} is an idiosyncratic error.[3] Vector $\mathbf{T}_{ig}$ includes teacher assignment indicators, $\mathbf{T}_{ig} = (T_{ig1}, \ldots, T_{igM})$, where indicator T_{igm} is equal to one if student i was assigned to teacher m's classroom in grade g, and is zero otherwise. The major focus of the value-added analysis is on estimating coefficients $\boldsymbol{\beta}$, which measure teachers' contribution to student learning, or 'value added.'

Studies vary on how teacher effects, $\boldsymbol{\beta}$, are modeled and estimated. If $\boldsymbol{\beta}$ is treated as a vector of random coefficients, then equation (1) is a mixed model, similar to the one estimated in Karl et al. (2013). If, in addition, the equation does not include observed student covariates, it

reduces to a random coefficients model considered by McCaffrey and Lockwood (2011). In this paper, we treat $\boldsymbol{\beta}$ as population parameters that have to be estimated together with $\boldsymbol{\gamma}$ and τ_g. As argued in Guarino et al. (2015), this approach should be more robust to nonrandom teacher assignment, where students with particular observed characteristics may have a higher or lower probability of being matched with better teachers. In this context, $\boldsymbol{\beta}$ is sometimes referred to as 'teacher fixed effects.' In the present paper, we will usually use the terms 'teacher effects' and 'teacher productivity.'

Similar to McCaffrey and Lockwood (2011) and Karl et al. (2013), we assume that the unobserved student effect and idiosyncratic error are independent of the observed student covariates and teacher assignment. However, classroom assignment is allowed to be correlated with the observed student characteristics, $\mathbf{x}_{ig}$. Moreover, assume

$$c_i \sim Normal(0, \sigma_c^2), \tag{2}$$
$$u_{ig} \sim Normal(0, \sigma_u^2).$$

The serial correlation in $\{u_{ig}\}_{g=1}^{G}$ is permitted. Thus, the composite errors, $c_i + u_{ig}$, may be correlated across grades due to both the existence of time-constant unobservables (e.g., student innate abilities) and serial correlation in idiosyncratic errors (e.g., shocks to student test scores).

As mentioned in the Introduction, student achievement data are often incomplete due to various reasons. Students might move to a public school from a private school, or vice versa, which would lead to missing values in test score data. When the analysis is limited to student performance in traditional public schools (which is often the case in the literature), student mobility between charter and traditional public schools also results in incomplete student records. Moreover, students may be missing scores for other reasons, such as absenteeism and

grade retention. As discussed in the literature (Heckman, 1976, 1979; McCaffrey and Lockwood, 2011; Wooldridge, 2010, among others), estimating equation (1) using the observed data produces no biases if student test scores are missing at random (MAR). However, because in practice the MAR assumption can fail, a correction may be needed.

Correction for the missing data problem has been widely discussed in the literature, including the seminal papers by Heckman (1976, 1979) and Rubin (1979). However, to the best of our knowledge, only a few papers propose selection correction in the context of value-added estimation. McCaffrey and Lockwood (2011) use a Bayesian approach to account for nonrandom selection. In their model, the probability of selection and the number of observed student test scores for student *i* depend on the student random effect, which is correlated with the student effect in the achievement equation. Such correlation causes nonrandom sorting of students into classrooms and is expected to cause biases in estimated teacher effects. Karl et al. (2013) consider a correlated random effects model, where both teacher and student random effects are included in the selection equation. These are allowed to be correlated with the teacher and student effects in the achievement model, so that nonrandom sorting may depend not only on the unobserved time-constant student characteristics, but also on the latent productivity of their teachers.

The model considered in this paper is different from the previous literature in two ways. First, in addition to allowing for correlation between the student random effects in the student achievement and selection equations, the model also incorporates a possibility that time-varying idiosyncratic factors in the two equations may be correlated.[4] This helps to accommodate cases, where the same time-specific shock may influence both the student performance and likelihood that the test score is observed. For example, a job loss by a parent during an economic crisis may

have a detrimental effect on the whole family, including the child's performance in school. It may also increase the probability that the student's test score will be missing, especially if the family has to move to a different state in search of better employment opportunities. As another example, the introduction of computerized testing, e.g., the Common Core assessments, may cause a temporary decline in student achievements, particularly among disadvantaged students, who are less familiar with computer devices. It may also lead to a temporary increase in missing test scores, where politically motivated parents opt their students out of these tests. As reported by national media, these numbers may not be trivial and could potentially affect a large number of students (New York Times, 2015, Washington Post, 2016; Politico, 2016). If the students who experience negative shocks to achievement are more likely to have missing scores, it would lead to overestimating the productivity of teachers in classes with larger proportions of missing score data. Conversely, in classes where the test scores are observed for almost everyone (i.e. both low- and high-performing students), the productivity rankings of those teachers would be underestimated.

The second distinctive feature of our model is the introduction of school-level factors that can impact the likelihood of selection. One such factor may be the presence of charter schools in the area where a particular traditional public school is located. The existing studies argue that the charter school location is nonrandom (Bettinger, 2005; Imberman, 2011; Ni, 2009, among others). Moreover, families that choose to transition their children to charter schools may differ from those that decide to keep their kids in traditional public schools. For example, Ladd et al. (2015) find that in North Carolina, parents of charter school students are more likely to have college education. If charter schools attract better students, then teachers in the nearby traditional public schools may be adversely affected. On the other hand, the productivity rankings of

teachers in the other schools (those that do not have charters nearby) are likely to be overestimated. Private schools are expected to have similar effects. Because the proximity to charter and private schools can change over time (due to school openings and closings), their influence on student attendance in traditional public schools may not be accurately captured by school fixed effects. Therefore, selection correction may be needed.

To formalize the discussion, we write the selection model as

$$s_{ig} = 1[\mu_g + \mathbf{x}_{ig}\boldsymbol{\delta} + \mathbf{z}_{ig}\boldsymbol{\varphi} + a_i + v_{ig} > 0], \qquad g = 1, \ldots, G, \qquad (3)$$

$$a_i \sim Normal(0, \sigma_a^2),$$

$$v_{ig} \sim Normal(0, \sigma_v^2),$$

where $1[\cdot]$ is an indicator function equal to one if the expression in brackets is true, and is zero otherwise. Vector $\mathbf{z}_{ig}$ includes characteristics of the school attended by student *i* in grade *g*, such as proximity to charter and private schools. Similar to definitions in equation (1), μ_g is a grade-specific intercept, a_i is the unobserved student random effect, and v_{ig} is the idiosyncratic error, where (a_i, v_{ig}) is assumed to be independent of $(\mathbf{x}_{ig}, \mathbf{T}_{ig}, \mathbf{z}_{is})$. As mentioned above, both $\mathrm{Cov}(c_i, a_i)$ and $\mathrm{Cov}(u_{ig}, v_{ig})$ may be different from zero. Moreover, $\{v_{ig}\}_{g=1}^{G}$ may be serially dependent.

As seen from the formulation of equations (1) and (3), it is assumed that variables in $\mathbf{z}_{ig}$ influence the probability of observing the test score, but have no direct effect on student achievement. In the case of charter school penetration, this may or may not be true, since the empirical evidence is mixed (Bettinger, 2005; Bifulco and Ladd, 2006; Booker et al., 2008; Imberman, 2011; Ni, 2009; Sass, 2006; Winters, 2012). In what follows, we will assume that this assumption holds.[5]

When student random effects and idiosyncratic errors in equations (1) and (3) are correlated, estimating (1) using the observed test score data would generally produce inconsistent estimates of teacher effects. In the next Section, we discuss estimation methods that address the potential selection problem that arises when data are missing not at random.

3. Estimation

Equations (1) and (3) can be estimated jointly using the maximum likelihood estimator. However, a simpler estimator had been proposed in the literature. Let $\varepsilon_{ig} = c_i + u_{ig}$ and $e_{ig} = a_i + v_{ig}$ be the composite errors in equations (1) and (3), respectively. By construction, $(\varepsilon_{ig}, e_{ig})$ has a bivariate normal distribution. As a normalization, assume that $\mathrm{Var}(e_{ig}) = 1$, and denote $\rho \equiv \mathrm{Corr}(\varepsilon_{ig}, e_{ig})$. Heckman (1976, 1979) has shown that under these assumptions, the main equation of interest, i.e. student achievement model, has to be augmented by the correction term to ensure that the error in the observed population is independent of the included covariates. Specifically, for the observed data, the achievement model can be written as

$$y_{ig} = \tau_g + \mathbf{x}_{ig}\boldsymbol{\gamma} + \mathbf{T}_{ig}\boldsymbol{\beta} + \chi\lambda_{ig} + \eta_{ig}, \qquad g = 1, \ldots, G, \tag{4}$$

where $\chi = \rho\sigma_\varepsilon$, $\sigma_\varepsilon = \sqrt{\sigma_\varepsilon^2}$, $\sigma_\epsilon^2 \equiv \mathrm{Var}(\varepsilon_{ig})$, and λ_{ig} is the inverse Mill's ratio (correction term),

$$\lambda_{ig} = \frac{\phi(\mu_g + \mathbf{x}_{ig}\boldsymbol{\delta} + \mathbf{z}_{ig}\boldsymbol{\varphi})}{1 - \Phi(\mu_g + \mathbf{x}_{ig}\boldsymbol{\delta} + \mathbf{z}_{ig}\boldsymbol{\varphi})}. \tag{5}$$

Because including the inverse Mills ratio removes the selection bias, parameters in equation (4) can be consistently estimated using the ordinary least squares (OLS) estimator. The only complication is that parameters in the selection equation are not known, and neither is λ_{ig}. As a solution, the model is estimated in two steps. First, using the fact that e_{ig} has a standard

normal distribution, the selection equation is estimated by pooled probit. Then, estimated parameters in the selection equation are used to compute the inverse Mills ratio,

$$\hat{\lambda}_{ig} = \frac{\phi(\hat{\mu}_g + \mathbf{x}_{ig}\hat{\boldsymbol{\delta}} + \mathbf{z}_{ig}\hat{\boldsymbol{\varphi}})}{1 - \Phi(\hat{\mu}_g + \mathbf{x}_{ig}\hat{\boldsymbol{\delta}} + \mathbf{z}_{ig}\hat{\boldsymbol{\varphi}})}, \quad (6)$$

where $\hat{\mu}_g$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\varphi}}$ denote estimated parameters. Subsequently, $\hat{\lambda}_{ig}$ is used in equation (4) instead of λ_{ig}. Notice that if the errors in the achievement and selection equations are not correlated (i.e. $\rho = 0$), the correction term disappears, so that the original model (1) can be used to consistently estimate teacher effects. Hence, it is easy to check for the presence of the selection bias by testing the hypothesis that the coefficient on $\hat{\lambda}_{ig}$ is equal to zero. Note that when performing the test, it is necessary to compute standard errors that are robust to serial correlation because η_{ig} are correlated across grades.

Equation (4) is obtained under the assumption that the error variance in both equations is constant across grades. In reality, the variance may vary. For example, one might expect that in later grades the unexplained variation in student test scores may be larger due to the increased complexity of the material. Furthermore, the variation in unobservables determining the probability of missing test scores may be higher in grades that coincide with the timing of structural moves. Therefore, it is useful to consider a more flexible specification proposed by Wooldridge (1995). Specifically, let all the above assumptions hold, but allow σ_{ug}^2 and σ_{vg}^2 to vary by g. Then, variances of the composite errors in the achievement and selection equations may also be different across grades.

Under these weaker variance assumptions, Wooldridge (1995) proposes a modified estimation procedure. First, the selection equation is estimated separately for each g, and $\hat{\lambda}_{ig}$ is computed. Then, $\hat{\lambda}_{ig}$ are interacted with grade indicators and interactions are included as

additional covariates in equation (1). A test for the selection bias can be easily performed by testing the joint significance of the interaction terms. Similar to the discussion above, one should use a robust test statistic that accounts for the presence of serial correlation in the errors.

The greater flexibility of Wooldridge's approach makes it more attractive. Moreover, Wooldridge (1995) shows that the normality of the error in the main (achievement) equation is not required. The correction is valid as long as the conditional mean of ε_{ig} is a linear function of the error in the selection equation, e_{ig}. The disadvantage of the model is that it does not permit estimating the error correlation, ρ. A more traditional Heckman-type correction imposes stronger assumptions, but permits estimating ρ, so that the size of the correlation between the errors in the two equations can be examined. In our analysis of North Carolina data we estimate the model using both Heckman-type correction (which we will also call 'pooled correction') and Wooldridge's correction (or, 'grade-specific correction').

4. Simulations

To investigate the performance of the described estimation methods in the presence of missing data we employ simulations. Following the existing literature, we generate data using a relatively simple design that does not include student, school, and peer covariates. Specifically, student test scores for grades 6 through 8 were generated using the following equations:

$$y_{ig} = \mathbf{T}_{ig}\boldsymbol{\beta} + c_i + u_{ig}, \qquad g = 6, 7, 8, \tag{7}$$

where variable definitions are the same as in Section 2. The student unobserved time-constant effect, c_i, and idiosyncratic error, u_{ig}, are drawn from zero-mean normal distributions with variances $\sigma_c^2 = 0.7$ and $\sigma_u^2 = 0.3$, respectively. Teacher effects, $\boldsymbol{\beta}$, were generated using the normal distribution with zero mean and variance equal to 0.0625.

We generate data for 2,500 students, each appearing in grades 6-8. In each grade, students are randomly assigned into classes of 25 students and linked to 100 teachers. Teachers are different in each grade, so that there are 300 teachers in total. We also perform simulations using four cohorts of students (10,000 students), which allows us to study the improvements in the precision of value added estimates when the number of students per teachers grows.

To assign missing values to student test scores, we generate the selection indicator,

$$s_{ig} = 1[\mu + \varphi z_{ig} + a_i + v_{ig} > 0], \quad g = 6, 7, 8, \tag{8}$$

where a_i, and v_{ig} are drawn from zero-mean normal distributions. Similar to the student achievement model, $\sigma_a^2 = 0.7$ and $\sigma_v^2 = 0.3$. Student test scores are set to missing if the selection indicator in the same grade is zero. We vary the proportion of missing test scores by changing the intercept in (8). In simulations, μ can be 0, 0.5, or 1.

Covariate z_{ig} is created to match the variable that we employ in the analysis of the actual student data. In our estimation, z_{ig} is the proximity of the traditional public school to private and charter schools. A higher concentration of alternative educational institutions in the area should make it easier for students to change schools and, therefore, is expected to influence the probability of test scores being missing. To generate an analogous variable in simulated data, we randomly assign teachers to 50 schools, two teachers per grade in each school. Then, z_{ig} is generated as a school-level variable that has normal distribution with unit mean and unit variance. Note that although z_{ig} is constant within a school, it may change over time for a given student, who may be randomly assigned to a teacher in a different school when progressing from one grade to the next. However, the possibility of teacher mobility is excluded. When considering multiple cohorts, teachers stay in the same schools in all years.

To assess the sensitivity of value-added estimates to missing data problem, we vary the strength of the correlation between errors in the achievement and selection equations. Initially, we set $\rho \equiv \mathrm{Corr}(c_i, a_i) = \mathrm{Corr}(u_{ig}, v_{ig}) = 0$, which corresponds to a scenario where test scores are missing at random. Then, we increase ρ to 0.2, 0.4, 0.6 and 0.8, and study how the correlation between the true teacher productivity and value-added estimates changes when using either OLS without selection adjustment or Heckman-type pooled correction that accounts for non-random sample selection. Because in simulated data idiosyncratic errors have constant variance, we do not apply Wooldridge's grade-specific correction. The experiments are performed separately for three values of φ (0, 0.5, and 1), which helps to examine the importance of having covariate z_{it} in the selection equation. Notably, varying φ also affects the percent of missing test score values. For each combination of parameter values, simulations were performed using 100 replications.

Average Spearman's rank correlations for a single student cohort of simulated data (N=2,500) are presented in Table 1. As we discussed earlier, the percent of missing test scores decreases as the intercept in the selection equation (μ) goes up. The same is true for the coefficient on the school-level variable z_{ig}. Notably, when z_{ig} does not affect selection ($\varphi = 0$), the uncorrected OLS and selection correction procedure produce identical estimates of teacher productivity. This happens due to the invariance of the correction term across *i*. Indeed, if $\varphi = 0$, then $\lambda_{ig} = \lambda = \phi(\mu)/[1 - \Phi(\mu)]$ for all *i* and *g*, and it becomes a part of the intercept in (4), making the OLS and selection corrected estimates of teacher effects identical. Note that in this case the correction term captures differences at the group level: students with observed scores vs. those with missing scores. There is no heterogeneity in the selection bias within each group, as long as z_{ig} has not partial effect on the probability of selection. Moreover, for $\varphi = 0$,

Table 1. Average Spearman's Rank Correlations between Estimated Teacher Effects and True Teacher Productivity, Simulated Data, N = 2,500 (One Cohort)

		%		Correlation between unobservables (ρ)				
φ	μ	miss.	Corr. Between **T**	0	0.2	0.4	0.6	0.8
0	0	50.0	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.640	0.639	0.651	0.680	0.732
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.640	0.639	0.651	0.680	0.732
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0	0.5	30.9	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.703	0.704	0.710	0.729	0.769
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.703	0.704	0.710	0.729	0.769
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0	1	15.9	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.739	0.741	0.743	0.751	0.781
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.739	0.741	0.743	0.751	0.781
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0.5	0	32.8	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.687	0.682	0.664	0.658	0.646
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.597	0.593	0.587	0.619	0.668
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.863	0.829	0.796	0.808	0.782
0.5	0.5	18.6	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.731	0.726	0.714	0.706	0.700
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.671	0.663	0.663	0.666	0.721
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.916	0.893	0.870	0.844	0.828
0.5	1	9.0	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.752	0.749	0.740	0.737	0.741
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.695	0.696	0.679	0.704	0.738
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.923	0.915	0.887	0.905	0.911
1	0	24.1	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.697	0.686	0.650	0.614	0.571
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.646	0.640	0.619	0.616	0.636
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.920	0.905	0.847	0.876	0.831
1	0.5	14.5	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.687	0.682	0.664	0.658	0.646
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.597	0.593	0.587	0.619	0.668
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.863	0.829	0.796	0.808	0.782
1	1	7.8	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.752	0.741	0.737	0.721	0.696
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.711	0.704	0.712	0.710	0.724
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.944	0.931	0.911	0.872	0.868

Note: $\text{Corr}(T_{tr}, \hat{T}_{OLS})$ is the correlation between true teacher effects and OLS estimates of teacher effects. $\text{Corr}(T_{tr}, \hat{T}_{cor})$ is the correlation between true teacher effects and the estimates of teacher effects produced by the selection correction procedure described in Section 3.

Spearman's correlations with the true teacher effects are on average higher when there is more nonrandomness in the missing score data (ρ is larger). This is also explained by $\lambda_{ig} = \lambda$ being an intercept in this case. When $\rho = 0$, λ has no partial effect on achievement, and the error variance

is large. On the other hand, when $\rho \neq 0$, $\chi\lambda$ (the intercept) captures part of the error, so the error variance decreases and the estimates of teacher effects become more accurate. Finally, for any given value of ρ, the OLS and corrected estimates are more accurate (more strongly correlated with the true effects) when the proportion of missing values is smaller, which is as expected.

When z_{ig} has a nonzero partial effect on selection (middle and bottom panels in Table 1), uncorrected and corrected estimates of teacher effects are different, and more so as the degree of nonrandomness increases (correlations between $\hat{T}_{OLS}$ and $\hat{T}_{cor}$ tend to decrease as ρ increases). When selection is random ($\rho = 0$), OLS produces more accurate results (more strongly correlated with true effects). This is likely due to selection correction unnecessarily introducing additional noise in the estimation (due to including $\hat{\lambda}_{ig}$ in the equation). However, when ρ grows, OLS becomes less accurate (less correlated with true effects), while the performance of the correction procedure improves, as the selection correction becomes more useful in correcting the bias. As a result, at $\rho = 0.8$, the correction procedure tends to produce higher correlations with the true effects than usual OLS, especially when the proportion of missing scores is relatively large.

Table 2 displays simulation results for four student cohorts (N = 10,000). Correlations with the true teacher effects tend to increase as the number of students per teacher grows, which is likely due to improvements in the precision of the estimation. Other general patterns are similar to those in Table 1, but are now even clearer. For high values of ρ, the correction procedure outperforms OLS more often as the sample size grows. Moreover, the correction does better when the coefficient on z_{ig} is larger. For example, when $\varphi = 1$ and N = 10,000, simulations suggest that correcting for selection bias is preferred to OLS estimation even for moderate levels of nonrandomness ($\rho \geq 0.4$). It is also apparent that accounting for nonrandom

selection is more beneficial when the proportion of missing values is substantial (differences between OLS and corrected estimates are more notable).

Table 2. Average Spearman's Rank Correlations between Estimated Teacher Effects and True Teacher Productivity, Simulated Data, N = 10,000 (Four Cohorts)

		%		Correlation between unobservables (ρ)				
φ	μ	miss.	Corr. Between **T**	0	0.2	0.4	0.6	0.8
0	0	50.0	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.855	0.862	0.869	0.884	0.907
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.855	0.862	0.869	0.884	0.907
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0	0.5	30.9	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.889	0.893	0.899	0.909	0.922
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.889	0.893	0.899	0.909	0.922
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0	1	15.9	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.907	0.909	0.914	0.919	0.926
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.907	0.909	0.914	0.919	0.926
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	1.000	1.000	1.000	1.000	1.000
0.5	0	32.8	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.879	0.870	0.838	0.785	0.723
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.761	0.753	0.779	0.797	0.808
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.861	0.834	0.801	0.733	0.729
0.5	0.5	18.6	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.902	0.895	0.876	0.841	0.796
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.788	0.810	0.816	0.822	0.849
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.871	0.889	0.861	0.812	0.778
0.5	1	9.0	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.913	0.910	0.902	0.885	0.860
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.837	0.829	0.841	0.858	0.854
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.915	0.902	0.894	0.889	0.866
1	0	24.1	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.880	0.855	0.792	0.701	0.616
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.787	0.757	0.812	0.819	0.811
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.886	0.820	0.799	0.732	0.610
1	0.5	14.5	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.900	0.885	0.844	0.781	0.716
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.833	0.829	0.849	0.843	0.854
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.917	0.896	0.866	0.789	0.747
1	1	7.8	$\text{Corr}(T_{tr}, \hat{T}_{OLS})$	0.914	0.904	0.877	0.844	0.809
			$\text{Corr}(T_{tr}, \hat{T}_{cor})$	0.856	0.860	0.868	0.865	0.881
			$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor})$	0.933	0.928	0.922	0.875	0.866

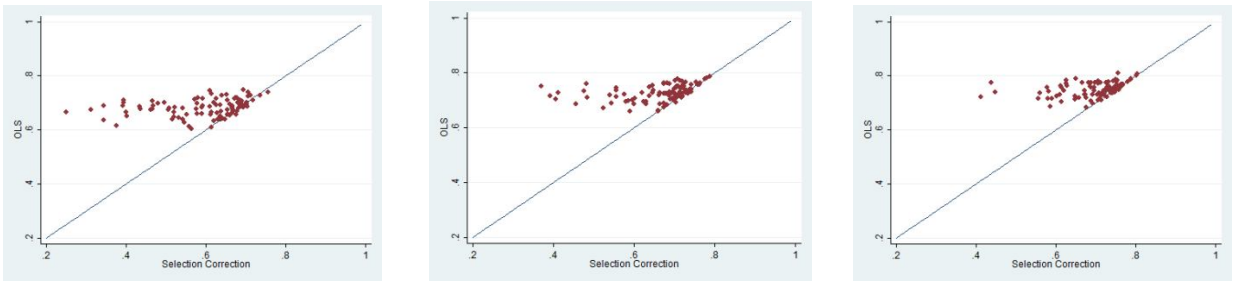
Note: $\text{Corr}(T_{tr}, \hat{T}_{OLS})$ is the correlaton between true teacher effects and OLS estimates of teacher effects. $\text{Corr}(T_{tr}, \hat{T}_{cor})$ is the correlaton between true teacher effects and the estimates of teacher effects produced by the selection correction procedure.

While Tables 1 and 2 present average correlations, more information on the relative performance of the two estimation methods is presented in Figures 1 through 4. The figures graph correlations between OLS and actual teacher effects ($\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$, vertical axis) against correlations between the corrected estimates and actual teacher effects ($\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$, horizontal axis). Single cohort relationships are displayed in Figures 1 and 2. When the selection problem is relatively mild (ρ is close to 0), OLS produces stronger correlations with true teacher effects in the overwhelming majority of the cases, especially when the proportion of missing values is not that large ($\varphi > 0$). As the selection becomes more and more nonrandom (ρ increases), corrected estimates correlate better with the true effects, but are always noisier than OLS estimates. Even when $\rho = 0.8$, where correction does better than OLS in the majority of the cases, there is substantial variation in $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ in the simulated data. In contrast, the variance of $\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$ is roughly constant and relatively small for all combinations of parameter values. Increasing the impact of variable z_{ig} on selection helps to reduce the variation in the correction procedure correlations (Figure 2 vs. Figure 1), but low correlations are still somewhat likely.

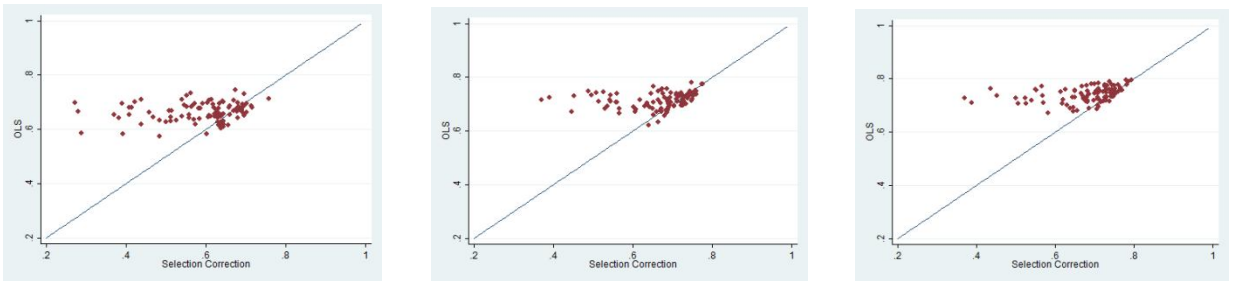
When the number of students per teacher grows (Figures 3 and 4), the correction procedure produces high correlations more frequently. Although the variation in $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ remains relatively high, the correction is much more likely to produce better results in large samples, particularly when both μ and ρ are large. For example, for N=10,000, $\mu = 1$, and $\rho = 0.8$, the correction procedure outperforms OLS in the overwhelming majority of the cases. Similar to what was found previously, the bias in the OLS estimates tends to be smaller when the proportion of missing scores is small ($\varphi = 1$).

Figure 1. Spearman's Rank Correlations Between True Teacher Effects and Corrected and Not Corrected Estimates, Simulated Data, $\mu = 0.5$, N=2,500 (One Cohort)

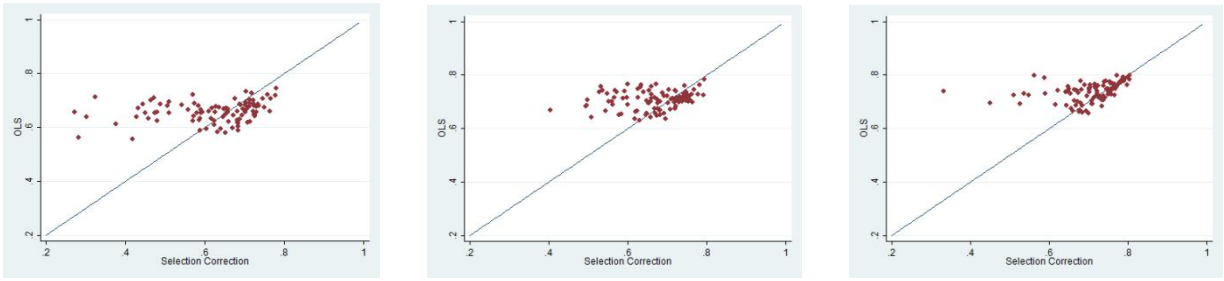
$\mu = 0.5, \varphi = 0, \rho = 0.2$ $\mu = 0.5, \varphi = 0.5, \rho = 0.2$ $\mu = 0.5, \varphi = 1, \rho = 0.2$



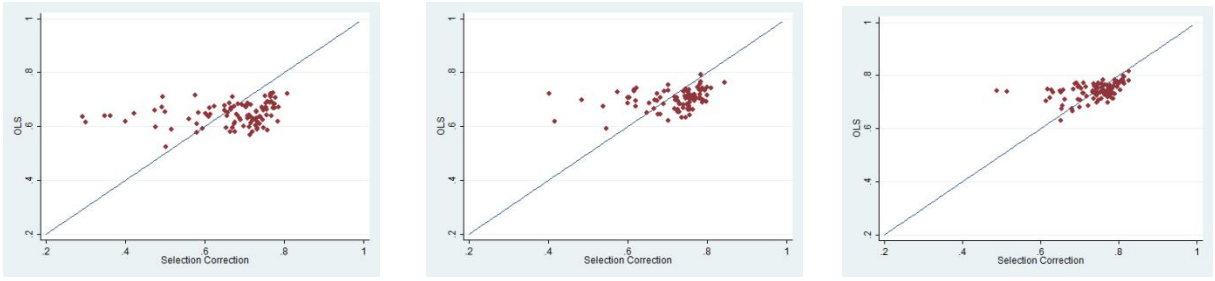
$\mu = 0.5, \varphi = 0, \rho = 0.4$ $\mu = 0.5, \varphi = 0.5, \rho = 0.4$ $\mu = 0.5, \varphi = 1, \rho = 0.4$



$\mu = 0.5, \varphi = 0, \rho = 0.6$ $\mu = 0.5, \varphi = 0.5, \rho = 0.6$ $\mu = 0.5, \varphi = 1, \rho = 0.6$



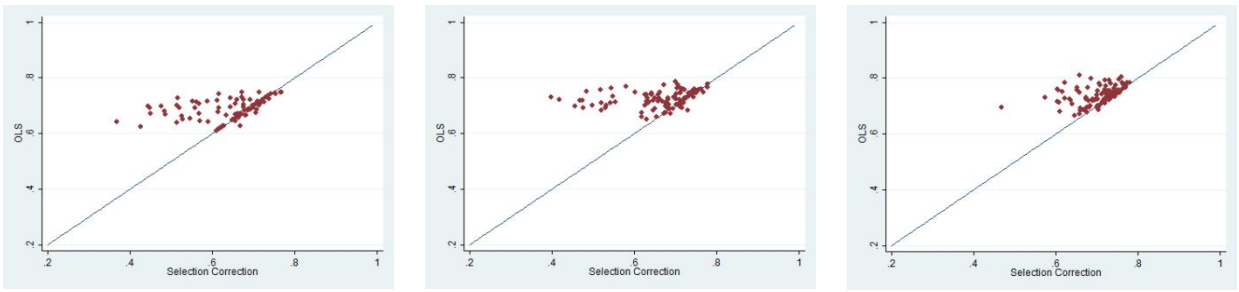
$\mu = 0.5, \varphi = 0, \rho = 0.8$ $\mu = 0.5, \varphi = 0.5, \rho = 0.8$ $\mu = 0.5, \varphi = 1, \rho = 0.8$



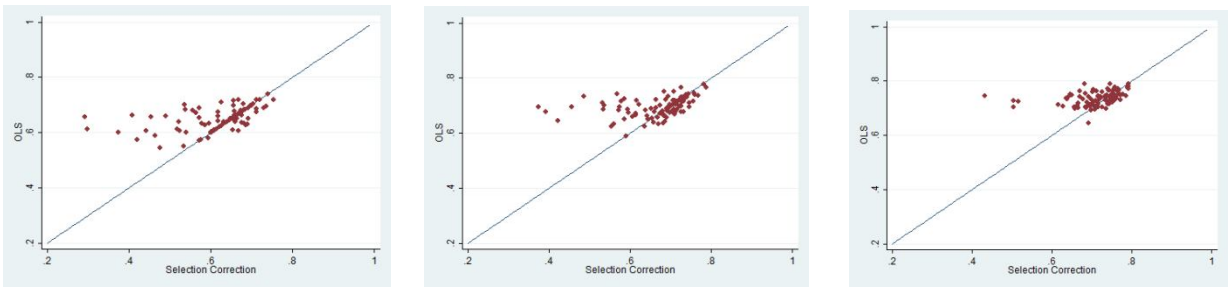
Note: $\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$ is on the vertical axis, $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ is on the horizontal axis.

Figure 2. Spearman's Rank Correlations Between True Teacher Effects and Corrected and Not Corrected Estimates, Simulated Data, $\mu = 1$, N=2,500 (One Cohort)

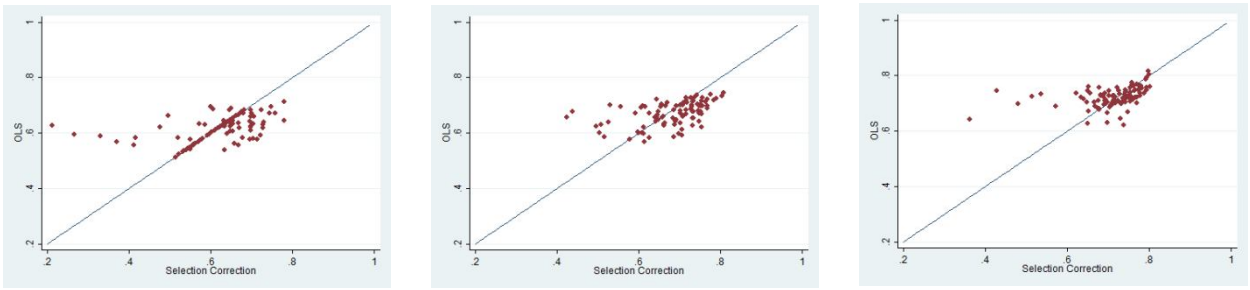
$\mu = 1, \varphi = 0, \rho = 0.2$ $\mu = 1, \varphi = 0.5, \rho = 0.2$ $\mu = 1, \varphi = 1, \rho = 0.2$



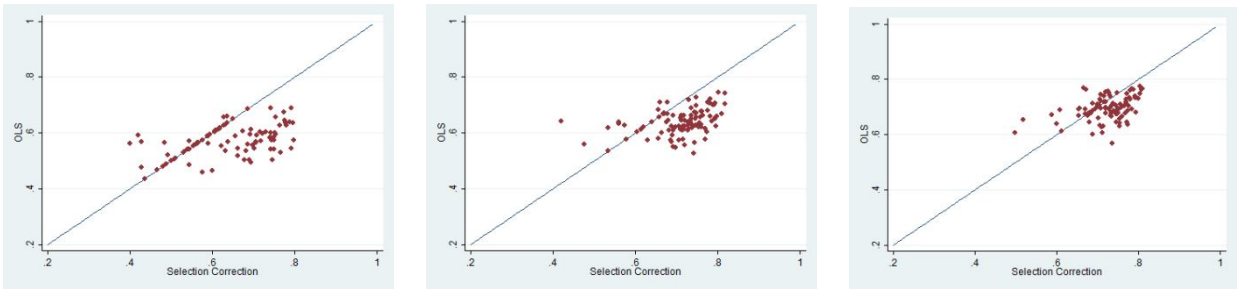
$\mu = 1, \varphi = 0, \rho = 0.4$ $\mu = 1, \varphi = 0.5, \rho = 0.4$ $\mu = 1, \varphi = 1, \rho = 0.4$



$\mu = 1, \varphi = 0, \rho = 0.6$ $\mu = 1, \varphi = 0.5, \rho = 0.6$ $\mu = 1, \varphi = 1, \rho = 0.6$



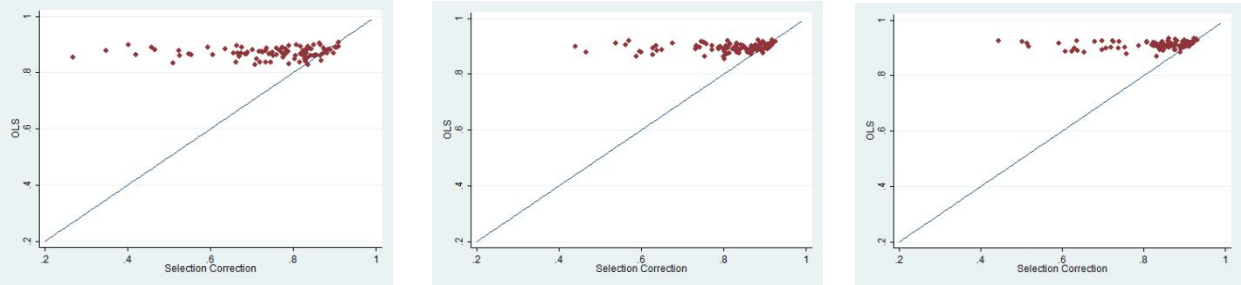
$\mu = 1, \varphi = 0, \rho = 0.8$ $\mu = 1, \varphi = 0.5, \rho = 0.8$ $\mu = 1, \varphi = 1, \rho = 0.8$



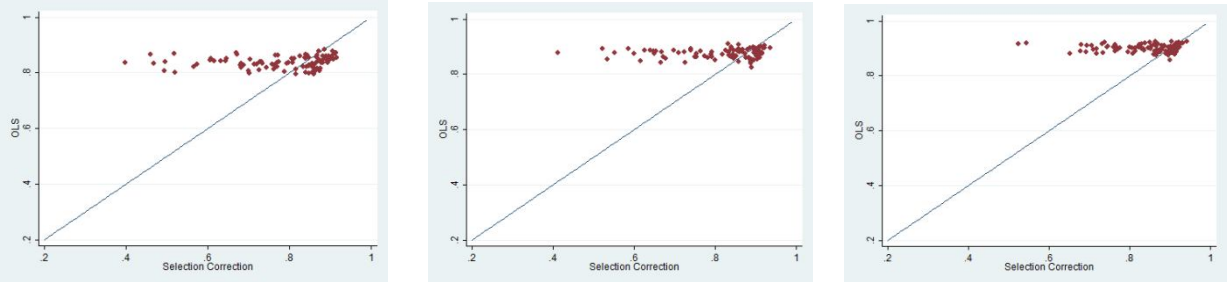
Note: $\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$ is on the vertical axis, $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ is on the horizontal axis.

Figure 3. Spearman's Rank Correlations Between True Teacher Effects and Corrected and Not Corrected Estimates, Simulated Data, $\mu = 0.5$, N = 10,000 (Four Cohorts)

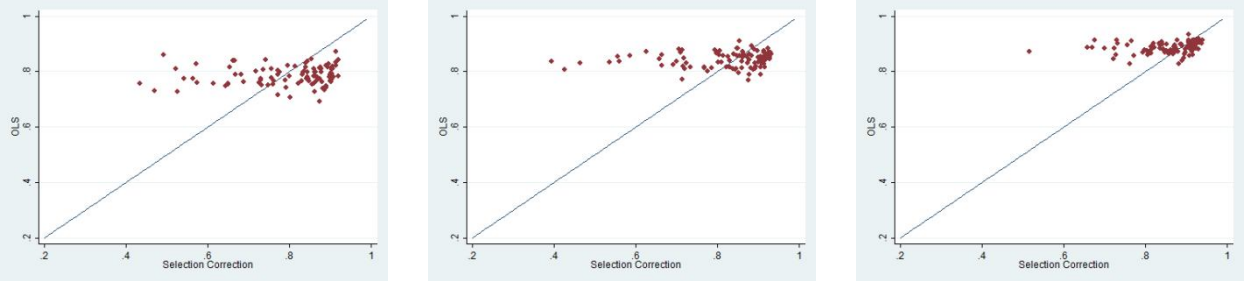
$\mu = 0.5, \varphi = 0, \rho = 0.2$ | $\mu = 0.5, \varphi = 0.5, \rho = 0.2$ | $\mu = 0.5, \varphi = 1, \rho = 0.2$



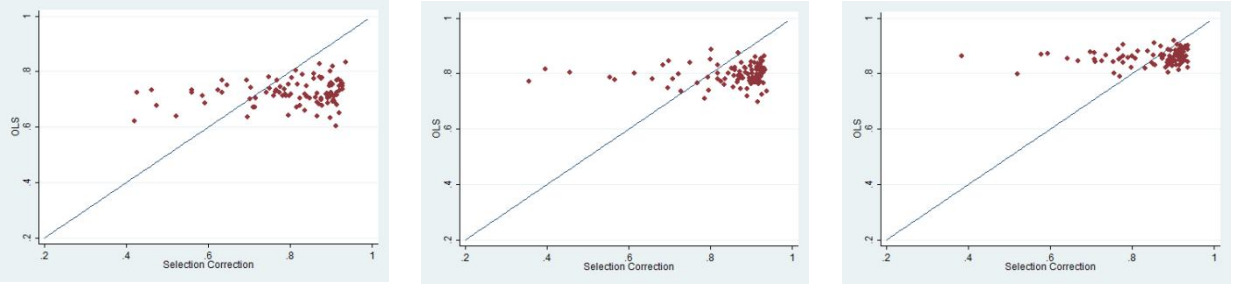
$\mu = 0.5, \varphi = 0, \rho = 0.4$ | $\mu = 0.5, \varphi = 0.5, \rho = 0.4$ | $\mu = 0.5, \varphi = 1, \rho = 0.4$



$\mu = 0.5, \varphi = 0, \rho = 0.6$ | $\mu = 0.5, \varphi = 0.5, \rho = 0.6$ | $\mu = 0.5, \varphi = 1, \rho = 0.6$



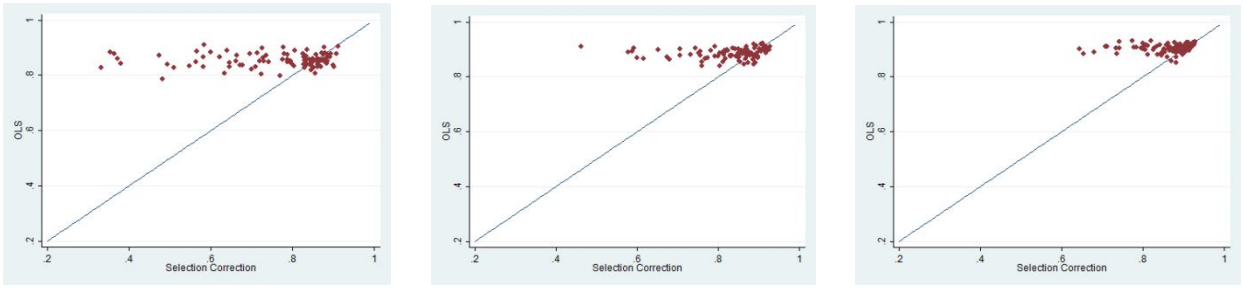
$\mu = 0.5, \varphi = 0, \rho = 0.8$ | $\mu = 0.5, \varphi = 0.5, \rho = 0.8$ | $\mu = 0.5, \varphi = 1, \rho = 0.8$



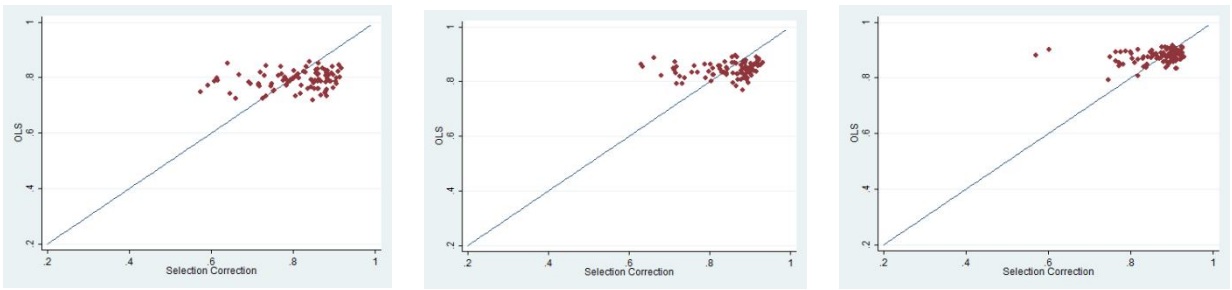
Note: $\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$ is on the vertical axis, $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ is on the horizontal axis.

Figure 4. Spearman's Rank Correlations Between True Teacher Effects and Corrected and Not Corrected Estimates, Simulated Data, $\mu = 1$, N = 10,000 (Four Cohorts)

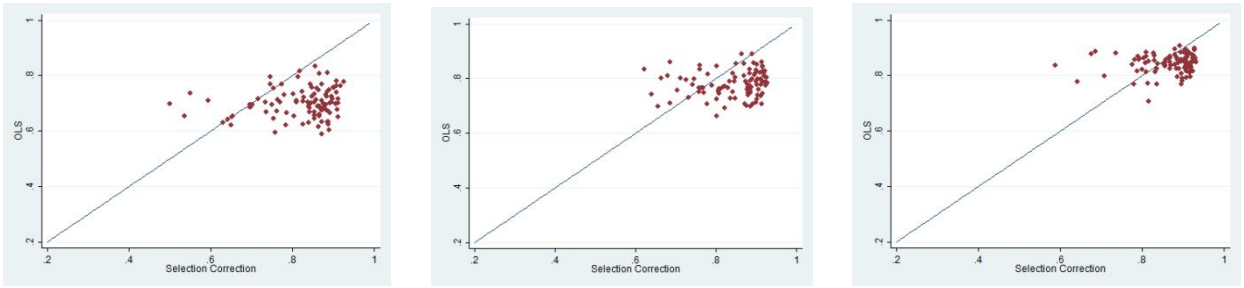
$\mu = 1, \varphi = 0, \rho = 0.2$ $\mu = 1, \varphi = 0.5, \rho = 0.2$ $\mu = 1, \varphi = 1, \rho = 0.2$



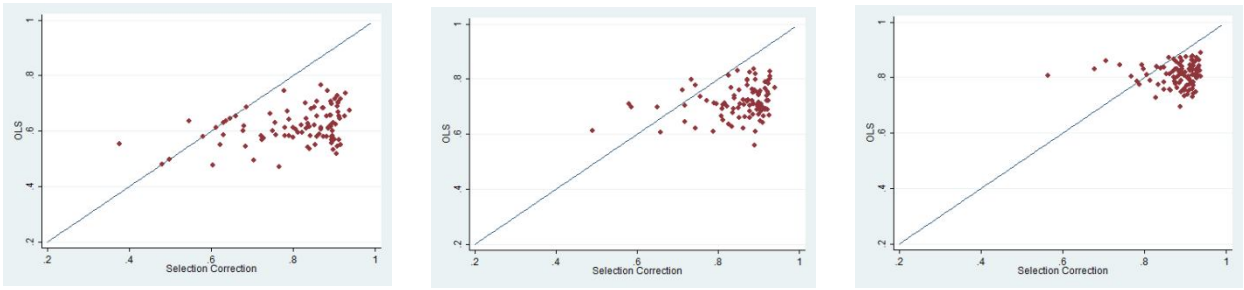
$\mu = 1, \varphi = 0, \rho = 0.4$ $\mu = 1, \varphi = 0.5, \rho = 0.4$ $\mu = 1, \varphi = 1, \rho = 0.4$



$\mu = 1, \varphi = 0, \rho = 0.6$ $\mu = 1, \varphi = 0.5, \rho = 0.6$ $\mu = 1, \varphi = 1, \rho = 0.6$



$\mu = 1, \varphi = 0, \rho = 0.8$ $\mu = 1, \varphi = 0.5, \rho = 0.8$ $\mu = 1, \varphi = 1, \rho = 0.8$



Note: $\mathrm{Corr}(T_{tr}, \hat{T}_{OLS})$ is on the vertical axis, $\mathrm{Corr}(T_{tr}, \hat{T}_{cor})$ is on the horizontal axis.

To summarize, simulations suggest that the OLS estimates of teacher effects are less noisy, but may be seriously biased. Generally, the results indicate that correcting for nonrandom selection is worthwhile when the correlation between the errors in the main and selection equations is relatively sizeable, the coefficient on z_{ig} in the selection equation is large, there are many missing test scores in the data, and when multiple cohorts of students are considered. We will use these findings when interpreting the estimates of teacher effects presented in the next section.

5. Observational Data Analysis

5.1 Data

In this Section, we investigate the missing data problem using the observed actual data. To do so, we draw on matched student-teacher records from the North Carolina Department of Public Instruction (NCDPI). North Carolina is an ideal setting for this type of work for two reasons: first, the state has a long history of standardized testing which dates back to 1997. Each year students from grades 3 to 8 are tested on the End of Grade assessments (EOGs) and high school students are tested on the End of Course assessments (EOCs). Second, the state was among the first to implement a state wide value-added system. North Carolina was awarded the federal Race to the Top (RTTP) grants in 2010, and as part of the RTTP implementation, the state developed a system of value-added scores to determine the impact of individual teachers. Because teacher evaluation is partially based on student achievement, any potential bias caused by missing data may have a substantial impact on individual teachers.

In this study we focus on math achievement in middle schools for a number of reasons. First, there is a clearer link between materials taught in math classes and contents covered in the

end of grade math tests, which permits a more accurate measurement of teacher contribution. In contrast, in many North Carolina schools and districts students may enroll in "Reading" and "English Language Arts" courses, both of which include materials covered in reading assessments. Moreover, because students in middle schools can move between classrooms, the same teacher may teach multiple classes in the same year. This increases the number of students taught by each instructor, which helps to improve precision. It also helps to distinguish teacher effects from common classroom shocks that may affect student achievement. Furthermore, the accuracy of the estimated teacher effects can be improved thanks to averaging the impact of other classroom-specific factors (e.g. peer effects) across multiple classes and cohorts of students. Following the existing literature, we limit our analysis to estimating teacher effects in traditional public schools.

In addition to the grade and subject restrictions, we also excluded cases where it was not possible to identify relevant teachers. For instance, we dropped cases where a student was taught by more than one teacher, attended multiple schools, or had inconsistent race or gender entries over time. All cases where a student repeated the grade were also dropped.

Because it is computationally challenging to estimate models that include all teachers (N=2,941) in our sample, we randomly selected 40 districts out of 119 districts serving middle school grades. Last, the state rolled out its Common Core aligned assessment in 2014 and replaced the old EOGs in 2015. Since the Common Core assessments are harder than the EOGs, results from the two tests are not directly comparable and we did not include data for these years. Our final analytical sample includes 380,182 students and 2,941 teachers in grades 6 to 8 from 2007 to 2012 school years.

From 2007 to 2012 school years, about 21 percent of students in grades 6-8 are missing their test scores in a given year. Not surprisingly, a larger share of students (35 percent) had missing data problem at least once during their middle school years. Most of our missing cases are situations where we do not have students' complete records that include the schools they attended, grades they were in, or teachers they were assigned to.[6] Although some of the missing data are due to students moving out of system (e.g., to charter or private schools), or out of state (e.g., to schools in other states), in other cases data may not be available for unknown reasons. Because we do not have enough information to distinguish between different types of missing data, our selection correction is agnostic with respect to the source of the problem.

As mentioned earlier, when performing correction we include charter and private schools concentration measures as additional covariates in the selection equation. The higher variety of schooling options is expected to increase student mobility, which in turn should influence the likelihood of observing student test scores in nearby traditional public schools. As a measure of concentration we use the number of charter (or private) schools within 30 minutes driving distance from a given traditional public school. We also check the robustness of the findings by considering the number of charter (private) schools within 2.5, 5, or 10 miles driving distance, which are commonly used measures in the literature. In cases where the test score is missing due to student mobility (i.e. when the current school information is not available), we use the values of the concentration measures from the year before (the school that the student most recently attended) or after (the school that the student would attend in the following year) as a proxy. While imperfect, this measure should capture the extent of school competition reasonably well.

In our data, information on charter schools in available only for years 2007-2011. To maximize the sample size, we use 2011 charter school location data for year 2012. If the number

of school openings and closings in a given year is small, the resulting distortions in the data should be minimal. Indeed, when we exclude 2012 from the analysis, the results are very similar. For private schools, the data are available only in 2004. Therefore, we use 2004 private schools concentration measures in all years, but in addition include their interactions with year indicators to allow the impact to change over time. Because private schools measures are much less accurate, we also consider specifications where these measures are excluded from the selection equation.

Table 3. Summary statistics, North Carolina data, 2007-2012.

Variable	N	Mean
White	464,984	57.37 %
Black	464,984	28.35 %
Hispanic	464,984	12.07 %
Asian	464,984	1.59 %
American Indian	464,984	0.62 %
Female	464,984	48.88 %
Math test score	366,812	360.01
		(8.61)
# Charter schools within 30 minutes	464,984	2.51
		(3.03)
# Charter schools within 2.5 miles	464,984	0.21
		(0.55)
# Charter schools within 5 miles	464,984	0.44
		(0.97)
# Charter schools within 10 miles	464,984	1.12
		(1.76)
# Private schools within 30 minutes	464,984	13.60
		(13.66)
# Private schools within 2.5 miles	464,984	0.59
		(1.22)
# Private schools within 5 miles	464,984	1.89
		(2.96)
# Private schools within 10 miles	464,984	4.86
		(6.26)

Note: Standard deviations are in parentheses under the sample means.

Sample summary statistics are provided in Table 3. As expected, the majority of the students are white, with African Americans being the second largest racial/ethnic group. The sample is roughly balanced with respect to gender. The number of charter schools within 2.5 and 5 miles driving distance is on average small, while the numbers tend to be the largest for the 30 minutes driving time measure. Not surprisingly, the number of the nearby private schools tends to be larger than the number of charter schools.

Missing data patterns are summarized in Table 4. The missing data problem seems to have improved in most recent years, largely due to the enhancement of the statewide data system. This pattern is observed in all grades with the exception of grade 8, where changes are more sporadic. Notably, the percent of missing test scores is very low in grade 6 in 2012 and grade 8 in 2007, which is due to these years coinciding with the beginning and the end of the sample period. For example, the 6 graders who attend a charter or private school in 2012, but transition to a traditional public school in 2013 or 2014, are not in the sample. If they were included, they would be counted as students with missing 6 grade scores, which would increase the percentage in Table 5. Similarly, 6 (7) graders who attend a traditional public NC school in 2005 (2006), but move to a different type of school or location in 2007, are not included in our data, which brings the percent of the missing values among the 8 graders in 2007 down.

There appears to be noticeable variation by student race/ethnicity and gender. For instance, the missing rate is slightly higher among African American, Hispanic, and American Indian students, and lower among girls (Table 4). The uneven distribution across racial/ethnic and demographic groups suggests that the observability of test scores may be systematically related to student characteristics.

Table 4. Percent of students missing test scores by year, grade, race, and gender, North Carolina data, 2007 - 2012

year	2007	2008	2009	2010	2011	2012	All years
Overall	25.67%	27.82%	25.44%	19.07%	18.27%	11.65%	21.11%
Grade 6	37.96%	34.86%	27.96%	19.99%	15.32%	2.98%	23.41%
Grade 7	28.53%	30.72%	25.80%	18.82%	18.30%	12.02%	22.10%
Grade 8	3.24%	16.10%	22.41%	18.37%	21.19%	18.88%	17.57%
Male	26.48%	29.01%	27.31%	20.46%	19.39%	12.67%	22.36%
Female	24.81%	26.58%	23.48%	17.61%	17.10%	10.59%	19.81%
White	20.23%	25.59%	23.84%	17.75%	16.85%	10.79%	19.11%
Black	34.35%	29.91%	27.47%	20.84%	20.55%	13.64%	24.28%
Hispanic	32.79%	33.73%	28.16%	20.80%	19.09%	11.25%	22.94%
Asian	24.83%	23.39%	26.43%	19.80%	21.36%	12.32%	21.21%
American Indian	32.97%	45.80%	26.65%	22.37%	20.90%	11.75%	25.92%

5.2 Estimation Results

Here we discuss the results from estimating the models presented in Section 2. We employ OLS to estimate equation (1) using the observed sample, and also perform selection correction by estimating equation (4). We implement both pooled (Heckman-type) and grade-specific (Wooldridge's) corrections, where vary the set of covariates in the selection equation. Student characteristics (race/ethnicity and gender) and grade-by-year indicators are included in both achievement and selection equations. Teacher and school indicators are included in the achievement model, while charter (and, in some specifications, private) schools concentration measures are included in the selection equation. In all regressions standard errors are adjusted for serial correlation at the student level.

To gain a better understanding of the relationship between the probability of observing test scores and student characteristics, it is useful to look at the estimates in the selection equation. Because the errors in the selection equation are assumed to have normal distribution,

we use probit to estimate equation (3). The results from the pooled probit regressions are reported in Table 5. Similar to descriptive results, the probability of the test score being observed is lower among racial and ethnic minorities and higher among females. Specifically, the probability of observing a test score is about 4.4 percentage points lower among African American and Hispanic students than among white students. The probability is even lower for Native American students. Among females, the probability of observing the test score is about 2.6 percentage points higher than among males. Importantly, all concentration measures are highly statistically significant, although their estimated effects are small. For example, for every

Table 5. Pooled probit estimates of partial effects on the probability that the test score is observed; both charter and private schools are included.

	z_{ig} is # schools within 30 min (1)	z_{ig} is # schools within 2.5 miles (2)	z_{ig} is # schools within 5 miles (3)	z_{ig} is # schools within 10 miles (4)
Black	-0.044***	-0.045***	-0.042***	-0.043***
	(0.002)	(0.002)	(0.002)	(0.002)
Hispanic	-0.044***	-0.043***	-0.041***	-0.042***
	(0.002)	(0.002)	(0.002)	(0.002)
Asian	-0.025***	-0.024***	-0.023***	-0.025***
	(0.005)	(0.005)	(0.005)	(0.005)
American Indian	-0.068***	-0.067***	-0.066***	-0.067***
	(0.009)	(0.009)	(0.009)	(0.009)
Female	0.026***	0.026***	0.026***	0.026***
	(0.001)	(0.001)	(0.001)	(0.001)
# Charter schools	-0.010***	-0.025***	-0.023***	-0.015***
	(0.000)	(0.001)	(0.001)	(0.001)
# Private schools	0.002***	0.006***	0.004***	0.003***
	(0.000)	(0.001)	(0.000)	(0.000)
Number of observations	464,984	464,984	464,984	464,984

Note: The dependent variable is an indicator equal to one if student's test score is observed in a given grade-year. P-values are in parentheses under estimated partial effects. Standard errors are adjusted for clustering at the student level. All equations include year and grade indicators. *** Indicates significance at the 1% level.

additional charter school within 30 minutes from the student's traditional public school, the probability of observing the test score decreases by about 1 percentage point, which is likely due to students transitioning from traditional public schools to nearby charters. Private schools are estimated to have a positive effect, which may be due to low competition between private and traditional public schools as they tend to serve different student populations.

When private schools concentration measures are excluded (Table 6), the estimates are qualitatively the same. The size of the estimated partial effects of student characteristics changes slightly, but the signs are preserved. Charter concentration measures are still highly statistically significant, although their effects decrease in magnitude.

Table 6. Pooled probit estimates of partial effects on the probability that the test score is observed; only charter schools are included.

	z_{ig} is # schools within 30 min	z_{ig} is # schools within 2.5 miles	z_{ig} is # schools within 5 miles	z_{ig} is # schools within 10 miles
	(1)	(2)	(3)	(4)
Black	-0.046***	-0.045***	-0.042***	-0.044***
	(0.002)	(0.002)	(0.002)	(0.002)
Hispanic	-0.043***	-0.042***	-0.040***	-0.042***
	(0.002)	(0.002)	(0.002)	(0.002)
Asian	-0.021***	-0.023***	-0.020***	-0.021***
	(0.005)	(0.005)	(0.005)	(0.005)
American Indian	-0.074***	-0.067***	-0.069***	-0.072***
	(0.009)	(0.009)	(0.009)	(0.009)
Female	0.026***	0.026***	0.026***	0.026***
	(0.001)	(0.001)	(0.001)	(0.001)
# Charter schools	-0.003***	-0.019***	-0.013***	-0.005***
	(0.000)	(0.001)	(0.001)	(0.000)
Number of observations	464,984	464,984	464,984	464,984

Note: The dependent variable is an indicator equal to one if student's test score is observed in a given grade-year. P-values are in parentheses under estimated partial effects. Standard errors are adjusted for clustering at the student level. All equations include year and grade indicators. *** Indicates significance at the 1% level.

As the main part of our analysis, we consider estimates of teacher effects. We use first-stage results to compute the selection correction term and include it along with other covariates

in the main equation to obtain estimated teacher effects using the pooled selection correction ($\hat{T}_{cor,P}$). We also re-estimate the selection equation separately for each year-grade combination, and compute the corresponding correction terms. The correction terms and their interactions with grade-by-year indicators are then included in the achievement model to estimate the teacher effects using the grade-specific selection correction ($\hat{T}_{cor,GS}$). Then, we ignore selection and estimate the original model (1) by OLS to obtain the OLS estimates of teacher effects ($\hat{T}_{OLS}$). Finally, we compare results.

Table 7. Spearman's rank correlation coefficients for corrected and non-corrected estimates of teacher effects.

Variables included in $\mathbf{z}_{ig}$	$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor,P})$	$\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor,GS})$	$\text{Corr}(\hat{T}_{cor,P}, \hat{T}_{cor,GS})$
Charter & private schools included			
# schools within 30 min	0.9997	0.9995	0.9994
# schools within 2.5 miles	0.9997	0.9984	0.9989
# schools within 5 miles	0.9995	0.9987	0.9990
# schools within 10 miles	0.9997	0.9991	0.9992
Only charter schools included			
# schools within 30 min	0.9997	0.9991	0.9992
# schools within 2.5 miles	0.9997	0.9985	0.9989
# schools within 5 miles	0.9996	0.9989	0.9991
# schools within 10 miles	0.9997	0.9993	0.9992

Note: $\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor,P})$ is the correlaton between the OLS estimates of teacher effects and estimated teacher effects obtained using pooled selection correction. $\text{Corr}(\hat{T}_{OLS}, \hat{T}_{cor,GS})$ is the correlaton between the OLS estimates of teacher effects and estimated teacher effects obtained using grade-specific selection correction. $\text{Corr}(\hat{T}_{cor,P}, \hat{T}_{cor,GS})$ is the correlaton between the estimated teacher effects obtained using pooled and grade-specific selection corrections.

Table 7 reports correlation coefficients for the estimated teacher effects obtained using different estimation methods. For all specifications, the Spearman's rank correlation coefficients are above 0.99, implying that, similar to previous work (McCaffrey and Lockwood, 2011; Karl et al. 2013), correction has a very minor impact on the estimates of teacher productivity. Even

though the estimates from the grade-specific selection correction appear to correlate less with the uncorrected OLS estimates, they still appear to be very similar.

Table 8. Results from first stage probit regressions and tests for selection bias

Variables included in z_{ig}	Pooled correction	Grade-specific correction			
	$H_0: \chi = 0$	$H_0: \boldsymbol{\chi} = 0$	avg. $\lvert\hat{\varphi}\rvert$ charter	avg.$\lvert\hat{\varphi}\rvert$ private	% cases with $\mathbf{z}_{ig}$ sign. at 5%
	(1)	(2)	(3)	(4)	(5)
Charter & private schools included					
# schools within 30 min	t = -9.28 (0.000)	$F_{18,196529}$ = 12.36 (0.000)	0.083	0.015	88.9
# schools within 2.5 miles	t = -8.80 (0.000)	$F_{18,196529}$ = 15.04 (0.000)	0.114	0.030	72.2
# schools within 5 miles	t = -8.99 (0.000)	$F_{18,196529}$ = 14.62 (0.000)	0.123	0.027	94.4
# schools within 10 miles	t = -8.81 (0.000)	$F_{18,196529}$ = 14.59 (0.000)	0.099	0.023	88.9
Only charter schools included					
# schools within 30 min	t = -8.63 (0.000)	$F_{18,196529}$ = 15.28 (0.000)	0.026	-	72.2
# schools within 2.5 miles	t = -8.58 (0.000)	$F_{18,196529}$ = 15.28 (0.000)	0.094	-	61.1
# schools within 5 miles	t = -8.68 (0.000)	$F_{18,196529}$ = 15.78 (0.000)	0.071	-	72.2
# schools within 10 miles	t = -8.52 (0.000)	$F_{18,196529}$ = 15.19 (0.000)	0.048	-	83.3

Note: P-values are in parentheses under the test statistics.

To untangle the underlying reasons for finding no difference, we use our results in Section 4 and additional information in Table 8. Our simulation results in Section 4 suggested that correcting for selection was most beneficial when the percent of missing values was large (μ is small), the errors in the selection and achievement equations were strongly correlated (ρ is large), additional variables $\mathbf{z}_{ig}$ (such as charter school concentration measures) have large partial

effects in the selection equation (φ is large), and the number of students per teacher is large. In the North Carolina data, student test scores are missing for a sizeable part of the sample (21%). Among teachers whose productivity is estimated, the average number of students per teacher is about 125.[7] Furthermore, the results in Table 8 indicate that ρ is statistically different from zero. Indeed, the null hypothesis that $\chi = 0$ is strongly rejected in both pooled and grade-specific correction equations (columns 1 and 2 in Table 8). This is also consistent with the fact that the estimated correlation coefficient between the errors in the two equations was close to negative one in the pooled selection correction model.

A factor that appears to be driving the results is the small size of the coefficients on $\mathbf{z}_{ig}$ variables. When estimating the selection equation separately for each year-grade combination, the coefficient on the charter concentration measure is on average between 0.026 and 0.123 in magnitude, depending on the specification (column 3 in Table 8).[8] Coefficient estimates on the private school concentration measures are even smaller (column 4). Moreover, the results in column (5), Table 8, indicate that in some specifications the concentration measures are not statistically significant at the 5% significance level. As discussed earlier, all concentration measures are highly significant in the pooled probit regressions, but their estimated effects are very small (Tables 5 and 6).

The good news is that under the scenario similar to the one observed in the NC data, simulations produced the OLS and selection corrected estimates of teacher productivity that were highly correlated with true teacher effects. In other words, for the value-added model considered in this paper, the simulations and actual data results suggest that it should be safe to ignore selection in the NC data.

7. Conclusions

In this paper, we explore the missing data problem in the context of VAM estimation. We extend the models used in the previous literature by allowing the idiosyncratic shocks to student achievement and time-varying shocks to score observability (i.e. selection) to be correlated. Furthermore, we introduce school-level variables (e.g. proximity to charter and private schools) in the selection equation. Such variables may have differential effects on teachers in different schools, which may cause biases in the estimates of teacher productivity.

In order to explore the causes and magnitude of the biases that the missing data problem may bring, we perform simulations and study the performance of several estimators under different selection scenarios. We find that the usual OLS estimates of teacher productivity are less noisy, but can be seriously biased if the errors in the achievement and selection equations are strongly correlated and the proportion of missing test scores is large. The presented selection correction approach produces less precise estimates, but can help to reduce the selection bias. Specifically, correcting for selection bias is preferred to ignoring the missing data problem when the proportion of incomplete data is large, correlation between the errors in the selection and achievement equations is high, the school-level variables have large partial effects on the probability of selection, and the number of students assigned to individual teachers is large.

Using 2007-2012 data for 40 randomly selected districts in North Carolina, we assess the importance of selection correction in the observational data. We find that missing data is a very pervasive problem in our sample (in approximately 21% of student-year cases the test scores are not observed), and the errors in the achievement and selection equations are strongly correlated. While the probability of observing test scores is significantly related to the presence of charter and private schools in the area, their estimated effects are rather small in magnitude. As a result,

the correction produces the estimated teacher effects that are very similar to the uncorrected OLS estimates, with rank correlation coefficients exceeding 0.99 in all specifications. This is similar to the results reported by McCaffrey and Lockwood (2011) and Karl et al. (2013) and suggests that correcting for nonrandom selection has little practical impact on the estimated teacher effects when using the NC data.[9]

While correction does not produce any practical benefits using NC data, the models proposed in this paper could be used in other settings, as more districts and states move on to evaluate teacher performances using value-added models. In such situations, our models can be used to assess whether the selection correction is needed, and how it can be implemented.

Notes

[1] The missing pattern by and large echoed those observed in McCaffrey and Lockwood (2011).

[2] The existing literature suggests that mobility and absences are negatively related to student performance (Dunn at al. 2003; Mehana and Reynolds, 2004; Rumberger, 2003).

[3] The model is formulated for a single cohort, but can be easily extended to multiple cohorts, as we do in our analysis of observational data.

[4] We do not include teacher effects in the selection equation because teacher assignment is often not observed for students with missing score data. Even though past teachers are known for some students with missing test scores, many lack that information in our data.

[5] Some previous studies using North Carolina data, e.g., Bifulco and Ladd (2006), found that after controlling for student and school fixed effects, charter schools do not appear to have any significant effects on student achievement.

[6] Among students with the available school and teacher information, only 3 to 4% had missing score data, which is similar to Karl et al. (2013).

[7] The number is relatively large due to the availability of multiple cohorts. Also, teacher indicators were generated only for the teachers who had 20 or more students with non-missing test scores.

[8] Because the coefficients on the concentration measures sometimes have different signs in different grades and years, we compute the average of their absolute values.

[9] Karl et al. (2013) report substantially lower correlations for the model where lagged teacher effects are included in the selection equation – a specification that we are not able to consider due to data limitations.

References

Bettinger, E. (2005) "The Effects of Charter Schools on Charter Students and Public Schools." *Economics of Education Review,* 24, 133-147.

Bifulco, R., and Helen Ladd (2006) "The Impact of Charter Schools on Student Achievement: Evidence from North Carolina," *Education Finance and Policy,* 1, 50-90.

Booker, K., Scott M. Gilpatric, Timothy Gronberg, and Dennis Jansen (2008) "The Effect of Charter Schools on Traditional Public School Students in Taxas: Are Children Who Stay Left Behind?", *Journal of Urban Economics,* 64(1), 123-145.

Clukey, K. (2016, July 29). In Spite of State Efforts, Test Opt-out Rates Remain High. *Politico*. Retrieved from http://www.politico.com/states/new-york/albany/story/2016/07/opt-out-numbers-for-2016-104370

Dunn, M., Kadane, J. B. and Garrow, J. (2003) "Comparing the harm done by mobility and class absence: Missing students and missing data," *Journal of Educational and Behavioral Statistics* 28**,** 269–288.

Greene, J.P., Greg Forster, and Marcus Winters (2003) "Apples to Apples: An Evaluation of Charter Schools Serving General Student Populations"

Guarino Cassandra M, Michelle Maxfield, Mark D. Reckase, Paul N. Thompson, and Jeffrey M. Wooldridge (2015) "An Evaluation of Empirical Bayes's Estimation of Value-Added Teacher Performance Measures," *Journal of Educational and Behavioral Statistics* 40(2), 190-222.

Harris, Elizabeth, A., and F. Fessenden (2015, May 20). "Opt out" Becomes Anti-Test Rallying Cry in New York State. *New York Times*. Retrieved from https://www.nytimes.com/2015/05/21/nyregion/opt-out-movement-against-common-core-testing-grows-in-new-york-state.html?_r=0

Heckman, James J. (1976) "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economics and Social Measurement* 5(4), 475-492.

Heckman, James J. (1979) "Sample Selection Bias as a Specification Error," *Econometrica* 47(1), 153-61.

Hoxby, C. (2000) "Does Competition among Public Schools Benefit Students and Taxpayers?" *American Economic Review,* 90(5), 1209-1238.

Imberman, S.A. (2011) "The Effect of Charter Schools on Achievement and Behavior of Public School Students," *Journal of Public Economics,* 95(7), 850-863.

Karl, A., Y. Yang, and S. Lohr (2013) "A Correlated Random Effects Model for Nonignorable Missing Data in Value-added Assessment of Teacher Effects," *Journal of Educational and Behavioral Statistic* 38(6), 577-603.

Kupermintz, H. (2003) Teacher Effects and Teacher Effectiveness: A Validity Investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis* 25, 287– 298.

Ladd, Helen F., Charles T. Clotfelter and John B. Holbein (2015) "The Growing Segmentation of the Charter School Sector in North Carolina," *NBER Working Paper* No. 21078.

Little, Roderick J.A. (1995) "Modeling the Drop-Out Mechanism in Repeated-Measures Studies," *Journal of the American Statistical Association* 90(431), 1112-1121.

McCaffrey, D. and J.R. Lockwood (2011) "Missing Data in Value-Added Modeling of Teacher Effects," *The Annals of Applied Statistics* 5(2), 773-797.

McCaffrey, D. F., Lockwood, J. R.,Koretz, D.M. and Hamilton, L. S. (2003) *Evaluating Value-Added Models for Teacher Accountability* (*MG*-158-*EDU*). RAND, Santa Monica, CA.

Mehana, M. and Reynolds, A. J. (2004) "School mobility and achievement: A meta-analysis," *Children and Youth Services Review* 26**,** 93–119.

National Council on Teacher Quality (2015) State of the States 2015: Evaluating Teaching, Leading and Learning. http://www.nctq.org/dmsView/StateofStates2015

Ni, Y. (2009) "The Impact of Charter Schools on the Efficacy of Traditional Public Schools: Evidence from Michigan," *Economics of Education Review,* 28(5), 571-584.

Rubin, Ronald B. (1979) "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association* 76(366), 318-328.

Sass, T. (2006) "Charter Schools and Student Achievement in Florida," *Education Finance and Policy,* 1(1), 91-122.

Strauss, Valerie. (2016, April 6). How Many Students Are Refusing to Take Common Core Tests This Year? *Washington Post*. Retrieved from https://www.washingtonpost.com/news/answer-sheet/wp/2016/04/06/how-many-students-are-refusing-to-take-common-core-tests-this-year/?utm_term=.3ebd409c4d39

U.S. Department of Education. (2001). NCLB: Improving Teacher Quality State Grants. https://www2.ed.gov/programs/teacherqual/index.html

U.S. Department of Education. (2009). State and Local Implementation of the *No Child Left Behind Act*, Volume VIII – Teacher Quality Under *NCLB:* Final Report. https://www2.ed.gov/rschstat/eval/teaching/nclb-final/report.pdf

U.S. Department of Education. (2009). Race to the Top. https://www2.ed.gov/programs/racetothetop/factsheet.html

U.S. Department of Education. (2011). NCLB Flexibility and Waivers. https://www2.ed.gov/nclb/freedom/local/flexibility/index.html

Winters, Marcus A. (2012) "Measuring the Effects of Charter Schools on Public School Student Achievement in an Urban Environment: Evidence from New York City." Economics of Education Review, 31, 293-301.

Wooldridge, J.M. (1995) "Selection Corrections for Panel Data Models under Conditional Mean Independence Assumptions," *Journal of Econometrics* 68(1), 115-132.

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