The Difficulty of Worker Monitoring in the Service Sector: Simulation and Empirical Evidence of the Value-Added of Middle School Teachers

Douglas Harris
Drew Anderson
Xiaodong Fan

University of Wisconsin at Madison

February 16, 2011

Presented at the Michigan State University, Seminar in the Economics of Education

Abstract: Accurately monitoring workers in some service sector jobs presents distinctive challenges because: (a) workers interact with, and receive utility directly from, heterogeneous clients; and (b) client contributions to output vary dramatically and may be correlated with the same client characteristics that affect worker utility. The selection bias problem this causes is particularly severe with public sector services where clients are not voluntary participants (e.g., students are required to attend school) and therefore market indicators of productivity are less informative. This may explain, for example, why it appears that elementary school students are non-randomly assigned to teachers, which has led to concern about the validity of “value-added” teacher productivity measures. We investigate these problems by studying how students and teachers sort into tracks in Florida public middle schools. Using theory and evidence, we generate a data set in which teacher productivity is known. We will use these simulated data to identify value-added specifications that minimize bias relative to actual teacher productivity.

Acknowledgements: The authors gratefully acknowledge financial support from the U.S. Department of Education, Institute for Education Sciences. We also recognize important contributions and research assistance of Robert Kelchen and Rachel Feldman.
1. Introduction

Economists have long considered how workers’ private information about their skill and effort complicates incentive design and leads to market inefficiencies (Spence 1973; Akerlof 1976; Greenwald 1986). Firms can acquire information through supervision to more accurately gauge productivity, but these efforts are costly, especially in the service sector. First, the clients themselves potentially contribute to the outcomes for which workers are held accountable. For example some medical patients follow the directions of their doctors and others do not. Some patients exercise and others do not. These behaviors, along with the actions of their doctors, contribute to health outputs, but none of these client contributions is easy to monitor. We are therefore left to assume, almost certainly falsely, that client contributions are randomly distributed among medical personnel.

Second, service workers directly interact with their clients and they may receive utility not just from the wages that generate consumption, but directly from client interactions themselves. Lawyers, for example, may prefer working with clients who are wealthy business people rather than drug dealers, ceteris paribus. By itself, this is not a problem—workers can simply sort themselves out according to the strength of their preferences, with lawyers having weaker preferences over client characteristics receiving a compensating differential for working with less desirable clients. If a more direct measure of worker productivity were desired (beyond revenue generation), then this too could be accomplished so long as client types are independent of client contributions to output so that individual workers have similar average client contributions over time. Alternatively, if client types and client contributions are perfectly collinear, then, in large firms, workers could be compared with others serving the same types of clients.¹

Even in the unlikely event that the above conditions hold, the situation is even more complicated in the public service sector. First, in many public service sector jobs, it is more likely that

¹ This situation would be further complicated if workers’ marginal products varied by client type. The evidence does not support this scenario in teaching. See later discussion.
one’s preferred clients are quite different from one’s actual clients. People generally prefer interacting with others who are generally like themselves in appearance, language, etc., yet public sector workers often hold advanced degrees and may interact with clients ranging from high school dropouts to students at elite universities. For social workers, probation officers, and public health practitioners, clients might range from violent criminals to college students who went too far at a party. The wide divergence between worker and client characteristics is simply the nature of public sector services.

Second, in the private sector, clients often request and pay more for the services of specific workers. This is not an option in the public sector where the workers are paid by a third party (the government). We cannot even use market indicators like number of requests for particular workers because the stated objective of the public service worker is often to force clients to do things they do not want to do (e.g., the probation officer making sure that the client does not leave the city and violate the conditions of parole). Finally, the public sector has historically been known for nepotism and other forms of corruption that lead workers to be hired for political reasons rather than any objective notion of productivity. For all of these reasons, civil service rules provide strong job protections for public service sector workers. These rules place a premium on objectivity in productivity measurement, preventing the use of subjective market indicators. The result is that the productivity of parole officers can be measured by criminal recidivism and parole violations, but not based on whether parolees think they did a good job. Likewise, social worker productivity could be measured by whether their clients’ children are going to school or the number of times the police are called to a home.

These distinctive features of the public sector make measuring performance even more difficult than in private sector services. We cannot assume that public service workers have no preferences over the large variation in client characteristics, or that those large variations are un-
related to the variation in client contributions. In fact, we would certainly expect that the clients contributing the least to their own outputs are exactly the ones that workers will least prefer to work with. And the variation in client contributions can be very large. In short, measuring productivity in the service sector is difficult, and doing so in the public service sector is even harder.²

Public middle schools represent useful sites for studying the monitoring problem in the service sector in general and public services in particular. There is ample evidence that teachers have preferences over whom they teach, e.g., preferring to teach students with backgrounds similar to the teacher’s own (Boyd, Lankford, Loeb, and Wyckoff, 2005; Finley, 1984). Further, it is well known that students and their families contribute a great deal to student learning (e.g., Coleman, 1966; Fryer and Levitt, 2004; Harris, 2007). Taken together, this evidence suggests that the students who teachers typically least prefer to teach are also the students (and their families) who contribute the least to their own learning outcomes.³ But yet some teachers have stronger preferences about student characteristics than others and it would make little sense to punish teachers simply because they are willing to, or even prefer, teaching more disadvantaged clients who, along with their families, contribute less to their own learning.

A plethora of panel data studies have tried to solve this problem by estimating teacher productivity based on student test scores, i.e., teacher “value-added.” Results suggest a wide distribution of productivity (Kane & Staiger, 2008; Hanushek, 2011) and policymakers have used this finding to support merit-based compensation, tenure, and promotion decisions. However, there appears to be considerable bias and imprecision in value-added estimates that limits their usefulness in monitoring and incentive design (Harris, 2009, 2011) with questions about the scale of the

---
² An additional problem in the public sector is that policymakers desire not just efficiency but equity in service delivery. So, the efficient solution to the monitoring problem might yield similar productivity by client type if the equity is concerned over the outcome. While we do not directly address equity concerns in this paper, our modeling of the sorting process is motivated by observed patterns in the teacher distribution by race/ethnicity and family income.
³ This is partly why some districts pay teachers more to work in schools serving minority and low-income students, sometimes called “combat pay.” In theory, this can reduce inequities in learning outcomes.
standardized tests (Ballou, 2009), large year-to-year fluctuations in value-added scores for individual teachers (McCaffrey, Sass, Lockwood, & Mihaly, 2009; Kodel & Betts, 2009), and variation in results across different testing instruments within the same subject (Papay, forthcoming). The results are also extremely sensitive to specification and the underlying assumptions of value-added modeling are all rejected (Harris, Sass, & Semykina, 2010).

The selection bias problem has received particular attention (Rothstein, 2009, 2010), though prior work has not attempted to model the sorting process that generates the bias. One prior study has attempted to find econometric specifications that generate the least bias, but does not attempt to model the underlying sorting process (Guarino, Reckase, and Wooldridge, 2010). We start by applying the Rothstein falsification tests to show that the assumptions necessary for value-added estimates to be unbiased do not hold in our sample of Florida middle schools. Then, we provide descriptive statistics about the sorting process and test some simple theories about how the sorting process works. That is, we try to get inside the non-random sorting process that Rothstein has identified.

In future work, as in Guarino, Reckase, and Wooldridge (2010), we plan to use a simulation approach to model the sorting mechanism, but we approach this differently by creating a data generating process calibrated with actual data to re-create key real-world equilibrium conditions, including the Rothstein falsification test results; the calibration of the model and other data is carried out using a rich statewide data set for the entire state of Florida. We will then apply multiple standard value-added specifications to see which best addresses the selection bias problem. Finally, we will apply the best specifications to the data and compare the bias and variance of the results with those from more standard specifications.
2. Data

Florida has an unusually rich and complete statewide panel data system that has been the basis for many value-added analyses (e.g., Harris and Sass, 2009, 2010). The data cover all public school students throughout the state and include student-level achievement test data for both math and reading in each of grades 3-10. Unlike most other statewide databases, we can precisely match students and their teachers to specific classrooms at all grade levels. We can determine the specific classroom assignments of middle school and high school students, who typically rotate through classrooms during the day for different subjects. This enables us to better separate the effects of teachers from students.

We restrict our analysis to math because it is generally believed that reading scores are influenced by non-school behaviors that are outside the control of educators, as well as other classes such as social studies, that involve reading but where developing reading is not the primary purpose. We also restrict our analysis to middle schools, which we define as institutions that offer grades 6th-8th exclusively. Middle schools are a useful target of analysis for several reasons. Students have test scores in each middle school grade, as well as prior grades (a key basis for tracking as we will show).

---

4 The state of Florida administered two sets of reading and math tests to all 3rd through 10th graders in Florida. The “Sunshine State Standards” Florida Comprehensive Achievement Test (FCAT-SSS) is a criterion-based exam designed to test for the skills that students are expected to master at each grade level. The second test is the FCAT Norm-Referenced Test (FCAT-NRT), a version of the Stanford Achievement Test used throughout the country. Version 9 of the Stanford test (the Stanford-9) was used in Florida through the 2003/2004 school year. Version 10 of the Stanford test (the Stanford-10) has been used since the 2004/05 school year. To equate the two versions of the exams we convert Stanford-10 scores into Stanford-9 equivalent scores based on the conversion tables in Harcourt (2003). The scores on the Stanford-9 are scaled so that a one-point increase in the score at one place on the scale is equivalent to a one-point increase anywhere else on the scale. The Stanford-9 is a vertically scaled exam, thus scale scores typically increase with the grade level. We use only the FCAT-SSS in this version of the paper, but will also apply the FCAT-NRT as sensitivity analysis in subsequent work.

5 Koedel (2009) provides some evidence that social studies teachers influence reading test scores at the high school level.
Also, unlike elementary schools, students are explicitly tracked into remedial, general, and advanced (remedial, middle, advanced) tracks—a key aspect of the sorting process.\footnote{There are some additional codes in the Florida data warehouse, but these comprise only 1.2 percent of total courses and are dropped.} Elementary and high schools, in contrast, do not test in every grade, and elementary schools lack explicit cross-course tracking.\footnote{In the Florida data, the remedial track is the only one in which the grade level is not also indicated in the data. Therefore, we assigned the grade level based on the most frequently occurring grade level of students in the course. We find these courses are very much separated by grade; typically, almost all students in a given intensive course are in the same grade (98\% on average).} This implies there are nine types of math courses in each middle school: 6th grade remedial track, 6th grade general track, and so on for the three tracks and three grades.

Focusing on middle school math is also supported by the results of Harris and Sass (forthcoming) who use value-added analysis to study the returns to teacher experience and formal training in terms. They argue that there are several reasons to believe that the assumptions necessary for valid value-added estimation are more likely to hold in middle school math and, consistent with this, they find results more frequently of the expected sign than in other grade/subject combinations.

To avoid atypical classroom settings including jointly taught classes, we consider only courses in which 10-40 students are enrolled and there is only one “primary instructor” for the class. Finally, we eliminate charter schools from the analysis since they may have differing curricular emphases and student-peer and student-teacher interactions may differ in fundamental ways from traditional public schools.

We restrict our analysis to students who receive math instruction from two teachers or less in a given year; although all students enrolled in a course are included in the measurement of peer-
group characteristics. We drop approximately 6.7% of our observations for having more than two math teachers in a given year.\footnote{In future analysis, we will apply weights in the value-added estimation that address the fact that many middle school students have two teachers in a given year.}

The resulting data file, summarized in Table 1, is an unbalanced panel of 1,323,702 unique students observed in Florida middle schools over a 6-year period for the years 2003-2004 through 2008-2009, resulting in 2,431,000 student/year observations. We observe 12,442 math teachers (some of whom also teach reading or other subjects), with an average of 10.2 years of teaching experience. The sample consists of 513 middle schools with a median school size of 832 students. Table 2 provides additional information broken down by course track. We observe 334,235 unique courses (for instance, one section of 6th grade advanced math counts as one course). The table shows the distribution of courses across tracks, standardized test scores, and the class sizes by track.

3. Endogeneity and Dynamic Tracking in Middle Schools

We focus in this paper on how student and teacher sorting complicates productivity monitoring, particularly with “value-added” measures. This section provides a brief introduction to how value-added measures are typically estimated. This shows more concretely what assumptions are necessary for there to be no bias and provides a basis for identifying model specifications for our future empirical work. We also discuss the important work of Rothstein (2009, 2010) and his evidence that students and teachers are non-randomly sorted.

3.1 Common Value-Added Models

Following prior work (Boardman and Murnane, 1979; Harris and Sass, 2006; Harris, Sass, and Semykina, 2010; Todd and Wolpin, 2003), we begin with a general cumulative model of student achievement:
\[ A_{it} = A_i[\mathbf{X}_{it}, \mathbf{F}_{it}, \mathbf{E}_{it}, \mu_{i0}, \epsilon_{it}] \]  

where \( A_{it} \) is the achievement level for individual \( i \) at the end of their \( t \)th year of life, and \( \mathbf{X}_{it}, \mathbf{F}_{it} \) and \( \mathbf{E}_{it} \) represent the entire histories of individual, family and school-based educational inputs, respectively. The term \( \mu_{i0} \) is a composite variable representing time-invariant characteristics an individual is endowed with at birth (such as innate ability), and \( \epsilon_{it} \) is an idiosyncratic error.

To make the estimation computationally feasible, it is common to assume that the cumulative achievement function, \( A_i[\cdot] \) is linear and additively separable, \(^9\) family inputs are constant over time and are captured by a student-specific fixed component, \(^10\) input effects decay geometrically, and the cumulative achievement function does not vary with the grade level.

One key assumption that differentiates the various empirical specifications of the cumulative achievement model is the treatment of the impact of the individual-specific effect on achievement. One approach is to assume that time-invariant student/family inputs decay at the same rate as other inputs. In this case, the individual-specific effect drops out of the achievement equation:

\[ A_{it} = \alpha \mathbf{X}_{it} + \beta \mathbf{E}_{it} + \lambda A_{it-1} + \eta_{it} \]  

The lagged test score thus serves as a sufficient statistic for the time-constant student/family inputs as well as for the historical time-varying student and school-based inputs. OLS estimates of equation (8) would be consistent so long as the \( \eta_{it} \) are serially independent. If the \( \eta_{it} \) are not serially

\(^9\) Figlio (1999) and Harris (2007) explore the impact of relaxing the assumption of additive separability by estimating a translog education production function.

\(^10\) In general, one could consider models with uncorrelated unobserved heterogeneity. However, it is likely that the observed inputs (e.g. teacher and school quality) are correlated with the unobserved student effect, which would lead to biased estimates in a random-effects framework. Therefore, in what follows, we assume that the unobserved heterogeneity may be correlated with the observed inputs and focus on a student/family fixed effect.
independent, one could use instrumental variable (IV) estimation techniques, employing \( A_{t-2} \) and longer lags as instruments for \( A_{t-1} \).\(^{11}\)

An alternative is to assume that the marginal effect of the individual-specific component is constant over time. Current achievement is then:

\[
A_i = \alpha X_{it} + \beta E_{it} + \lambda A_{it-1} + \gamma_i + \eta_i \tag{3}
\]

where \( \gamma_i \) is an individual student fixed effect. If \( \eta_{it} \) are serially independent, equation (3) can be consistently estimated by first differencing (FD) to remove the individual effect and instrumenting for \( \Delta A_{it-1} \) using \( A_{t-2} \) and longer lags as instruments.\(^{12}\)

A final factor that distinguishes commonly estimated value-added specifications is the assumed rate of decay, \( \lambda \). There are three typical choices: unrestricted decay, no decay, or complete decay. If no constraints are placed on \( \lambda \), we are left with either equation (2) or equation (3), depending on whether the fixed student/family inputs are assumed to decay at the same rate as do the time-varying student and school-based inputs or their effect is assumed constant over time. Ordinary least squares (OLS) estimation of the resulting equation (3) is problematic. Since \( \eta_{it} \) is a function of the lagged error, \( \varepsilon_{it-1} \), the lagged achievement term, \( A_{it-1} \), may be correlated with \( \eta_{it} \), in which case OLS estimates of equation (3) will be biased.

Alternatively, one can assume that there is no decay in the effect of past schooling inputs on current achievement, i.e. \((1-\lambda)=0\) or \(\lambda=1\). Setting \(\lambda=1\) implies that the effect of each input must be

---

\(^{11}\) This requires an AR(1) process.

\(^{12}\) The lagged difference could also be used as an instrument, but it has a disadvantage of imposing restrictions on parameters (in the first-stage regression, where \( \Delta A_{it-1} \) is the dependent variable, \( A_{it-1} \) is forced to have the same coefficient as \( A_{it} \)). Using \( A_{t-2} \) and \( A_{t-3} \) instead of the difference is more flexible and is likely to produce higher correlation with the instrumented variable (stronger instruments). So, using lags (rather than differences) is generally preferred.
independent of when it is applied, such that school inputs each have an immediate one-time impact on achievement that does not decay over time. Given this assumption, the coefficient on lagged achievement in equations (2) and (3) is unity.\(^{13}\) One can then subtract \(A_{it-1}\) from both sides of equation (3) to obtain:

\[
\Delta A = A_t - A_{t-1} = \alpha X_t + \beta E_t + \gamma_i + \eta_t
\] (4)

A fourth alternative is to assume there is no effect of lagged inputs on current achievement. In other words, there is immediate and complete decay. In this case, lagged achievement drops out of the achievement function and equations (2) and (3) become:

\[
A_t = \alpha X_t + \beta E_t + \gamma_i + \eta_t
\] (5)

where the individual-specific effect is assumed to decay at a different rate than other inputs and therefore does not drop out of the equation. In comparison to other studies, Rothstein (2009) considers all of the above except for (3). Guarino, Reckase, and Wooldridge (2010) consider (3) and (4), as well as two variations of (2) that assume \(\lambda = 1\) (what they call “pooled OLS” and “random effects”).

Harris, Sass, and Semykina (2010) develop and run tests of many of the above assumptions and reject all of those tested. But again our key concern here is that students and teachers may be assigned in ways that violate the strict exogeneity assumptions.

3.2 Dynamic Tracking and Sorting on Unobservables

A series of papers by Rothstein (2009, 2010) focuses on the central assumption in all econometric studies of treatment effects, and of value-added models in particular: teacher

\(^{13}\) Alternatively, the model can be derived by starting with a model of student learning gains (rather than levels) and assuming that there is no persistence of past schooling inputs on learning gains.
assignments have to be orthogonal to all other factors determining achievement gains in all grades, conditional on all covariates. He tests this via the application of the falsification test developed by Chamberlain (1984). The intuition of the test is that future teachers cannot cause past outcomes. Formally, the sufficient strict exogeneity condition is:

$$E[u_{it}|E_{it}, E_{i,T-1}, ... E_{i0}, X_{it}, X_{i,T-1}, ... X_{i0}, \mu_{i0}] = 0$$ (6)

Rothstein proposes a test of this assumption along the following lines (adapted to middle school grades):

$$Y_{it}^6 = \alpha_i + X_{it}\theta + Y_{it-1}^5 + T_{it}\varphi + T_{it+1}\beta + u_{it}$$ (7)

where, for simplicity, teachers represent the only school input so that $E_i$ in (6) is replaced by a vector of teacher assignments, $T_i$. We can then use an $F$-test for whether $H_0: \beta_1=...=\beta_{\kappa}$. Similarly, if the future teacher effects are small, then they will have remedial variance than current teacher effects; therefore, the ratio of the variances presents an additional test statistic.

Equation (7) can be difficult to compute because of the multiple layers of fixed effects. It is even more difficult to compute when, as in (4), one of the covariates is an individual effect. Rothstein addresses this partly by limiting analysis to students who do not switch schools and by using correlated random effects estimation, treating the student effect as a random draw that may be correlated with the other covariates. Koedel and Betts (2009) instead address the problem by subtracting equation (7) from the analogous equation when 7th grade achievement is the dependent variable, estimating the resulting first-differenced equation, and carrying out a similar set of
hypothesis tests. This approach reflects the logic that a two-period first-differenced equation is equivalent to a fixed effects model.

The results of Koedel and Betts (2009) and Rothstein (2009, 2010) vary for at least three reasons: (a) different samples and therefore possibly different sorting mechanisms; (b) somewhat different statistical tests (see above); and (c) Rothstein only uses one cohort whereas Koedel and Betts’s version uses multiple cohorts. This last point appears particularly important as the ratio of the teacher effect standard deviations drops from one-half to one-third when additional cohorts are added. To our knowledge, no one has tested whether the Koedel and Betts and Rothstein tests yield similar results with the same sample.

The only application of the falsification test to middle (and high) schools that we are aware of is Harris and Sass (forthcoming) who use the Koedel and Betts approach. They also use the same Florida data, making this especially pertinent to the present analysis. They reject exogeneity in middle school math in eight of Florida’s 68 districts (the test would not run in nine districts). Many of the districts where exogeneity was not rejected had very small samples and many also have $p$-values that were nearly statistically significant.

All of the above tests are “conservative” in the sense that they include school indicators as part of equation (7) and therefore only capture student and teacher sorting within schools. Given the tremendous variation in measured outcomes across schools, and evidence that teachers gravitate to higher-scoring schools as they progress through their careers, cross-school sorting is also likely to play a role and add to the bias. Specifically, the falsification test might reject exogeneity even more strongly when the school indicators are removed, but again this has not apparently been tested.

While there is relatively little evidence on dynamic sorting, especially in middle schools, this prior evidence strongly points to rejection of strict exogeneity and therefore bias in teacher value-added estimates. This suggests a need to better understand and model the sorting process.
3.3 Modeling Teacher and Student Sorting

Unfortunately, there is extremely little empirical research from which to begin modeling the sorting process. The conventional wisdom is that teachers prefer working with academically strong students and there is some qualitative evidence to support this (Finley, 1984). Likewise, it is widely believed that students are tracked based on their academic ability, especially their test scores and grades (Oakes, 1986). But this only begins to touch the surface of the potential ways that the strict exogeneity assumption might be violated. We therefore begin by asking some very basic descriptive questions and posing and testing theories. Understanding the assignment process facilitates a realistic data generating process for the simulations.

First, how much variation is there in the availability of courses in various tracks across schools? This is an important question because we hypothesize, and the data support, the fact that the number of courses in each track available within a school is an important factor affecting teacher and student sorting into tracks. However, because many value-added models are estimated based on within-school variation, and to keep the complexity of the later theoretical model to a manageable level, we consider this question about school-level course availability only briefly.

Second, for a given school course allocation, how specialized are teachers in the tracks, subjects, and grade levels they teach? What teacher characteristics (e.g., experience and degrees) and school characteristics (e.g., school size and overall course offerings), affect the degree of specialization? In short, how well can we model the sorting of teachers into tracks within schools? We ask similar questions of students, although students, unlike teachers, are only in one track at a time. How well can we predict, through factors such as prior test scores and demographics, which types of courses students will end up in (for a given school-level course allocation)?
3.4 Variation in Tracks across Schools

Figure 1 provides a kernel density plot of the ratio of the number of students to the number of courses in each track in 6th grade math (by school). Notice that the distribution is far more compressed with respect to general track courses. That is, the number of general track courses is explained largely by the number of students in the school, and the curve peaks at around 20 students per general track course. The curve has a long right tail due to schools with higher proportions of remedial and advanced courses and therefore fewer general track courses. These other courses are much more unevenly distributed across schools than the regular courses.

To better understand the variation in course availability, we regressed the student-to-advanced-course ratio from Figure 1 on a variety of factors we hypothesized to be related to it: 5th grade student achievement levels, 5th grade test score variance, and a vector of student demographics (race and income indicators). These are all statistically significant and collectively explain 42 percent of the school-level variation in advanced track courses. Thus, while course availability is predictable, we interpret this as evidence that other factors also play a role, including student course grades from prior years (these are not available in the data), the supply of teachers able to teach advanced courses, and unobserved differences in parents and students.14

3.5 Teacher Sorting into Tracks within Schools

3.5.1 Cross-Sectional Evidence on Tracking

We hypothesize that teachers will tend to specialize in specific grades and tracks because there is a fixed cost to preparing for an additional course (developing lesson plans, learning how to

---

14 An additional factor affecting this distribution is an amendment to Florida’s constitution in 2002 that restricts class sizes to 22 students in elementary and middle school grades. The amendment was worded to allow a gradual implementation, beginning roughly at the beginning of our panel and continuing to nearly full implementation at the end of the panel. Interestingly, the average class size is above this maximum for advanced track courses, even though with a hard cap in every classroom it should be well below 22. This is partly because we have restricted the analysis to schools with more than 10 students and a higher proportion of advanced track courses are below this cutoff.
explain key concepts and procedures, and so on). Table 3 shows the number of teachers who teach only in a given grade/track combination, followed by the number who teach in one other track, two other tracks and so on. Several patterns are worth noting. First, teachers do typically specialize. More than half of the teachers in every row teach either one or two grade/track combinations (see first two columns). This pattern is even more pronounced among teachers who have at least one general track course where, by this definition, 75 percent or more of teachers are specialized. This is apparently because there are simply more general track courses to be taught, making specialization more feasible.

We can further understand the specialization by analyzing which specific other courses teachers teach. Table 4 presents the proportion of teachers of a given course (indicated by the row) who also teach the course indicated by the column (the diagonals are all 1.0 for this reason). We can see that when teachers diversify their teaching, they do so in predictable ways, tending to stay within the same grade. For example, most teachers teaching an advanced course in any grade also teach a regular course within the same grade. A similar, though less pronounced, pattern emerges for teachers in the remedial track. Likewise, very few teachers in any of the 6th grade course tracks also teach in any 8th grade course track, or vice versa. The few exceptions to this arise within tracks; for example, of the 6th grade remedial math teachers, 20 percent also teach 8th grade remedial math.

So, there are two potential deviations from complete specialization. Teachers can diversify across grades, which is somewhat rare because, we theorize, all tracks within a given grade are, at least in theory, operating under the same set of academic standards. Therefore, the cost of adding a course in another grades is even more costly than doing so within a grade because of the need to learn an additional set of academic standards. For this reason, and perhaps others, teachers are typically hired to serve in a particular grade. There is still some diversification across grades because

---

15 Federal law requires that all states to set standards in math, reading, and science, by grade level.
16 The authors wish to thank Debbi Harris for pointing this out.
some schools “loop” students with the deliberate intent of having each student have the same teachers across grades. Also, some schools are too small to allow strict grade-level specialization.

3.5.2 Teacher Sorting Over Time

We are able to explain above a great deal of the cross-sectional variation in the teacher sorting process. However, there is good reason to expect that there are additional important dynamics, with respect to both teachers’ individual careers and teacher turnover within schools. We hypothesize first that teachers sort into courses based on their experience levels. One reason is that more experienced teachers have more authority over their work (including the courses they teach), a fact that is common to organizations in general but especially public sector organizations where the importance of seniority is institutionalized through experience-based salary schedules and tenure rules. Thus, if more experienced teachers prefer teaching advanced courses (ceteris paribus), then we would expect the teachers of advanced courses to have more experience. A second reason we predict such a pattern is that more experienced teachers may be more effective in teaching advanced courses, and parents are likely to place the most pressure on teachers and schools principals for effective instruction in these advance courses. Figure 2 supports our hypothesis, plotting the average proportion of advanced track courses by teacher experience (based on simple pooled cross-sections). As predicted, the proportion of advanced-track courses almost doubles from the lowest to highest years of experience.\footnote{Few teachers have more than 35 years of experience; therefore, we do not view the outliers on the right hand side as significantly affecting our conclusion about the role of experience.}

To better understand the reason teacher experience might affect sorting, we tested whether years of experience is a stronger predictor than experience rank, as the latter should be more aligned with “seniority” whereas small differences in experience are unlikely to indicate important differences in teacher productivity. We find that teacher experience rank within the school is the stronger predictor of the two, supporting the seniority hypothesis.

\footnote{Few teachers have more than 35 years of experience; therefore, we do not view the outliers on the right hand side as significantly affecting our conclusion about the role of experience.}
A related factor likely to affect the teaching specialization is the departure of a teacher. On one hand, our hypothesis that teachers specialize suggests that teachers would tend to continue with their prior courses. However, if teachers also generally prefer teaching advanced track courses, and if seniority plays a role, then we might expect more senior teachers (who are not already completely specialized) to pick up at least some of the advanced track courses opened up by the departing teacher. It is not clear ex ante which of these two tendencies—diversification costs and preferences for better students—will dominate.

Roughly 10 percent of our sample (N=17,072 teachers) were in schools in which at least one math teacher left in the prior year. We regressed the year-to-year change in the proportion of advanced track courses taught on a measure of courses taught in the previous year by departing teachers (separately by track). On average, each of the retained teachers picks up roughly 10 percent of the departed teacher’s advanced track course load. Since the average school has 7 math teachers remaining from the previous year, this implies that 70 percent of the departed teachers’ loads are re-allocated and the remainder is picked up by a new teacher. We carried out the same exercise, changing the dependent variable to the proportion of courses in the middle track.\(^{18}\)

The degree of re-allocation drops when the remaining teacher is more experienced. This suggests that the preference to continue teaching the same courses dominates the preference to teach more advanced courses. As an aside, it is worth pointing out that this strategy may also be optimal from the school’s standpoint: If the returns to experience are grade-track specific as some evidence suggests (Ost, 2010), then having continuing teachers teach the same courses will yield higher productivity. For the new teacher, all the courses will be new to some degree, so this teacher will have remedial productivity (on the average) no matter what they teach.

\(^{18}\) It is also possible that schools change their course availability after a departing teacher leaves. To test this, we regressed the change in the number of advanced track courses in a given school with a departing teacher on the change in the number of departed teachers in the prior year, as well as the change in the number of students in the school.
3.6 Student Sorting into Tracks within Schools

There is more prior evidence on student tracking and this evidence all points to the fact that students sort based on prior test scores and grades, as well as family socio-economic status (Oakes, 1986). Since we have measures of these factors in our data, we expect student sorting into courses to be much more predictable than it was above for teachers. We regressed the student math course track in each grade on a vector of prior year test scores (6th grade track on 5th grade scores, 7th grade track on 5th and 6th grade scores and so on).

Tables 5A and 5B describe the transitions across tracks, combined 6th-7th grade transitions with 7th-8th grade transitions. Because of the nature of the analysis, students who are not observed in at least two contiguous grades drop out of the analysis. The first table shows the proportions of students making the transitions. As expected, the vast majority of students stay in the same track over time. This is less true for the remedial students who, the majority of whom switch to the general track in the subsequent year. (This may have to do with district or state policies related to remediation that we need to explore.) But 83 (75) percent of the general (advanced) track students stay in the same track over time.

Table 5B expresses similar information, but in terms of the number of students in each cell. This highlights the fact that the vast majority of students are in the general track to start with, which implies further that the tendency to stay in the same track applies to nearly all students in the sample. More than 600,000 students were in the general track in both periods compared with only 140,000 who moved in either direction.

3.7 Conclusions and Stylized Facts

We draw several conclusions from this analysis that inform the subsequent theoretical model for the simulation. First, since we know from prior research that more experienced teachers have higher value-added (Harris & Sass, forthcoming), the fact that experienced teachers are more likely
to teach advanced courses means that, on the average, advanced track courses are probably taught by more productive teachers. (The two positive separate correlations do not guarantee a positive combined correlation between teacher productivity and course track, as discussed later.)

Second, we find some support for the theory that a key factor driving teacher sorting is teacher experience, and the seniority and authority that comes with it. This pattern is likely reinforced by the fact that principals prefer placing their most experienced teachers in advanced track courses to minimize complaints from parents. In future work, we will test whether principals tend to assign the highest value-added teachers to advanced track courses when they have the opportunity. If not, then this would suggest that this reinforce the teacher authority hypothesis.

Third, the fixed cost of a new course preparation appears to be quite important. These costs are apparently greater across grades than within grades (across tracks) as there is noticeably less diversification across grades than across tracks within grades. Overall, we find a fairly high degree of specialization.

Fourth, as prior research suggests, the best measureable predictor of students’ course tracks is the prior year test score. We hypothesize that most of the remaining variation is caused by a combination of exogenous variation in course availability that affects students’ opportunities to take advanced and remedial courses and unobserved student and parent characteristics (e.g., motivation).

4. Sorting Model and Data Generating Process

We will use the descriptive information and stylized facts above to create a data generating process that mimics how the students, teachers, and other school actors behave. After discussing the basic notation, we describe a simple static model, followed by a dynamic model. Throughout the section, we report additional results based on the Florida data. This section is very preliminary.
Notation:

- Course $c$’s track: $m_c \in R$
- Teacher $i$’s productivity at time $t$: $v_{it} \in R$
- Teacher $i$’s seniority at time $t$: $s_{it} \in R$. Higher value of $s_{it}$ means closer to the top of the ranking. We assume a positive correlation between teacher’s value-added and seniority, $\rho_{sv} \in (0,1]$.
- Student $j$’s ability or motivation at time $t$: $\gamma_{it} \in R$. Unobservable to econometricians.
- Student $j$’s standardized test score at the end of time $t$: $A_{it} \in R$.

Further, student $j$’s standardized test score at the end of time $t$

$$A_{it} = v_{it} + \gamma_{it} + \eta_{it}$$

$$\eta \sim N(0, \sigma_{\eta})$$

In all of our models, we assume there is no complementarity between teacher productivity and student ability (Lockwood and McCaffrey, 2009)$^{19}$, implying that this is a zero sum game in which some students “win” and get high-value-added teachers and learn more, but these are equally balanced by achievement losses among students who “lose.”

4.1 Static model with no teacher turnover

The agents in the model are middle school students and parents (whom we treat as one unit called “students”), teachers, and school principals. Students make decisions about which courses to take and teachers make decisions only about which courses to teach. Each student must take one course and each teacher teaches one course. For simplicity, there are four students, two teachers, and two courses in each school such that each course has two students; students take standardized

---

$^{19}$ Lockwood and McCaffrey specifically find little evidence that teacher value-added varies based on measureable students characteristics. Even if this is true, it is still possible that the are complementarities between teacher value-added and unmeasured student characteristics.
tests on the first day of class and the last day of class in each school year; the tests perfectly measure their knowledge and skills.

Teachers choose their course in order of seniority, and all teachers choose a higher track courses when it is their turn to select if there are any such courses remaining. When there is a tie in seniority, the decision is made by a coin flip. This leads to $\rho_{ms} \in (0,1]$.

Consistent with prior empirical work, we assume that teacher’s seniority is positively correlated with teacher value-added (Harris and Sass, forthcoming). If higher ability students end up in higher track courses, then this may indirectly create a positive correlation between teacher value-added and student ability. In this case, including teacher experience as a covariate is insufficient to account for the effect of experience on productivity and the effect of experience on seniority ranking. If we assume sufficiently high correlations such that

$$\rho_{vs}^2 + \rho_{ms}^2 > 1$$

then we have $\rho_{vm} > 0$ (Langford, Schwertman, & Owens, 2010). That is, teachers with high value-added will more likely teach high track courses. Notice that (9) is a sufficient condition rather than a necessary one.

Students are initially assigned (by school leaders) to courses based strictly on their prior test scores. If we assume that students have (unobserved) motivation levels such that low-motivation students receive more utility from remedial track courses and high-motivation students receive greater utility from higher track courses. Specifically we assume students choose $m$ as follows:

$$\min_m |y - m - \varepsilon_m|$$

(10)
where $\epsilon_m$ is unobservable (to econometricians) and independent of $m$. The optimal solution is $y = m + \epsilon_m$. This implies the following correlation between student ability and track:

$$
\rho_{ym} = \frac{\text{cov}(y,m)}{\sqrt{\text{var}(y)\text{var}(m)}} = \frac{\text{cov}(m + \epsilon_m, m)}{\sqrt{\text{var}(m)\text{var}(\epsilon_m)\text{var}(m)}} = \frac{\sigma_m^2}{\sqrt{\sigma_m^2 (\sigma_m^2 + \sigma_{\epsilon_m}^2)}} = \frac{1}{\sqrt{1 + (\sigma_{\epsilon_m}/\sigma_m)^2}} \tag{11}
$$

Again if the correlations are high enough $\left(\rho_{ym}^2 + \rho_{ym}^2 > 1\right)$, then there is “positive selection” with high-ability teachers tending to be matched with high-ability students $\left(\rho_{y} > 0\right)$.

Because more motivated students tend to have higher test scores, a failure to measure unobserved motivation means that higher track teachers (who are also high value-added) will appear to have even higher value-added than they really do and make low-value-added teachers look worse than they are. In other words, this would have the effect of increasing the measured dispersion of teacher value-added.

4.2 Dynamic model with teacher turnover

The previous model is a static, one-period model in which teachers never leave. Here we extend the model to allow for the most senior teacher to retire in each period. This teacher is always replaced by a low value-added, low seniority teacher. The continuing teachers are therefore each bumped up the seniority chain.

One implication from such setup is that every remaining teachers will switch to the immediate higher track courses. However we might not observe frequent switching. If we assume teachers receive negative utility from switching courses because of the cost of preparing a new course, then it is possible that the teacher who is most senior in the second period (after being bumped up) may not choose to teach motivated students even if the teacher was teaching less
motivated students. Whether the continuing teacher switches to the higher track depends on whether the utility value of teaching those students exceeds the utility cost of switching courses.

Assume a teacher’s life-cycle includes $T$ periods (or years). At period $t$ the teacher solves the following problem

$$V_t(m_{t-1}) = \max_{m_t} \{m_t - c(m_t - m_{t-1}) + \beta V_{t+1}(m_t)\}$$  \hfill (12)

Here we assume that each period the most senior teacher retires and one least senior new teacher comes in. So at the beginning of the period $t$, the most senior teacher of those remaining gets to pick the course, followed by the second to most senior teacher and so on so forth. The existence of the switching cost $C(\cdot)$ will induce the teacher to switch courses smoothly if we assume the switching cost is a increasing convex function of the distance between two courses (e.g., switching from remedial to general is less costly than switching from remedial to advanced). However, for now, we assume the switching cost is just a fixed cost, independent of the distance between two courses and it has to be paid in every switch, then the teacher would rather switch courses as infrequently as possible. We will never observe a teacher teaching a high track course switches to a remedial track course under this set of assumptions.\(^{20}\)

The maximization problem becomes

$$V_t(m_{t-1}) = \max_{m_t} \{m_t - c \cdot 1(m_t \neq m_{t-1}) + \beta V_{t+1}(m_t)\}$$  \hfill (13)

The benefit of switching is

$$\left( m_t - m_{t-1} \right) + \beta \left[ V_{t+1}(m_t) - V_{t+1}(m_{t-1}) \right]$$  \hfill (14)

\(^{20}\) Another issue is that we assume even if the teacher switches back to one of the courses she/he has taught several years before the fixed cost of switching has to be paid as well. It could be that the curriculum has changed. This is also for technical simplification, otherwise we have to keep track the whole history of the courses taught by each teacher which is very computationally challenging.
Whenever the benefit exceeds the cost the teacher chooses to switch.

New (replacement) teachers select into schools and then teach whatever course they are assigned. If the continuing teacher from the first period decides not to switch, then the replacement teacher is automatically assigned to teach the course vacated by the retiring teacher.

All newly hired teachers have no authority. Perhaps after period 1, these teachers switch to high-ability with some probability, driven be a Markov of process. The process of switching from low- to high-ability is independent of which courses were taught. So this generates our assumption that $\rho_{vs} \in (0,1]$. \(^{21}\)

### 4.2.1 Two-teacher case

Assume there are only 2 teachers and 2 courses, one advanced track and one remedial track. The more senior teacher A teaches the high track course and the less senior teacher B teaches the low track course. With probability $p$ teachers leave the current school (retires or switching to another school). Now assume teacher A leaves and one replacement teacher C comes. Teacher B becomes more senior in the school so she gets to pick the course first according to seniority. The value of teaching the low track course is

$$V_l = m_l + \beta(1-p)EV + \epsilon_l$$ \hspace{1cm} (15)

The value of teaching the high track course is

$$V_h = \frac{m_h}{1-\beta(1-p)} - c + \epsilon_h$$ \hspace{1cm} (16)

The $\epsilon_h$ and $\epsilon_l$ are observed by the teacher but not the econometricians. This error term accommodates the fact that not all the more senior teacher chooses the high track course. We assume $\epsilon_h$ and $\epsilon_l$ follow extreme value distribution. We also assume if the teacher chooses to teach the high

---

\(^{21}\) In the dynamic model, we also have to think about the fact that, if students are assigned to tracks based on their prior scores, then 7th grade assignments will depend on sorting in 6th grade. We may need to model this as well. Perhaps there is some trajectory of assignments such that low-motivation students “pull apart” from high-motivation students over time.
track course, she never switches to teaching the remedial track one. For simplicity we assume infinite horizon.

The expected value is

$$EV = \int \max \left\{ m_l + \beta (1-p)EV + \varepsilon \nu \frac{m_h}{1-\beta (1-p)} - c + \varepsilon_h \right\} d\varepsilon$$  \hspace{1cm} (17)$$

The extreme value distribution implies

$$EV = \sigma_e \ln \left( \exp \left( \frac{m_l + \beta (1-p)EV}{\sigma_e} \right) + \exp \left( \frac{m_h}{1-\beta (1-p)} - c \right) \right)$$  \hspace{1cm} (18)$$

We can numerically solve EV from this nonlinear equation. So the probability of switching is

$$Pr(s) = \frac{\exp \left( \frac{m_h}{1-\beta (1-p)} - c \right)}{\exp \left( m_l + \beta (1-p)EV \right) + \exp \left( \frac{m_h}{1-\beta (1-p)} - c \right)}$$  \hspace{1cm} (19)$$

Assume the fixed cost of switching is a function of teachers’ characteristics, for instance, gender, education, age, and others,

$$c = b_0 + b_1 \exp$$  \hspace{1cm} (20)$$

Notice here we assume $c_t \approx c_{t+1}$ to simplify the computation. If $c_t$ does not include the time-variant variables then this holds with equality.

We estimate $p$ from the Florida data. We also normalize $m_l = 0$. The next step is to estimate $\beta, m_h, \sigma_e, b_0, b_1$ from the maximum likelihood estimation

$$\max \sum_{t=1}^{N} D_s \ln \left( \min \left\{ 1, \frac{N_t}{N_h} \right\} \times Pr(s) \right) + (1 - D_s) \ln \left( 1 - \min \left\{ 1, \frac{N_t}{N_h} \right\} \times Pr(s) \right)$$  \hspace{1cm} (21)$$
where \( D_s = 1 \) if switching from low track to high track, otherwise 0. \( N_h \) is the number of senior teachers and \( N_v \) is the number of vacant high track courses. Remember we assume only two types of course tracks and two types of seniority, high or low for both teachers.

We have to specify a relationship between teacher experience and teacher value-added. Based on prior evidence (Harris and Sass, forthcoming), we assume teacher value-added is produced according to a production function

\[
v_t = f(exp_t) + \varepsilon_{v,t} \\
\varepsilon_{v} \sim N(0, \sigma_{\varepsilon_v})
\]

where \( exp_t \) is the experience and \( \varepsilon_v \) is the iid unobserved component term. We assume a polynomial form of \( f(exp_t) \),

\[
f(exp_t) = b_{10} + b_{11}exp_t + b_{12}exp_t^2 + b_{13}exp_t^3
\]

In this setup teachers cannot really control their value-added since we don’t allow inputs. It could be taken as that this is already in the equilibrium. That is, there are other inputs and the optimal inputs are just functions of the state variable \( v_t \). Substituting optimal inputs into the production function yields the above law of motion for the teachers’ value added.

4.2.3 Estimation and Results

We used data from Florida to estimate the parameters. The parameters of interest are:

1. The distribution of student’s ability, \( \gamma \sim N(0, \sigma_{\gamma}^2) \).
2. Student test score unobservables, \( \eta \sim N(0, \sigma_{\eta}^2) \).
3. Student motivation unobservables, \( \varepsilon_m \sim N(0, \sigma_{\varepsilon_m}^2) \).
4. The parameters of the production function \((b_{10}, b_{11}, b_{12}, b_{13})\) and \(\sigma^2_e\).

5. The standard deviation of the extreme value distribution in the probability of switching \(\sigma_e\).

6. The switching cost \(c(\cdot)\), and the discount factor \(\beta\).

The parameters in 6 could be estimated separately. They can be estimated in the first step and in the second step they are used to generate teachers’ mobility.

We also define these variables from data

- Course tracking, \(m_c\), and derive \(\sigma_m\).
- Teacher seniority, \(s\).
- The value of teaching a course with tracking \(m_c\).
- The probability of leaving \(\pi\).

The estimation involves two steps. In the first step, we estimate parameters of data generating processes which include the probability of leaving \(\pi\) from data, the standard deviation of the extreme value distribution in the probability of switching \(\sigma_e\), and the switching cost \(c(\cdot)\) from the probability of switching \(Pr(s)\) and the maximum likelihood function.

The results, which we emphasize are very preliminary, are summarized in Table 6. The switching cost \(c = b_0 + b_1 \exp\) decreases with the experience. It implies that it is easier for more experienced teacher to switch courses. It could be that it is easier for more experienced teachers to prepare for new courses. In the second step, we estimate the remaining parameters, specifically 1 – 5 listed above. From the production function of the test score, we have

\[
A_{lt} = v_{lt} + \gamma_l + \eta_{lt}
\]

\[
= b_{10} + b_{11} \exp_{lt} + b_{12} \exp_{lt}^2 + b_{13} \exp_{lt}^3 + \epsilon_{v,lt} + \gamma_l + \eta_{lt}
\]

(23)
Normalize $\mu_Y = 0$. The fixed effect estimation gives us unbiased estimation of $\{b_{1i}\}_{i=0}^{3}$. Since we assume that $\varepsilon_{v,t}$ and $\eta_{it}$ are iid, it gives us unbiased estimation of $\{b_{1i}\}_{i=0}^{3}$. Define $\hat{\theta}_{it} = \hat{b}_{10} + \hat{b}_{11} \exp_t + \hat{b}_{12} \exp_t^2 + \hat{b}_{13} \exp_t^3$, and $\hat{\eta}_{it} = \varepsilon_{v,t} + \eta_{it}$. Rewrite the equation as

$$A_{it} = \hat{\theta}_{it} + \gamma_i + \hat{\eta}_{it} \quad (24)$$

Cunha, Flavio, Heckman, and Schennach (2010) show that we can separately identify the distribution of $\gamma_i$ and $\hat{\eta}_{it}$.

Define $\hat{A}_{it} = A_{it} - \hat{\theta}_{it}$. This implies that

$$\text{var}(\hat{A}_{it}) = \sigma^2_Y + \sigma^2_{\hat{\eta}_t} \quad (25)$$

$$\text{cov}(\hat{A}_{it}, \hat{A}_{it+1}) = \sigma^2_Y \quad (26)$$

Given we have data at three time points ($t = 6, 7, 8$), we use the average moments

$$\hat{\sigma}^2_Y = \frac{1}{3} \sum_{t=6}^{8} \text{var}(\hat{A}_{jt}) = 0.520 \quad (27)$$

$$\hat{\sigma}^2_{\hat{\eta}} = \frac{1}{3} \sum_{t=6}^{8} \text{cov}(\hat{A}_{jt}, \hat{A}_{jt+1}) - \hat{\sigma}^2_Y = 0.248 \quad (28)$$

Notice $\sigma^2_{\hat{\eta}} = \sigma^2_{\varepsilon_v} + \sigma^2_{\hat{\eta}}$. We cannot separately identify each component here. Let’s assume $\hat{\sigma}^2_{\varepsilon_v} = \hat{\sigma}^2_{\hat{\eta}} = 0.124$. The only parameter of interest which is not estimated is the distribution of the student motivation unobservable, $\varepsilon_m \sim N(0, \sigma^2_{\varepsilon_m})$.

Again, we emphasize that these results are very preliminary and incomplete. In future work we will complete the parameter estimation and use these results for the data generating process.
6. Conclusion

Monitoring public service sector workers presents a distinctive challenge because workers have preferences over which clients they interact and because the marginal product of client effort to objectively measured outputs can be large relative to the worker’s marginal product. If the client characteristics over which workers have preferences are difficult to observe yet correlated with client marginal products, and if workers have some control over whom they work with, then even sophisticated econometrics may be insufficient to eliminate bias in productivity measures.

In this study, we have considered the particular case of sorting of workers and clients in Florida middle schools. Combined with some, albeit sparse, prior evidence, our results yield important insights into the sorting process in these organizations. In future we work, we will continue forward with both the empirical analysis and the simulation exercise. Ultimately, the goal is to provide evidence on optimal value-added specifications and, more broadly, to provide a sense of how well econometrics can solve the distinctive challenge that monitoring poses in the service sector and particularly public service sector institutions such as schools.
References


32


Figure 1: Distribution of Course Availability Across Schools

Figure 2: Distribution of the Teacher Track Ratios (6th Grade)
Figure 3: Model of Student and Teacher Sorting
### Table 1: Sample Composition

<table>
<thead>
<tr>
<th>Group</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>51</td>
</tr>
<tr>
<td>Low-income*</td>
<td>52</td>
</tr>
<tr>
<td>White</td>
<td>45</td>
</tr>
<tr>
<td>Hispanic</td>
<td>25</td>
</tr>
<tr>
<td>Black</td>
<td>24</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
</tr>
<tr>
<td>Other race</td>
<td>3</td>
</tr>
</tbody>
</table>

* Low income is defined as being eligible for free or reduced price lunch.

### Table 2: Course Tracks

<table>
<thead>
<tr>
<th></th>
<th>Intensive</th>
<th>Regular</th>
<th>Advanced</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of classrooms</td>
<td>11</td>
<td>66</td>
<td>23</td>
<td>100</td>
</tr>
<tr>
<td>Mean class size (standard dev.)</td>
<td>17.9 (5.1)</td>
<td>21.9 (5.8)</td>
<td>22.9 (5.7)</td>
<td>21.7 (5.8)</td>
</tr>
<tr>
<td>Test score* mean</td>
<td>-0.69</td>
<td>-0.3</td>
<td>0.59</td>
<td>0.00</td>
</tr>
<tr>
<td>Test score standard deviation</td>
<td>0.82</td>
<td>0.81</td>
<td>0.67</td>
<td>1.00</td>
</tr>
<tr>
<td>Test score min</td>
<td>-6.4</td>
<td>-6.4</td>
<td>-5</td>
<td>-6.4</td>
</tr>
<tr>
<td>Test score max</td>
<td>3.5</td>
<td>3.5</td>
<td>3.9</td>
<td>3.9</td>
</tr>
</tbody>
</table>

* Test scores are standardized by grade level / year and are computed amongst students who registered for at least one class at the specified tier.
Table 3: Number of Teachers in Each Grade/Track Teaching Courses in Other Grades/Tracks

<table>
<thead>
<tr>
<th></th>
<th>0 Other</th>
<th>1 Other</th>
<th>2 Other</th>
<th>3 Other</th>
<th>4 Other</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th Rem.</td>
<td>106</td>
<td>188</td>
<td>209</td>
<td>35</td>
<td>8</td>
<td>546</td>
</tr>
<tr>
<td>6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>972</td>
<td>1154</td>
<td>393</td>
<td>77</td>
<td>13</td>
<td>2,609</td>
</tr>
<tr>
<td>6th Adv.</td>
<td>177</td>
<td>855</td>
<td>211</td>
<td>44</td>
<td>6</td>
<td>1,293</td>
</tr>
<tr>
<td>7th Rem.</td>
<td>96</td>
<td>243</td>
<td>215</td>
<td>51</td>
<td>10</td>
<td>615</td>
</tr>
<tr>
<td>7&lt;sup&gt;th&lt;/sup&gt;</td>
<td>555</td>
<td>1150</td>
<td>455</td>
<td>93</td>
<td>17</td>
<td>2,270</td>
</tr>
<tr>
<td>7th Adv.</td>
<td>155</td>
<td>757</td>
<td>241</td>
<td>54</td>
<td>6</td>
<td>1,213</td>
</tr>
<tr>
<td>8th Rem.</td>
<td>319</td>
<td>359</td>
<td>226</td>
<td>43</td>
<td>11</td>
<td>958</td>
</tr>
<tr>
<td>8&lt;sup&gt;th&lt;/sup&gt;</td>
<td>744</td>
<td>751</td>
<td>322</td>
<td>63</td>
<td>13</td>
<td>1,893</td>
</tr>
<tr>
<td>8th Adv.</td>
<td>115</td>
<td>377</td>
<td>101</td>
<td>16</td>
<td>1</td>
<td>610</td>
</tr>
<tr>
<td>Col. Total</td>
<td>3,239</td>
<td>5,834</td>
<td>2,373</td>
<td>476</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Proportion of Teachers Teaching in Each Grade (Row) Also Teaching in Another Grade/Subject (Column)

<table>
<thead>
<tr>
<th></th>
<th>6th Rem.</th>
<th>6th</th>
<th>6th Adv.</th>
<th>7th Rem.</th>
<th>7&lt;sup&gt;th&lt;/sup&gt;</th>
<th>7th Adv.</th>
<th>8th Rem.</th>
<th>8&lt;sup&gt;th&lt;/sup&gt;</th>
<th>8th Adv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th Rem.</td>
<td>1.00</td>
<td>0.49</td>
<td>0.24</td>
<td>0.28</td>
<td>0.10</td>
<td>0.03</td>
<td>0.20</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>6&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.10</td>
<td>1.00</td>
<td>0.36</td>
<td>0.03</td>
<td>0.17</td>
<td>0.03</td>
<td>0.05</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>6th Adv.</td>
<td>0.10</td>
<td>0.73</td>
<td>1.00</td>
<td>0.01</td>
<td>0.08</td>
<td>0.10</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>7th Rem.</td>
<td>0.25</td>
<td>0.12</td>
<td>0.03</td>
<td>1.00</td>
<td>0.43</td>
<td>0.20</td>
<td>0.31</td>
<td>0.09</td>
<td>0.02</td>
</tr>
<tr>
<td>7&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.02</td>
<td>0.20</td>
<td>0.05</td>
<td>0.12</td>
<td>1.00</td>
<td>0.37</td>
<td>0.07</td>
<td>0.22</td>
<td>0.03</td>
</tr>
<tr>
<td>7th Adv.</td>
<td>0.01</td>
<td>0.06</td>
<td>0.11</td>
<td>0.10</td>
<td>0.70</td>
<td>1.00</td>
<td>0.05</td>
<td>0.10</td>
<td>0.06</td>
</tr>
<tr>
<td>8th Rem.</td>
<td>0.11</td>
<td>0.14</td>
<td>0.04</td>
<td>0.20</td>
<td>0.17</td>
<td>0.06</td>
<td>1.00</td>
<td>0.26</td>
<td>0.07</td>
</tr>
<tr>
<td>8&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.02</td>
<td>0.15</td>
<td>0.03</td>
<td>0.03</td>
<td>0.26</td>
<td>0.07</td>
<td>0.13</td>
<td>1.00</td>
<td>0.19</td>
</tr>
<tr>
<td>8th Adv.</td>
<td>0.01</td>
<td>0.04</td>
<td>0.04</td>
<td>0.02</td>
<td>0.10</td>
<td>0.12</td>
<td>0.11</td>
<td>0.60</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Notes: First row represents the proportion of teachers teaching at least one course of 6<sup>th</sup> grade remedial math and one course from the respective column. The denominator is the row total from Table 3. Cells are highlighted in yellow (green) if the proportion is in the range of 0.20-0.39 (0.40-0.99).
Table 5A: Student Transitions Across Tracks Over Time (Proportions)

Next Track

<table>
<thead>
<tr>
<th>Current Track</th>
<th>Remedial</th>
<th>General</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remedial</td>
<td>0.35</td>
<td>0.61</td>
<td>0.04</td>
</tr>
<tr>
<td>General</td>
<td>0.03</td>
<td>0.83</td>
<td>0.14</td>
</tr>
<tr>
<td>Advanced</td>
<td>0.01</td>
<td>0.24</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 5B: Student Transitions Across Tracks Over Time (Number of Students)

Next Track

<table>
<thead>
<tr>
<th>Current Track</th>
<th>Remedial</th>
<th>General</th>
<th>Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remedial</td>
<td>7,167</td>
<td>12,560</td>
<td>847</td>
</tr>
<tr>
<td>General</td>
<td>23,654</td>
<td>639,259</td>
<td>110,681</td>
</tr>
<tr>
<td>Advanced</td>
<td>3,132</td>
<td>61,174</td>
<td>194,240</td>
</tr>
</tbody>
</table>
### Table 6: Estimation of Model Parameters

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability of leaving $p$</td>
<td>0.489</td>
</tr>
<tr>
<td>$N = 23,873$</td>
<td>(0.500)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.974</td>
</tr>
<tr>
<td>$m_h$</td>
<td>2.718</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>1.534</td>
</tr>
<tr>
<td>$b_0$</td>
<td>2.032</td>
</tr>
<tr>
<td>$b_1$</td>
<td>-0.079</td>
</tr>
<tr>
<td>Observations</td>
<td>1,766</td>
</tr>
</tbody>
</table>

### Table 7: Returns to Experience in the Florida Data

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\exp(b_{11})$</td>
<td>0.006418**</td>
</tr>
<tr>
<td></td>
<td>(0.003)</td>
</tr>
<tr>
<td>$\exp^2(b_{12})$</td>
<td>-0.000322*</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>$\exp^3(b_{13})$</td>
<td>0.000004</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
</tr>
<tr>
<td>Constant ($b_{10}$)</td>
<td>-0.529947***</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
</tr>
<tr>
<td>Observations</td>
<td>21,166</td>
</tr>
<tr>
<td>Number of N</td>
<td>10,338</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Standard errors in parentheses

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