

## Teachers and Student Achievement in the Chicago Public High Schools

December 2003

Daniel Aaronson  
Federal Reserve Bank of Chicago

Lisa Barrow  
Federal Reserve Bank of Chicago

William Sander  
DePaul University

### Abstract

Using unique administrative data on Chicago public high school students and their teachers, we estimate the importance of teachers for student mathematical achievement. We find that teachers are educationally and statistically important. To be sure, sampling variation and other measurement issues can strongly influence estimates of teacher effects, and, in some cases, account for much of the dispersion in teacher quality. Even after correcting for these problems, we find that one semester with a teacher rated two standard deviations higher in quality could add 0.3 to 0.5 grade equivalents, or 25 to 45 percent of an average school year, to a student's math score performance. Additionally, our teacher quality ratings remain relatively stable for an individual instructor over time, are reasonably impervious to controlling for non-math teachers, and do not appear to be driven by classroom sorting or selective reporting of test scores. After relating our measured teacher effects to the standard observable characteristics of the instructor, we find that traditional human capital and demographic measures, including those used for compensation purposes, explain little of the total variation in teacher quality.

---

We thank the Chicago Public Schools and the Consortium on Chicago School Research at the University of Chicago for making the data available to us. We are particularly grateful to John Easton and Jenny Nagaoka for their help in putting together the data and answering our many follow-up questions. We thank Joe Altonji, Dave Card, Julie Cullen, Rajeev Dehejia, Tom DiCiccio, Eric French, Brian Jacob, Jeff Kling, Steve Rivkin, Ceci Rouse, Doug Staiger, Dan Sullivan, Chris Taber, seminar participants at the University of Chicago, DePaul University, the University of Illinois, the Federal Reserve Bank of Chicago, the ILR-Cornell Institute for Labor Market Policy/Princeton Industrial Relations Section Tenth Annual Policy Conference, and the Urban School Finance conference at UIC for helpful comments and discussions. The views expressed in this paper are the views of the authors and are not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Updated versions of this paper are available by contacting the authors at [daaronson@frbchi.org](mailto:daaronson@frbchi.org) or [lbarrow@frbchi.org](mailto:lbarrow@frbchi.org).

## 1. Introduction

The Coleman Report (Coleman et al. 1966) broke new ground in the empirical estimation of education production functions, concluding that family background and peers were more important than schools and teachers in determining educational outcomes such as test scores and graduation rates. While research since the Coleman Report generally supports the influence of family background, substantiation of the importance of other factors, particularly schools and teachers, has evolved slowly with the release of better data. Today, most researchers agree that schools and teachers matter.<sup>1</sup> However, how much they matter, the degree to which these effects vary across student populations, and whether measurable characteristics such as teacher education and experience affect student educational outcomes continue to be of considerable research and policy interest.

In this study, we use administrative data on students and teachers in Chicago public high schools to estimate the importance of teachers on student test score gains in mathematics, and then relate our measured teacher effects to observable characteristics of the instructors. Our data provide us with a key and unique advantage: the ability to link teachers with students in specific classrooms. In contrast, many other studies are able to match students to the average teacher in a grade or school. In addition, the administrative teacher records allow us to separate the effects of observed teacher characteristics from unobserved aspects of teacher quality.

Consistent with earlier studies, we find that teachers are important inputs in 9<sup>th</sup> grade math achievement. However, a certain degree of caution must be exercised in evaluating teacher

---

<sup>1</sup> Literature reviews include Hanushek (1996,1997,2002) and Greenwald, Hedges, and Laine (1996). A brief sampling of other recent work on teacher effects includes Rivkin, Hanushek, and Kain (2002), Jepsen and Rivkin (2001), Goldhaber and Brewer (1997), Jacob and Lefgren (2002), Angrist and Lavy (2001), and Rivers and Sanders (2002). The earliest studies on teacher quality were hampered by data availability and thus often relied on state or school-level variation. Hanushek, Rivkin and Taylor (1996) show that aggregation can result in flawed estimates of education production function parameters. Moreover, measurement error is compounded by proxies, such as student-teacher ratios and average experience, which do not fully capture the role of an instructor.

quality, as biases related to measurement, particularly from changes in exam scoring and the presence of small populations of students used to identify certain teachers, can critically influence results. Sampling variation, in particular, overstates our measures of teacher dispersion by up to 50 percent, consistent with an evaluation of North Carolina schools by Kane and Staiger (2002). Correcting for sampling error suggests that the variance in teacher quality in the Chicago public high schools is roughly 0.02 to 0.06 grade equivalents. That is, replacing a teacher with one that is rated two standard deviations higher in quality increases math test scores by 0.3 to 0.5 grade equivalents, or 25 to 45 percent of an average school year. Additionally, we show that our results are not likely to be driven by classroom sorting or selective use of test scores and that individual teacher ratings are relatively stable over time, reasonably impervious to controlling for non-math teachers, and consistent across many student subgroups.

Finally, the vast majority of the variation in teacher effects is unexplained by observable teacher characteristics, including those used for compensation. While some teacher attributes, notably undergraduate major, are consistently related to our quality measure, they explain at most 10 percent of the total variation in teacher quality. The teacher attributes come from administrative data, a subset of which determines teacher compensation. These facts highlight the disconnect between teacher pay and productivity, the difficulty in developing compensation schedules that reward teachers for good work based solely on standard administrative data, and the difficulty in prescribing recruitment strategies for hiring quality teachers.

While our study focuses on only one school district over a three-year period, this district serves a large population of minority and lower income students, typical of many large urban districts in the United States. Fifty-five percent of ninth graders in the Chicago public schools are African-American, 31 percent are Hispanic, and roughly 80 percent receive free or reduced-price

school lunch. Similarly, New York City, Los Angeles Unified, Houston Independent School District, and Philadelphia City serve student populations that are 80 to 90 percent nonwhite and roughly 70 to 80 percent eligible for free or reduced-price school lunch (Authors' calculations based on the Common Core of Data, 2001). Therefore, on these dimensions Chicago is quite representative of the school systems that are the focus of U.S. education policy.

## **2. Data and Background on Chicago Public High School Student Performance**

The quality of our data is a major strength of this study. Upon agreement with the Chicago Public Schools (CPS), the Consortium on Chicago School Research at the University of Chicago provided us with administrative records from the city's public high schools. These records include all students enrolled and teachers working in 88 CPS high schools from 1996-97 to 1998-99.<sup>2</sup> We concentrate on the performance of 9<sup>th</sup> graders in this paper.

Apart from offering a large sample of urban school children, the CPS administrative records provide several other useful features that rarely appear together in other studies. First, the student data include a history of pre-high school test scores that can be used as controls for past (latent) inputs. Second, classroom schedule detail allows student-teacher matches at a level that plausibly corresponds with what we think of as a teacher effect. Finally, the teacher records include specifics about human capital and demographics. These data allow us to decompose the total teacher effects into unobservable and observable factors, including those relied on for compensation decisions by the Chicago public school system. Next, we discuss these issues, and describe a few econometric issues related to each.

### *A. Test scores*

---

<sup>2</sup> Of the 88 schools, 6 are so small that they do not meet criteria on sample sizes that we describe below. These schools are generally more specialized, serving students who have not succeeded in the regular school programs.

Information on multiple test scores is vital as important family background measures, particularly income and parental education, are unavailable. While there are various ways to account for the cumulative effect of inputs that we cannot observe, in the results below we rely on estimating a general form of the value-added model of education production. In particular, we estimate the relationship between 9<sup>th</sup> grade math test scores and the variables of interest while controlling for initial achievement as measured by the 8<sup>th</sup> grade test score.

Chicago Public Schools administers the Iowa Test of Basic Skills (ITBS) during the spring in grades 3 through 8 and the Test of Achievement and Proficiency (TAP) exam during the spring for grades 9 and 11.<sup>3</sup> We observe both 8<sup>th</sup> and 9<sup>th</sup> grade test scores for the majority of ninth grade students, as shown in Table 1. The exams are used to measure whether students have achieved the skills that are appropriate for their grade. In fact, in the CPS a minimum grade-equivalent score on the ITBS is set as a requirement for promotion from 8<sup>th</sup> to 9<sup>th</sup> grade.

Restricting the sample to 9<sup>th</sup> graders is not limiting in terms of sample size, as 27,000 to 30,000 students are available per year. Eighth and 9<sup>th</sup> grade test score data are reported for between 75 and 78 percent of the sample, yielding a potential sample of around 64,000 over the three-year period. Our sample drops to 53,000 when we exclude students without 8<sup>th</sup> and 9<sup>th</sup> grade test scores, those without scores in consecutive school years, and those in the top and bottom one percent of score gains.<sup>4</sup>

The administrative files provide data for the math and reading sections of the TAP and ITBS. Unique student identifiers allow score gains to be computed. Table 2 displays descriptive statistics on math test scores from 1993 to 2000. Scores are reported as grade equivalents, a

---

<sup>3</sup> TAP testing was mandatory for grades 9 and 11 through 1998. 1999 was a transition year in which 9<sup>th</sup>, 10<sup>th</sup>, and 11<sup>th</sup> graders were tested. Starting in 2000, TAP testing is mandatory for grades 9 and 10.

<sup>4</sup> We discuss the effect of sample selection based on missing test score data below.

national normalization that assigns grade levels to test score results. For instance, a 9.7 implies that the student is performing at the level of a typical student in the 7<sup>th</sup> month of 9<sup>th</sup> grade.

Since the late 1980s when former Secretary of Education William Bennett called Chicago Public Schools the “worst in the nation,” substantial effort has been made to improve public schools in Chicago. Following the reforms, Hess (1999) and Roderick (2001) document some initial decline in test-score achievement followed by gains, especially in mathematics. This rise can be seen in table 2. From 1993 to 2000, 9<sup>th</sup> grade math test scores rose dramatically, such that the average test score in 2000 is a full year and two-month grade equivalents higher than it was in 1993. Eighth-grade scores have increased more modestly, from 7.5 in 1993 to 8.0 in 2000.

Girls and boys score similarly on the 8<sup>th</sup> and 9<sup>th</sup> grade math tests, with boys scoring about one month of a grade equivalent higher on the 9<sup>th</sup> grade test and girls scoring roughly two months of a grade equivalent higher on the 8<sup>th</sup> grade test. Low-income students, defined as those receiving free or reduced-price school lunch, score only 4 months lower on the 8<sup>th</sup> grade exam but just over one year lower on the 9<sup>th</sup> grade math test. Finally, significant racial and ethnic gaps exist with African-American and Hispanic students scoring between 8 months and one grade equivalent below whites on the 8<sup>th</sup> grade math test and roughly two grade equivalents behind whites on the 9<sup>th</sup> grade math test. Asian students have the highest average scores on 8<sup>th</sup> and 9<sup>th</sup>-grade tests, averaging roughly one year and six months higher on each.

The raw data suggest that racial and income test score gaps rise dramatically between the 8<sup>th</sup> and 9<sup>th</sup> grade. While we expect that higher-ability students may gain more in one year of education than lower-ability students, we also suspect the rising gap may be a function of the different exams. More generally, we are concerned about how differences in the 8<sup>th</sup> and 9<sup>th</sup> grade test score distributions may lead to misleading teacher effect estimates. In Figure 1, we plot

kernel density estimates of the 8<sup>th</sup> and 9<sup>th</sup> grade mathematics test scores. The 9<sup>th</sup> grade scores are skewed right while the 8<sup>th</sup> grade test score distribution is much more symmetric. As a consequence, controlling for 8<sup>th</sup> grade test scores in the regression of 9<sup>th</sup> grade test scores on teacher indicators and other student characteristics may not adequately control for the initial quality of a particular teacher's students. This may lead us to conclude that teachers with better than average students are superior instructors. Throughout the paper, we drop the top and bottom one percent of the students by change in test scores to partly account for this problem. We also discuss additional strategies, including using alternative measures of test scores, accounting for student attributes, and analyzing groups of students by initial ability.<sup>5</sup>

Finally, missing test score data may raise concerns about problems with selection. Approximately 11 percent of 9<sup>th</sup> graders do not have 8<sup>th</sup> grade math test scores and 17 percent do not have a 9<sup>th</sup> grade score. There are several possible explanations for this outcome: students might have transferred from another district, did not take the exam, or perhaps simply did not have scores appearing in the database. According to the administrative records, 86 percent of the students took the TAP (9<sup>th</sup> grade) test, and of this group, we observe scores for 98 percent. Missing data appear more likely for the subset of students who tend to be male, white or Hispanic, older, and designated as having special education status (and thus exempt from the test). Convincing exclusion restrictions are not available to adequately assess the importance of selection of this type.<sup>6</sup> However, later in the paper, we show that our quality measure is not

---

<sup>5</sup> The student controls that are available to us are somewhat limited but include sex, race/ethnicity, age, free and reduced lunch status, and designated guardian. Because of the paucity of family background information, one strategy we take is to use available address information to match census tract level income, adult education, and house value into the data.

<sup>6</sup> If selection is based on potential test score improvements because, for example, schools and teachers are somehow gaming the test score system by reporting only the most improved students' outcomes, we could overstate the impact of teacher quality (e.g. Jacob and Levitt 2001 and Figlio and Getzler 2002). Identification of a selection equation requires an exclusion restriction that is able to predict the propensity to have a test score in the administrative records but is not correlated with the educational production function's error term. There is no obvious candidate.

correlated with missing test scores, suggesting that this type of selection or gaming of the system is not unduly influencing our measure of teacher quality.<sup>7</sup>

### *B. Classroom scheduling and sorting*

The second important feature of our data is the detailed scheduling that allows us to construct the complete history of a student's class schedule while in the CPS high schools. The data include where (room number) and when (semester and period) a class met, the teacher assigned, the title of the class, and the level to which it was taught (i.e. AP, regular, etc.). Furthermore, we know the letter grade received and the number of classroom absences. Because teachers and students were matched to the same classroom, we believe we have more power to estimate teacher effects than is commonly available in administrative records where matching occurs at the school or grade level. Additionally, since we have this information for every student, we are able to calculate measures of peer characteristics in the classroom.

One natural concern in how we estimate teacher quality is whether there are lingering influences from the classroom sorting process. That is, the students with the most or least achievement potential may be purposely placed with certain instructors. The most likely scenario

---

One possibility is to take advantage of the clear difference in absences between test takers and nontakers. Absences is an obvious correlate of test taking since the propensity for being at school must be associated with taking an exam at school on a given day. But, of course, absences also proxy for ambition, drive, ability, and family circumstances. Therefore, we used a factor in school absences, distance to school, that might be uncorrelated with unobserved student ability. Students who live farther from school likely face additional costs associated with getting there. Since this restriction is not appropriate if distance proxies for latent aspects of a family that is willing to travel farther for their school of choice (Cullen, Jacob, and Levitt 2000), we also use a subsample who do not opt out of their neighborhood school. The actual measure used is a three-threshold spline in distance from the student's census tract to the school's census tract. We have also tried distance polynomials of various orders but found it made little difference. These distance variables appear to be useful predictors of the likelihood of 9th grade test score information being available. Point estimates on the Mill's ratio suggest that selection may be positively associated with achievement. Yet, our primary inferences are unaffected by this correction.

<sup>7</sup> Although more explicit teacher cheating as found in Jacob and Levitt (2001) would also lead to biased results, we think that teacher cheating among 9<sup>th</sup> grade teachers is unlikely as this is not a high stakes test. Additionally, trimming the large negative and large positive test score gains is likely to eliminate students with cheating teachers in either 8<sup>th</sup> or 9<sup>th</sup> grade.

involves parental lobbying which may be correlated with expected test score gains. But a school or teacher may also exert influence that results in nonrandom sorting of students.<sup>8</sup>

To evaluate the extent to which students may be sorted based on expected test score gains, we calculate average test score dispersion for the observed teacher assignments and for several counterfactual teacher assignments. In Table 3, we report the degree to which actual within-teacher variance in student pre-9<sup>th</sup> grade performance differs from simulated classrooms that are either assigned randomly or based on test score rank. We use three lagged test score measures for assignment: 8<sup>th</sup> grade test scores, 6<sup>th</sup> to 7<sup>th</sup> grade test score gains, and 7<sup>th</sup> to 8<sup>th</sup> grade test score gains. Each panel reports results for the three fall semesters in our data.<sup>9</sup> The top row of each panel, labeled “observed,” displays the observed average within-teacher variance of these measures. This is the baseline to which we compare the simulations. Each of the four subsequent rows assigns students to teachers based on pre-9<sup>th</sup> grade performance characteristics.

Row (2) displays the average within-teacher variance when students are perfectly sorted across teachers within their original school.<sup>10</sup> Such a within-school sorting mechanism reduces the within-teacher variance to less than one-tenth of the observed analog. In contrast, if we randomly assign students to classrooms within their original school, as shown in row (3), the average within-teacher variance is very close to the observed within-teacher variance. There is virtually no evidence that sorting occurs on past gains, with the observed variances being slightly larger than the simulated. The randomly assigned classrooms based on 8<sup>th</sup> grade scores tend to

---

<sup>8</sup> Informal discussions with a representative of the Chicago public school system suggest that parents have little influence on teacher selection and the process is not based on characteristics of the students, conditional on course level. Furthermore, our use of first-year high school students alleviates concerns since it may be difficult to evaluate new students, particularly on unobservable characteristics.

<sup>9</sup> The estimates for the spring semester are very similar.

<sup>10</sup> For example, within an individual school, there may be 10 teachers, each with classrooms of 20 students. In the simulation, the top 20 students, based on our three pre-9<sup>th</sup> grade measures, would be placed together, the next 20 together, and so forth. The number of teachers and schools, as well as any heterogeneity in classroom size is set equal to that observed in the data.

have within-teacher variances that are roughly 35 percent higher than the observed classrooms. But clearly, the observed teacher variance in lagged math scores is much closer to what we would expect with random sorting of students than what we would expect if students were sorted based on their past test performance.<sup>11</sup> Thus, we are more confident that teacher assignment is close to random and less likely to confound our estimates of teacher effects.

### *C. Teacher records*

Finally, we match student administrative records to teacher administrative records using school identifiers and eight-character teacher codes from the student data.<sup>12</sup> The administrative teacher file contains information on 6,890 teachers in CPS high schools between 1997 and 1999. Although these data do not provide information on courses taught, through the student files, we isolate 1,243 possible mathematics and computer science teachers. This list is further pared by excluding teachers who did not have at least 15 student-semesters during our sample period. These teachers are placed in the same “other” teacher group for estimation. Ultimately, we identify teacher effects for 856 math or computer science instructors, as well as an average effect for those placed in the “other” category. While the student and teacher samples are not as big as those used in some other administrative files, they allow for reasonably precise estimation.

Matching student and teacher records allows us to take advantage of the third feature of the data: the detailed demographic and human capital information supplied from the teacher administrative files. In particular, we can use a teacher's gender, ethnicity, experience, tenure, university attended, college major, advanced degree achievement, and teaching certification to

---

<sup>11</sup> These calculations are done using all levels of courses—honors, basic, regular, etc. Because most classes are “regular,” the results are very similar when we limit the analysis to regular level classes.

<sup>12</sup> Details about the matching are available in the data appendix. As is made clear there, we cannot match all teacher codes in the student data to teacher names in the teacher files.

decompose total teacher effects into those related to common observable traits of teachers and those, such as drive, passion, and connection with students that are unobserved.

Table 4 provides descriptive statistics of characteristics of the 645 teachers we can match to the administrative records. The average teacher is 45 years old and has been in the CPS for 13.3 years. Minority math and computer science teachers are underrepresented relative to the student population, as 37 percent are African-American and 9 percent Hispanic. Almost 85 percent are certified to teach high school, 38 percent are certified to be a substitute, and 10 to 12 percent are certified to teach bilingual, elementary, or special education classes. The majority of math teachers have a Master’s degree and many report a major in mathematics (47 percent) or education (19 percent).<sup>13</sup>

### 3. Basic Empirical Strategy

In the standard education production function, achievement,  $Y$ , of student  $i$  with teacher  $j$  in school  $k$  at time  $t$  is expressed as a function of cumulative own, family, and peer inputs,  $X$ , from age 0 to the current age, as well as, cumulative teacher and school inputs,  $S$ , from grades kindergarten through the current grade:

$$(1) \quad Y_{ijkt} = \beta \sum_{t=-5}^T X_{it} + \gamma \sum_{t=0}^T S_{ijkt} + \varepsilon_{ijkt}$$

The requirements to estimate (1) are substantial. Without a complete set of conditioning variables for  $X$  and  $S$ , omitted variables may bias estimates of the coefficients on observable inputs unless strong and unlikely assumptions about the covariance structure of observables and unobservables are maintained. Thus, alternative identification strategies are typically applied.

---

<sup>13</sup> Nationally, 55 percent of high school teachers have a Master’s degree, 66 percent have an academic degree (e.g. mathematics major), and 29 percent have a subject area education degree (U.S. Department of Education 2000).

A simple approach is to take advantage of multiple test scores. In particular, we estimate a general form of the value-added model by including 8<sup>th</sup> grade test scores as a covariate. Lagged test scores account for the cumulative inputs of prior years while allowing for a flexible autoregressive relationship in test scores. Controlling for past test scores is especially important with this data, as information on the family and pre-9<sup>th</sup> grade schooling is sparse.

The education production model is of the general form:

$$(2) \quad Y_{ijkt} = \alpha Y_{ijkt-1} + \beta X_{it} + \tau T_i + \theta_i + \mu_t + \rho_k + \varepsilon_{ijkt}$$

where  $\theta_i$ ,  $\mu_t$ ,  $\rho_k$  and  $\varepsilon_{ijkt}$  measure the unobserved impact of individuals, time, schools, and white noise. Each element of  $T_i$ ,  $T_{ij}$ , equals the number of semester classes taken with teacher  $j$  in 9<sup>th</sup> grade.  $\tau_j$  is the  $j$ th element of the vector  $\tau$  and represents the effect of one semester spent with teacher  $j$  in a math or computer science class. Relative to equation (1), the impacts of lagged schooling and other characteristics are now captured in the lagged test score measure. This strategy may still mismeasure teacher quality, however. For simplicity, assume that all students have only one teacher for one semester so that the number of student semesters for teacher  $j$  equals the number of students for teacher  $j$ ,  $N_j$ . In this case, estimates of  $\tau_j$  may be

$$\text{biased by } \rho_k + \frac{1}{N_j} \sum_{i=1}^{N_j} \theta_i + \frac{1}{N_j} \sum_{i=1}^{N_j} \varepsilon_{ijkt} .^{14}$$

The school term  $\rho_k$  is typically removed by including measures of the school quality, a common and general form of which is school fixed effects. School fixed effect estimation is useful to control for time-invariant school characteristics that covary with individual teacher quality, without having to attribute the school's contribution to specific measures. However,

---

<sup>14</sup> The time effects are easily captured by year indicators and therefore are not discussed further.

this strategy requires the identification of teacher effects to be based on differences in the number of semesters spent with a particular teacher and teachers that switch schools during our three-year period. For short time periods, such as a single year, there may be little identifying variation to work with. Thus, this cleaner measure of the contribution of mathematics teachers comes at the cost of potentially much identifying variation. For that reason, we show many results without allowing for school fixed effects.

Factors affecting test scores are often attributed to a student's family background. In the context of gains, however, time-invariant qualities are differenced out, leaving only factors that

are changing, such as divorce or a student's introduction to drugs, in  $\frac{1}{N_j} \sum_{i=1}^{N_j} \theta_i$ . Furthermore,

students must be assigned to teachers based on these changes in order to bias our teacher quality estimates.<sup>15</sup> Nevertheless, we test the robustness of our results to the inclusion of observable student, family, and peer traits because they may be correlated with behavioral changes that influence achievement and may account for group differences in gain trajectory, thus easing concerns about test score normalizations.

Finally, as the findings of Kane and Staiger (2002) make clear, the error term

$\frac{1}{N_j} \sum_{i=1}^{N_j} \varepsilon_{ijkt}$  is particularly troubling when fixed effect estimates are based on small populations

(small  $N_j$ ). In this case, sampling variation can overwhelm signal, causing a few good or bad draws to strongly influence the estimated teacher fixed effect. Consequently, the variance of the distribution of estimated  $\tau_j$  is most likely inflated.

---

<sup>15</sup> We do not discount the possibility of this type of sorting, especially for transition schools, which are available to students close to expulsion. School fixed effects pick this up but we also estimate the results excluding these schools.

This problem is illustrated in Figure 2. The chart computes  $\tau_j$  conditional on 8<sup>th</sup> grade math score, year indicators, and student, family, and peer attributes, as described below. What is notable is that the lowest and highest performing teachers are those with the fewest student semesters.  $\sum_i T_{ij}$  represents the number of student semesters taught by teacher  $j$  over the time period examined. As more student semesters are used to estimate the fixed effect, the importance of sampling variation declines and reliability improves. Regressing  $|\tau_j|$  on  $\sum_i T_{ij}$  summarizes this association. Such an exercise has a coefficient estimate of -0.00047 with a standard error of 0.00008, suggesting that number of student semesters is a critical correlate of the magnitude estimated teacher quality. The association declines as we raise the minimum threshold on  $\sum_i T_{ij}$ . Statistical significance disappears when  $\sum_i T_{ij} \geq 200$ .

To address the problem of sampling error, we analytically adjust the variance of  $\hat{\tau}_j$  for the size of the sampling error by assuming that the estimated teacher fixed effect is the sum of the actual teacher effect,  $\tau_j$ , plus some error. We use the mean of the square of the standard error estimates of  $\hat{\tau}_j$  as an estimate of the sampling variance and subtract this from the observed variance of  $\hat{\tau}_j$  to get an “adjusted” variance. We report both the variance of  $\hat{\tau}_j$  and the adjusted variance in the tables below. We also show how these values vary as we increase the minimum evaluation threshold,  $\sum_i T_{ij}$ . Sampling error largely disappears when the minimum is set high enough.<sup>16</sup>

---

<sup>16</sup> Note, however, that excluding teachers with small numbers of students is limiting because new teachers, particularly those for whom tenure decisions are being considered, cannot be examined.

In the section to follow, we present our baseline estimates that ignore the existence of most of these potential biases. Thus, they should be considered naïve. We then report results that attempt to deal with each potential bias. To the extent that real world evaluation might not account for these problems, this exercise could be considered a cautionary tale of the extent to which teacher quality estimates can be interpreted incorrectly.

Finally, we examine whether teacher quality can be explained by demographic and human capital attributes of teachers. Because of concerns raised by Moulton (1986) about the efficiency of OLS estimates in the presence of a school-specific fixed effect and because students are assigned multiple teachers per year, we do not include the teacher characteristics directly in equation (2). Rather, we employ a GLS estimator outlined in Borjas (1987) and Borjas and Sueyoshi (1994). This estimator regresses  $\hat{\tau}_j$  on teacher characteristics  $Z$ :

$$(3) \quad \hat{\tau}_j = \phi Z_j + u_j;$$

The variance of the errors is calculated as the covariance matrix derived from OLS estimates of (3) and the portion of equation (2)'s variance matrix related to the  $\hat{\tau}$  coefficient estimates,  $V$ .

$$(4) \quad \Omega = \sigma_u^2 I_j + V$$

The  $\Omega$  term in (4) is used to compute GLS estimates of the observable teacher effects.

## 4. Results

### A. The distribution of teacher quality

Our naïve baseline estimates of teacher quality are presented in table 5. In column (1) we present details on distribution of  $\hat{\tau}_j$ , specifically the variance and the 10<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, and 90<sup>th</sup> percentiles. We also list the p-value for an F-test of the joint significance of the teacher effects (i.e.  $\tau_k = 0$  for all k) and the p-value for an F-test of the other regressors. Since this is

our most parsimonious specification, the list of regressors is limited to year dummies and the 8<sup>th</sup> grade math score.<sup>17</sup> Clearly, we cannot rule out the importance of confounding changes in family, student, peer, and school influences, as well as random fluctuations in student performance across teachers. Rather, we report these estimates as a baseline for considering the importance of these biases.

Consequently, the estimated range of the teacher fixed effects is quite broad, perhaps implausibly so. The variance of  $\hat{\tau}_j$  is 0.21 with gaps between the 90<sup>th</sup> percentile and 10<sup>th</sup> percentile teacher of over 1 grade equivalent. Furthermore, approximately 0.47 grade equivalents separate average class gains between the 75<sup>th</sup> and 25<sup>th</sup> percentile teacher. An F-test of the joint significance of  $\hat{\tau}_j$  easily rejects no teacher effects at the highest significance level.

The robustness of these results can be explored by tracking the stability of individual teacher quality over time. To do so, we reestimate equation (2) but with t subscripts on  $\tau_j$ .<sup>18</sup> In table 6 we display the resulting transition matrix linking quartile rankings of  $\hat{\tau}_{jt}$  with quartile rankings of  $\hat{\tau}_{jt+1}$ . Quartile 1 represents the lowest 25 percent of teachers, as ranked by the teacher effect estimate, and quartile 4 the highest 25 percent. The table reports each cell's share

---

<sup>17</sup> Naturally, the key covariate in our production functions, regardless of specification, is the 8<sup>th</sup> grade test score. The t-statistic on this variable often exceeds 200. Yet the magnitude of the point estimate is somewhat surprising, in that it is often greater than 1. For example, in our sparsest specification, the coefficient on 8<sup>th</sup> grade test score is 1.30 (0.01). This suggests the math test score time-series may not be stationary. However, this is not likely to be a problem since we are working off of the cross-section. It would become an issue if we were to include longitudinal information on 10<sup>th</sup> or 11<sup>th</sup> grade. Nevertheless, a simple way to deal with nonstationarity is to estimate equation (3) in differenced form:

$$Y_{ikt} - Y_{ikt-1} = \alpha(Y_{it-1} - Y_{it-2}) + \beta(X_{it} - X_{it-1}) + \gamma(W_{ikt} - W_{ikt-1}) + \tau(T_{it} - T_{it-1}) + \varepsilon_{ijt} - \varepsilon_{ijt-1}$$

Such a specification will lead to inconsistent estimates because of the correlation between the error term with the lagged differenced dependent variable, but a common strategy to avoid this problem is to use the twice lagged differenced dependent variable,  $Y_{it-2} - Y_{it-3}$  as an instrument. This IV estimator broadly supports the results presented below.

<sup>18</sup> Of course, this amplifies sampling variability.

of a row's total or the fraction of teachers in quartile  $q$  in year  $t$  that move to each of the four quartiles in year  $t+1$ . If our estimates are consistent with some signal, whether it is quality or something correlated with quality, we would expect masses of teachers on the diagonals of the transition matrix where quality quartiles are constant across years. We expect cells farther from the diagonals to be monotonically less common. Particularly noisy estimates would not be able to reject the pure random assignment result that each cell would contain equal shares of teachers. In this rather extreme case, teachers would be randomly assigned a new quality ranking each year, and the correlation between this year's ranking and next would be 0.

Our results suggest a nontransitory component to the teacher quality measure. Of the teachers in the lowest quality quartile in year  $t$ , 40 percent remain in year  $t+1$ , 28 percent move into quartile 2, 24 percent into quartile 3 and 8 percent into the highest quartile. Of those in the highest quartile in year  $t$  (row 4), 60 percent remain the following year, 24 percent move one category down, and only 16 percent fall to the lowest two quartiles. A chi-square test easily rejects random assignment for each row.<sup>19</sup>

Moreover, we also explored to what extent teachers in the top and bottom deciles of the quality distribution continue to rank there the following year. Of the teachers in the top decile, 46 percent rank there the following year. This is highly significant relative to the random draw scenario whereby 10 percent would again appear in the top decile in consecutive years. However, of those teachers in the bottom decile, only 14 percent remain there the following year. Given our sample sizes, this is not significantly different from the random assignment baseline.

We believe the latter result is partly driven by greater turnover among teachers in the bottom decile. By definition, to appear in our transition matrix, a teacher must be in the

administrative records for two consecutive years. However, our teacher distributions are derived from the full population. Therefore, if poor performing teachers are more likely to leave the school system, this could bias our test; the random draw baseline would no longer be 10 percent. To investigate this possibility, we regress an indicator of whether the teacher appears in the teacher records in year  $t+1$  on whether she is ranked in the top or bottom decile of the quality distribution in year  $t$ .<sup>20</sup> We find that a teacher ranked at the bottom is 26 percent less likely (standard error of 4 percent) than a teacher ranked in the 10th to 90<sup>th</sup> percentile to appear in the administrative records the following year. In contrast, teacher turnover for those in the top decile is no different than turnover for the middle group. Once we account for this turnover behavior, the share of teachers remaining in the bottom decile of the teacher quality distribution is significant at standard levels.<sup>21</sup>

These results emphasize that teacher quality evaluated using parsimonious specifications with little attention to measurement issues still has an important persistent component. In fact, we find it encouraging that there is any signal to gauge. However, the transitory part, which is aggravated by sampling error when looking at estimates based on one year, is also apparent. Furthermore, the magnitude of the estimates is perhaps improbably large.

### *B. The impact of sampling error*

We next consider how sampling error may affect our results. We already attempt to improve the signal-to-noise ratio by throwing out students with grade changes in the extreme tails and by restricting identified teachers to those with more than 15 student semesters.

---

<sup>19</sup> Similarly, regressing contemporaneous teacher quality on lagged teacher quality results in a point estimate of 0.44 (0.05) for 1998 and 0.53 (0.06) for 1999. Limiting it to teachers in all three years, the coefficients (and standard errors) on lagged and twice lagged teacher quality are 0.48 (0.08) and 0.31 (0.08).

<sup>20</sup> Unfortunately, we cannot distinguish quits and layoffs, nor exits out of teaching from exits into other school systems.

<sup>21</sup> The adjustment assumes that the share of teachers that remain at the bottom decile under the random draw baseline is 7.4 percentage points, or 26 percent lower than 10 percentage points.

However, Kane and Staiger (2002) show that more than half of the variance in score gains from small North Carolina schools, which tend to be smaller than our  $\sum_i T_{ij}$ , and one-third of the variance in that state's larger schools are due to sampling variation. Figure 2 emphasizes the susceptibility of our results to these concerns as well.

The row labeled “adjusted variance” in table 5 presents an estimate of the variance of  $\tau_j$  after adjusting for sampling variation as described above. This modification reduces the variance from 0.208 to 0.172, suggesting that 17 percent of the variation in teacher quality arises from sampling error. We can confirm this result simply by adjusting for possible overweighting of unreliable observations. Column (2) reports the distribution of  $\hat{\tau}_j$ , when weighted by  $\sum_i T_{ij}$ .

The weighted variance of the teacher effects drops to 0.155, comparable in size to the adjusted variance reported in column (1). In either case, the main conclusion remains that dispersion in teacher quality is wide and educationally significant.<sup>22</sup>

### *C. The impact of family, student, and peer characteristics*

The results thus far report on specifications that are quite sparse. They do not fully capture heterogeneity in student, family, and peer background that could be correlated with particular teachers. Moreover, little has been done to account for variation introduced by exam normalization differences. To this end, table 7 reports results in which student, family, and peer group characteristics available in the administrative records are included. For comparison purposes, column (1) repeats the findings from table 5. Each column reports unadjusted,

---

<sup>22</sup> We also experimented with the common strategy of using 6<sup>th</sup> grade math scores as an instrumental variable for 8<sup>th</sup> grade scores. Classical measurement error in the 9<sup>th</sup> grade scores affects efficiency not consistency. The F-value on this IV regression is somewhat smaller than in our basic specification but the distribution of the teacher effects is relatively unchanged.

adjusted, and weighted variance estimates, as well as p-values for F-tests of the joint significance of the teacher effects and the other regressors as they are added to the production function.

Column (2) incorporates student characteristics including gender, race, age, designated guardian relationship (mom, dad, stepparent, other relative, or nonrelative), and free and reduced-price lunch eligibility. In addition, we include a measure of the student's average 9<sup>th</sup> grade math class size, as is standard in educational production analysis, as well as controls for whether the student changed high schools or repeats the 9<sup>th</sup> grade.<sup>23</sup> These controls reduce the size of the variance by roughly one-third but the estimates remain large and highly significant.

In column (3) we introduce additional student controls, primarily related to performance, school choice, and peer and neighborhood characteristics. The additional student regressors are the level and subject matter of math classes, the student's cumulative grade point average, class rank, disability status, and whether the school is outside of her residential neighborhood.<sup>24</sup> The neighborhood measures based on Census data for a student's residential census tract include median family income, median house value, and the fraction of adults that fall into five education categories; they are meant to proxy for latent parental influences. Again, like many of

---

<sup>23</sup> Jointly these background measures are quite significant; individually, the sex and race measures are the primary drivers. Female students gain 0.16 (0.01) grade equivalents less than males, and black and Hispanic students gain 0.49 (0.03) and 0.30 (0.03) less than nonblack, nonhispanic students. Accounting for student performance, neighborhood, and peer controls diminishes the racial differences slightly but the female gap nearly doubles. Students whose designated guardian is the father have, on average, 0.10 to 0.20 higher test score gains than do students with other guardians. Math class size has a positive and significant relationship with test scores; however, once we control for additional classroom characteristics, it switches sign and is usually insignificant.

<sup>24</sup> As math teachers can influence the student's study habits and performance outside the math class, our teacher effect estimates might be biased downward by introducing such controls.

We also experiment with additional controls for student ability, including 8<sup>th</sup> grade reading scores, 6<sup>th</sup> and 7<sup>th</sup> grade math scores, and the variance in 6<sup>th</sup>-8<sup>th</sup> grade math scores. When we control for 8<sup>th</sup> grade reading scores, the point estimate on the 8<sup>th</sup> grade math score declines by about 4 percent but the impact on the teacher effects is minimal. Including controls for 6<sup>th</sup>, 7<sup>th</sup>, and 8<sup>th</sup> grade math scores distributes the autoregressive component of math test scores between the three as follows: 0.28 (0.01), 0.42 (0.01), and 0.71 (0.01). However, again, there is no additional effect on our estimated teacher variances. We have also allowed 8<sup>th</sup> grade math test scores to enter in alternative formats, including as a spline and as indicator categories representing four quartiles of performance and indicators of below and above national norm performance. None of these specification choices are important, relative to the simple linear 8<sup>th</sup> grade score control. Finally, including the variance of junior high math scores is an important and interesting dimension of 9<sup>th</sup> grade achievement, but has no collateral impact on the teacher estimates.

the student controls, the value-added framework should account for permanent income gaps but will not account for differences in student growth rates by parental income or education. Finally, the math class peer characteristics include the 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentiles of math class absences, as a means of measuring how disruptive the classroom is, and the same percentiles of 8<sup>th</sup> grade math test scores as a measure of peer ability. Because teacher ability may influence classroom attendance patterns, peer absences could confound our estimates of interest, leading to downward biased teacher quality estimates.<sup>25</sup>

Adding peer and neighborhood covariates reduces the adjusted variance to 0.059, one-third the size of the naïve estimates reported in column (1).<sup>26</sup> Much of the attenuation comes from adding either own or peer performance measures. Nevertheless, regardless of the controls introduced, the dispersion in teacher quality remains high and statistically significant. The F-value of the joint test of the teacher effects drops below 5, to 4.6, only when the full set of controls are included, but remains statistically significant at the highest levels.

Once again, transition matrices for the full control specification clearly reject random quality draws. The quartile-rank matrix is reported in table 8. 40 percent of teachers ranking in the top 25 percent in one year rank in the top 25 percent in the following year. Another 23 percent slip down one category, 21 percent two categories and 16 percent to the bottom category.

---

<sup>25</sup> See Manski (1993) for a methodological discussion and Hoxby (2000) and Sacerdote (2001) for evidence. While we hesitate to place a causal interpretation on the peer measures, there is a statistical association between a student's performance and that of her peers. The point estimates (standard errors) on the 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentile of peer absences are 0.006 (0.005), -0.006 (0.002), and -0.005 (0.001). Including the level and subject of the class and own student's overall performance eliminates the influence of the median peer and cuts the 90<sup>th</sup> percentile student's influence in half. Thus it appears that the main effect is from missed class among the most absent of students. The point estimates on the 10<sup>th</sup>, 50<sup>th</sup>, and 90<sup>th</sup> percentile of 8<sup>th</sup> grade math scores are 0.052 (0.014), 0.205 (0.025), and 0.162 (0.020). These peer measures reduce the student's own 8<sup>th</sup> grade math test score influence by about 10 percent. Including the additional student performance and class type regressors reduces the peer 8<sup>th</sup> grade score estimates to 0.025 (0.013), 0.137 (0.025), and 0.117 (0.020), suggesting that high performers have the most influence on student performance.

<sup>26</sup> Arguably, part of the reduction in variance is excessive, as teachers may affect academic performance through an effect on absences or GPA. That said, eliminating student or peer measures based on 9<sup>th</sup> grade performance has a small impact (0.01).

All other rows are similarly monotonic and reject chi-square tests at standard significance levels.<sup>27</sup>

#### *D. Within-School Estimates*

Within-school variation in teacher quality is often preferred to the between-school variety as it eliminates concerns about school-level factors, including the principal, curriculum, school size or composition, quality of other teachers in the school, and latent family or neighborhood-level characteristics that might influence school choice. Because our results are based on achievement gains, we are generally concerned only with changes in these factors. However, school fixed-effect estimation minimizes potential bias caused by school choice, as only changes in parental demand for school quality confound our teacher estimates. Since all our students are in new schools in the 9<sup>th</sup> grade, many of the school and institutional factors may be relevant. Therefore, restricting the source of teacher variation to within-school differences will result in a more consistent, but less efficient, measure of the contribution of teachers.

Our primary method of controlling for school-level influences is school fixed effects estimation. As mentioned above, identification depends on differences in the intensity of teacher use by students within-schools as well as teachers switching schools during the sample period. We report these results in columns (4) and (5) of table 7. Relative to the analogous columns without school fixed effects the dispersions in teacher quality are similar, although variance magnitude and the precision of the estimates decline somewhat. The correlation of  $\tau_j$  with and without school fixed effects is 0.87. With the full set of controls, the adjusted variance drops from 0.059 (column 3) to 0.040 (column 5), implying that a two standard deviation improvement in teacher quality produces, on average, a 0.4 grade-equivalent test score gain; 0.09 grade

---

<sup>27</sup> 18 and 12 percent of those in the top and bottom deciles remain the next year. 22 and 23 percent rated in the top and bottom deciles in 1997 are still there in 1999. Again, turnover is high among the lowest performing teachers.

equivalents less than the estimate without school fixed effects. Again, an F-test rejects that the within-school teacher quality estimates jointly equal zero at the 1-percent level, although the level of the F-statistic drops to 2.7.

School fixed effect estimation has the distinct advantage of not having to attribute school performance to specific measurable characteristics. Because we are concerned about the loss of identifying variation, however, we alternatively tried controlling for additional characteristics. The controls include the student's 8<sup>th</sup> and/or 9<sup>th</sup> grade reading scores, as well as school-level characteristics—average 9<sup>th</sup> grade math and reading scores, average number of absences, the size of the school, and the type of school (neighborhood, selective, charter, alternative, special education)—to proxy for school quality. Controlling for students' 8<sup>th</sup> and 9<sup>th</sup> grade reading scores or just 9<sup>th</sup> grade reading scores lowers the adjusted variance in teacher quality to 0.047, relatively close to the 0.040 school fixed effects estimate. In contrast, school-level composition controls—school size, average 8<sup>th</sup> grade math and reading scores, average number of absences, and school type—lower the adjusted variance only to 0.051. These results suggest that aggregate school measures, which are commonly used in the older literature, may be a less satisfactory way to proxy for the school-level effects.

#### *E. Additional Robustness Checks*

This section provides additional evidence on the robustness of our results to the gaming of score reporting, sampling variability, test normalization, and the inclusion of other teachers in the math score production function.

#### Cream skimming

One concern is that teachers or schools discourage some students from taking exams because they are expected to perform poorly. For example, a teacher could increase test

performance, and thus her quality ranking, by only having the best students take the exam. In order to evaluate whether this may be occurring, we explored how our estimate of teacher quality relates to the share of a teacher's students missing 8<sup>th</sup> or 9<sup>th</sup> grade test scores. In both cases, the correlation is low (-0.04), opposite in sign to the cream skimming prediction, and not statistically significant.

Another way to game exam results is for teachers or schools to test students whose scores are not required to be reported and then report scores for those students who do well. To examine this possibility, we calculate the correlation between teacher quality and the share of students excluded from exam reporting. In this case, evidence is consistent with gaming of scores; the correlation is positive (0.08) and statistically significant at the 5 percent level. To gauge the importance of this finding for our results, we reran our statistical models dropping all students for whom test scores may be excluded from school and district reporting. This exclusion affected 6,371 students (12 percent of the full sample) but had no substantive impact on our results.

#### Sampling variability: Restrictions on student semester observations

A simple strategy for minimizing sampling variability is to restrict evaluation to teachers with a large number of student semesters. In table 9, we explore limiting assessment of teacher dispersion to teachers with at least 50 or 100 student semesters. We emphasize that a sampling restriction, while useful for its simplicity, can be costly in terms of inference. Obviously, the number of teachers for whom we can estimate quality is reduced. There may also be an issue about how representative the teachers are particularly since we overlook an important sample of teachers, namely new instructors with upcoming tenure decisions. Finally, sampling variation exists with large numbers of students as well, so we would not expect to offset concerns about measurement error completely by simply setting a high minimum  $\sum_i T_{ij}$ .

Columns (1) and (2) report results further increasing the minimum student semesters required to estimate teacher quality. In Panel A we include all covariates from the specification presented in column 3 of table 7. Panel B additionally includes school fixed effects. Using a 50 or 100 student-semester threshold and the full control specification, we find that the adjusted variance is roughly 0.04 without school fixed effects and 0.02 grade equivalents with school fixed effects. In both cases, the teacher effects are jointly statistically significant. Note that increasing the minimum student semesters from 15 to 100 increases the average number of student semesters per teacher from 106 to 196. Consequently, sampling variability drops from 0.032 grade equivalents (0.092 minus 0.059) for the 15 student threshold to 0.007 (0.049 minus 0.042) for the 100 student threshold.

#### More on test score normalization and the undue influence of outliers

The remaining columns of table 9 include several additional attempts to minimize the influence of outlier observations. Column (3) reports findings using national percentile rankings that are impervious to the normalization problem inherent in grade equivalent scores.<sup>28</sup> We find that the adjusted variance of  $\hat{\tau}_j$  is 4.25 percentile points, a result that is statistically and economically significant, broadly consistent with the grade equivalent results, and robust to exploiting only within-school variation.<sup>29</sup>

In the next column we simply trim the top and bottom 3 percent of the distribution of 8<sup>th</sup> to 9<sup>th</sup> grade math test gains from the student sample. We would clearly expect that this sample restriction would reduce the variance, as it eliminates roughly 2,600 students in the tails of the

---

<sup>28</sup> These rankings have the advantage of potentially greater consistency across tests so long as the reference population of test takers is constant. The publisher of the tests, Riverside Publishing, advertises the TAP as being “fully articulated” with the ITBS and useful for tracking student progress.

<sup>29</sup> Just under 2 percent of the sample is left or right censored, of which over 98 percent are at the lowest possible percentile score of 1. Estimates using a tobit to account for this censoring problem result in virtually identical coefficient estimates and estimates of the variance of the  $\hat{\tau}_j$ .

score distribution. Still, the adjusted teacher variance remains large in magnitude and statistically significant at 0.042 grade equivalents.<sup>30</sup>

Finally, we stratify the sample into ability groups based on 8<sup>th</sup> grade math test score and re-estimate the teacher effects within ability group.<sup>31</sup> Low ability students are defined as those in the bottom third of the Chicago public school 8<sup>th</sup> grade score distribution, at or below 7.5 grade equivalents. As reported in the final three columns of table 9, low ability students have a mean 9<sup>th</sup> grade score of 7.1 and a mean test score change of 0.54. High ability students are in the top third of the 8<sup>th</sup> grade test score distribution with scores above 8.7 (i.e. performing at or above national norms). These students have mean 9<sup>th</sup> grade scores of 11.8 and mean test score growth of 2.2 grade equivalents. All other students are classified as “middle” ability. The middle group has an average 9<sup>th</sup> grade test score of 8.7 grade equivalents and a mean change of 0.67. Looking at subgroups of students with more similar initial test scores should help reduce the possibility that teacher effect estimates are simply measuring test score growth related to normalization issues. As such, it can be considered another test of the robustness of the results. Moreover, it is of independent interest to document the effect of teachers on different student populations, particularly those achieving at the lowest and highest levels. The major drawback, of course, is that by limiting the sample to a particular subgroup we exacerbate the small sample size problem in estimating teacher quality.

Among all ability groups, we attribute large shares of the variance in estimated teacher effects to sampling variability. That said, a two standard deviation improvement in teacher

---

<sup>30</sup> We have also estimated the education production function using the robust estimator developed by Huber to account for outliers. The technique weights observations based on an initial regression and is useful for its high degree of efficiency in the face of heavy-tailed data. These results generate an even wider distribution of estimated teacher quality and are not reported in the paper.

<sup>31</sup> We have also stratified the sample by sex and race. Teacher dispersion is higher for male than female students by nearly 30 percent. By race/ethnicity, the adjusted variance is 0.056 for African-Americans and 0.055 for Hispanics. Unfortunately, small sample sizes preclude reliable estimation for other races.

quality is still worth a sizable gain in average test score growth, particularly among the middle and low achieving populations. A two standard deviation increase in teacher quality for one semester raises 9<sup>th</sup> grade test score performance by 0.28, 0.48, and 0.43 grade equivalents for low, middle, and high ability students. These are 52, 71, and 19 percent of average test score gains from 8<sup>th</sup> to 9<sup>th</sup> grade for each group.<sup>32</sup>

#### Including other teachers in the production function

We explore one final specification that takes advantage of the detailed classroom scheduling in our data by including a full set of English teacher semester counts, akin to the math teacher semester count,  $T_i$ , in equation (2). Assuming the classroom sorting mechanism is similar across subject areas (*e.g.*, parents who demand the best math teacher will also demand the best English teacher or schools will sort students into classrooms and assign classes to teachers based on the students' expected test score gains), the English teachers will pick up some sorting that may confound estimates of  $\tau$ . Moreover, the English teachers may help us gauge the importance of teacher externalities, *i.e.*, the proverbial superstar teacher who inspires students to do well not just in their class but in all classes. In the presence of student sorting by teacher quality, these spillover effects will exacerbate the bias in the math teacher quality estimates. Although we cannot separately identify classroom sorting from teacher spillovers, we are primarily interested in testing the robustness of our math teacher effects to such controls. Recall, however, that the table 3 results suggest classroom sorting is fairly minimal.

We report estimates including English teachers in table 10. For additional comparison, we also report variances of the English teacher effect estimates both with and without controls for the math teachers. Controlling for English teachers, the math teacher adjusted variance is

---

<sup>32</sup> Although not related directly to the teacher effects, the dynamics of the test scores differ across groups as well. The autoregressive component of math scores is substantially lower for the lowest achieving students (around 0.47)

somewhat smaller and less precisely estimated. That said, the dispersion in English teacher quality, at least in terms of their effect on math scores, is less than one-third that of math teachers. This is consistent with our expectation math teachers are a more important input in math achievement than are English teachers. Thus, we conclude that our results are robust to controls for additional teachers.<sup>33</sup>

## 5. Predicting Teacher Quality Based on Resume Characteristics

This final section relates our estimates of  $\tau_j$  to measurable characteristics of the instructors available in the CPS administrative records. Observable characteristics include common demographic and human capital characteristics such as teachers' gender, race, potential experience, tenure at CPS, advanced degrees (Masters or Ph.D.), undergraduate major, undergraduate college attended, and teaching certifications. We report select results in table 11. All are based on the full control specification reported in column 3 of table 7. We discuss common themes below.

First and foremost, the vast majority of the total variation in teacher quality is unexplained by observable teacher characteristics. At one extreme, a cubic in tenure and indicators for advanced degrees and teaching certifications explains at most 3 percent of the total variation, adjusting for the share of total variation due to sampling error.<sup>34</sup> That is, the characteristics on which compensation is based have extremely little power in explaining teacher

---

relative to middle and high ability students (1.3 and 1.4).

<sup>33</sup> Alternatively, we can substitute reading scores for math scores in a production function with both English and math teachers. In this case, the adjusted variances for math teachers equals 0.029 and for English teachers equals 0.020. The table 10 results are based on a specification without school fixed effects. Including school fixed effects, the variance of the math teacher effects is slightly smaller at 0.047, and the variance on the English teacher effects is somewhat smaller at 0.024. Precision is also affected by the limitation to within-school variation although both the math and English teacher effects are jointly significant at the 1 percent level.

<sup>34</sup> The  $R^2$  is an understatement of the explanatory power since up to 50 percent of the variation in  $\hat{\tau}_j$  is due to sampling error. If we simply multiply the total sum of squares by 50 percent to account for this, the  $R^2$  will double. However, it still remains below 10 percent in all cases.

quality dispersion. Including other teacher characteristics, changing the specifications for computing the teacher effects, and altering the minimum student-semester threshold have little impact on this inference. In all cases, the  $R^2$  barely exceeds 0.05.

Given a lack of compelling explanatory power, it is of little surprise that few human capital regressors are associated with teacher quality. A notable exception is math or science undergraduate degrees, which are associated with teacher quality of 0.06 to 0.08 grade equivalents higher.<sup>35</sup> The majority of standard education background characteristics, including certifications, advanced degrees, and graduating from a top university, are loosely, if at all, related to  $\hat{\tau}_j$ .<sup>36</sup>

Experience and tenure have little relation to  $\tau_j$  when introduced in levels (unreported) or higher order powers. In column (3) we look specifically at teachers with less than one or exactly one year of potential experience on average over the three years compared to teachers with more potential experience. Again we find no statistically significant difference. That said, teachers with less than one year of potential experience are associated with estimated quality that is 0.031

---

<sup>35</sup> Other studies that correlate specific human capital measures to teacher quality are mixed. Hanushek (1971) finds no relationship between teacher quality and experience or master's degree attainment. Rivkin, Hanushek, and Kain (2002) also find no link between education level and teacher quality, although they find a small positive relationship between the first two years of teacher experience and teacher quality. Summers and Wolfe (1977) find that student achievement is positively related to the teacher's undergraduate college while student achievement is negatively related to the teacher's test score on the National Teacher Examination test. In contrast, Hanushek (1971) finds that teacher verbal ability is positively related to student achievement for students from "blue-collar" families. Ferguson (1998) argues teacher test score performance is the most important predictor of a teacher's ability to raise student achievement. Goldhaber and Brewer (1997) find some evidence that teacher certification in mathematics or majoring in mathematics is positively related to teacher quality. Other work by Jacob and Lefgren (2002) and Angrist and Lavy (2001) find no evidence that human capital investment in the form of teacher in-service training influence student achievement.

<sup>36</sup> Bilingual certification is associated with lower student gains. However, this result is likely related to the difficulty of teaching children with English as a second language rather than an indictment of the certificate itself. Our inability to identify potentially important contributors to achievement such as student's native language with the data in hand is a problem with most administrative records. The bilingual result disappears when we just look at within-school variation suggesting that bilingual students are concentrated in particular schools.

Our data include the name of the undergraduate university. We aggregate universities into six categories, based on *U.S. News and World Reports'* rankings.

(standard error of 0.052) grade equivalents lower than teachers with more than one year of potential experience.<sup>37</sup>

Finally, the race/ethnicity of the teacher has no significant effect on overall student achievement although female teachers are associated with test scores roughly 0.045 grade equivalents higher. Moreover, we find little compelling evidence (unreported) that students perform better or worse with teachers that “look like them” with the exception of African-American male students.<sup>38</sup> For African-American male students, African-American teachers are associated with math test scores that are 0.11 (standard error of 0.04) grade equivalents higher; there is no statistically significant difference for African-American female students.

## **6. Conclusion**

The primary implication of our results is that teachers matter. While this has been obvious to those working in the school systems, it is only in the last few years that social scientists have had access to data necessary to verify and estimate the magnitude of these effects. In spite of the improved data, the literature remains somewhat in the dark about what makes a good teacher. Our results are consistent with related studies like Hanushek (1992) and Rivkin, Hanushek, and Kain (2002) who argue that unobservables are driving much of the dispersion in teacher quality. Traditional human capital measures have few robust associations with measures of teacher quality and explain a very small fraction of its wide dispersion. That our teacher measure has an autoregressive component implies that principals may eventually be able to identify quality; however, they are unlikely to have information on teacher quality when recruiting or for recent hires, where quality may be poorly inferred due to sampling variability.

---

<sup>37</sup> Combining the 0 and 1 year categories results in an estimate of -0.011 (0.023). When we include a similarly constructed measure of tenure, we find no independent effect of being new to the school system.

Moreover, while it is often argued that low achievement in Chicago is a result of inadequate resources (e.g. Kozol 1991), it is unclear that more money would have a large impact unless it is directed in the proper manner (Hanushek, Kain, and Rivkin 1999). One common proposal is to tie teacher pay more directly to performance, rather than the current system, which is based on measures that are unrelated to student achievement, namely, teacher education and tenure. That said, such a compensation scheme would require that serious attention be paid to important measurement problems associated with identifying quality.

---

<sup>38</sup> Goldhaber and Brewer (1997) find teacher quality higher among female and lower among African-American instructors. Ehrenberg, Goldhaber, and Brewer (1995) and Dee (2001) also look at teacher race and/or sex but instead focus on whether students perform better with teachers of their own race and/or sex.

## References

- Angrist, Joshua and Victor Lavy, 2001, "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools," *Journal of Labor Economics* 19:343-369.
- Borjas, George, 1987, "Self-Selection and the Earnings of Immigrants," *American Economic Review* 77:531-553.
- Borjas, George and Glenn Sueyoshi, 1994, "A Two-Stage Estimator for Probit Models with Structural Group Effects," *Journal of Econometrics* 64:165-182.
- Coleman, James S., et al. 1966, *Equality of Educational Opportunity*, Washington, D.C.: U.S. Government Printing Office.
- Cullen, Julie, Brian Jacob, and Steven Levitt, 2000, "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools," *Journal of Public Economics*, forthcoming.
- Dee, Thomas, 2001, "Teachers, Race, and Student Achievement in a Randomized Experiment," working paper, Swarthmore College.
- Ehrenberg, Ronald, Daniel Goldhaber, and Dominic Brewer, 1995, "Do teachers' race, gender, and ethnicity matter?" *Industrial and Labor Relations Review* 48:547-561.
- Ferguson, Ronald, 1998, "Paying for public education," *Harvard Journal of Legislation* 28:465-498.
- Figlio, David and Larry Getzler, 2002, "Accountability, Ability, and Disability: Gaming the System," Working paper, University of Florida.
- Goldhaber, Dan and Dominic Brewer, 1997, "Why don't school and teachers seem to matter?" *Journal of Human Resources* 32: 505-523.
- Greenwald, Rob, Larry Hedges, and Richard Laine, 1996, "The Effect of School Resources on Student Achievement," *Review of Educational Research* 66:361-396.
- Hanushek, Eric, 1971, "Teacher characteristics and gains in student achievement," *American Economic Review* 61: 280-288.
- Hanushek, Eric, 1992, "The Trade-off Between Child Quantity and Quality," *Journal of Political Economy* 100:84-117.

- Hanushek, Eric, 1996, "Measuring Investment in Education," *Journal of Economic Perspectives* 10:9-30.
- Hanushek, Eric, 1997, "Assessing the effects of school resources on student performance: An update," *Education Evaluation and Policy Analysis* 19:141-164.
- Hanushek, Eric, 2002, "Publicly Provided Education," In Auerbach and Feldstein (eds), *Handbook of Public Finance*, Amsterdam: North-Holland.
- Hanushek, Eric, Steven Rivkin, and Lori Taylor, 1996, "Aggregation and the Estimated Effects of School Resources," *Review of Economics and Statistics*, 78:611-627.
- Hanushek, Eric, John Kain, and Steven Rivkin, 1999, "Do Higher Salaries Buy Better Teachers?" Working paper, University of Texas at Dallas.
- Hess, G. 1999, "Understanding Achievement (and Other) Changes Under Chicago School Reform," *Educational Evaluation and Policy Analysis*, 21:67-83.
- Hoxby, Caroline, 2000, Peer Effects in the Classroom: Learning from Gender and Race Variation," NBER working paper 7867.
- Jacob, Brian and Lars Lefgren, 2002, "The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago," NBER Working Paper No. 8916.
- Jacob, Brian and Steven Levitt, 2001, "'Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, forthcoming.
- Jepsen, Christopher and Steven Rivkin, 2001, "What is the Tradeoff Between Smaller Classes and Teacher Quality," Working paper, Public Policy Institute of California.
- Kane, Thomas and Douglas Staiger, 2002, "The Promises and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives* 16:91-114.
- Kozol, Jonathan, 1991, *Savage Inequalities*. New York: Crown Publishers.
- Manski, Charles, 1993, "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies* 40:531-542.
- Moulton, Brent, 1986, "Random Group Effects and the Precision of Regression Estimates," *Journal of Econometrics* 32:385-397.
- Rivers, June and William Sanders, 2002, "Teacher Quality and Equity in Educational Opportunity: Findings and Policy Implications," in Izumi and Evers (eds.), *Teacher Quality*, Stanford, CA: Hoover Institution Press.

- Rivkin, Steven, Eric Hanushek, and John Kain, 2002, "Teachers, schools, and academic achievement," Working paper, University of Texas at Dallas.
- Roderick, Melissa, 2001, "Educational Trends and Issues in the Region, the State and the Nation," In L. B. Joseph (ed.), *Education Policy for the 21st Century: Challenges and Opportunities in Standards-Based Reform*, Chicago: University of Illinois Press.
- Sacerdote, Bruce, 2001, "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *Quarterly Journal of Economics* 116:681-704.
- Summers, Anita and Barbara Wolfe, 1977, "Do schools make a difference?" *American Economic Review* 67:639-652.
- U.S. Department of Education, National Center for Education Statistics, *The Condition of Education 2000*, NCES 2000-062, Washington, D.C.: U.S. Government Printing Office, 2000.

## Data Appendix

The student administrative records assign an eight-character identification to teachers. The first three characters are derived from the teacher's name (often the first 3 characters of the last name) and the latter five reflect the teacher's "position number" which is not necessarily unique. In the administrative student data, several teacher codes arise implausibly few times. When we can reasonably determine that the teacher code contains simple typographical errors, we recode it in the student data. Typically, we will observe identical teacher codes for all but a few students in the same classroom, during the same period, in the same semester, taking the same subject, and a course level other than special education. These cases we assume are typographical errors. Indeed, often the errors are quite obvious, as in the reversal of two numbers in the position code.

A second problem we face in the teacher data occurs because a teacher's position and school number may change over time. We assume that administrative teacher records with the same first and last name as well as the same birth date are the same teacher and adjust accordingly. Finally, we face the problem of matching the teacher codes in the student-level administrative records to the administrative data on teachers. We first match students to teachers using the school number, a three-letter name code, and the position number for the combinations that are unique in the teacher data.<sup>39</sup> Next, we match students to teachers using the school number, the first letter of the last name, and the position number, again for unique combinations. Next, we match students and teachers using unique school number and position number combinations, and finally we match any remaining students and teachers using unique combinations of school number and the first three letters of the last name.

---

<sup>39</sup> Note, some three-letter teacher codes were assigned manually for cases in which the teacher code did not correspond to the first 3 letters of the teacher's last name.

**Table 1**  
**Descriptive Statistics for the Student Data**

	<u>All Students</u>		<u>Students with 8th and 9th grade math scores</u>		<u>Students with 8th and 9th grade test scores 1 year apart</u>	
	<u>Mean</u>	<u>Std Dev</u>	<u>Mean</u>	<u>Std Dev</u>	<u>Mean</u>	<u>Std Dev</u>
<u>Sample Size</u>						
Total	84,190		64,457		52,991	
1997	29,302		21,992		17,941	
1998	27,340		20,905		16,936	
1999	27,548		21,560		18,114	
<u>Test scores (grade equivalents)</u>						
Math, 9th grade	9.07	2.74	9.05	2.71	9.21	2.64
Math, 8th grade	7.75	1.55	7.90	1.50	8.07	1.41
Math change, 8th to 9th grade	1.15	1.89	1.15	1.89	1.14	1.75
Reading comprehension, 9th grade	8.50	2.94	8.50	2.89	8.63	2.88
Reading comprehension, 8th grade	7.64	1.94	7.82	1.88	8.01	1.80
Reading change, 8th to 9th grade	0.66	2.02	0.67	2.02	0.62	1.95
<u>Demographics</u>						
Age	14.8	0.8	14.7	0.8	14.6	0.7
Female	0.497	0.500	0.511	0.500	0.522	0.500
Asian	0.035	0.184	0.033	0.179	0.036	0.185
African-American	0.549	0.498	0.571	0.495	0.562	0.496
Hispanic	0.311	0.463	0.304	0.460	0.307	0.461
Native American	0.002	0.047	0.002	0.046	0.002	0.046
Eligible for free school lunch	0.703	0.457	0.721	0.448	0.728	0.445
Eligible for reduced-price school lunch	0.091	0.288	0.097	0.295	0.103	0.303
Legal Guardian:						
Dad	0.241	0.428	0.244	0.429	0.253	0.435
Mom	0.620	0.485	0.627	0.484	0.619	0.486
Nonrelative	0.041	0.197	0.039	0.195	0.037	0.189
Other relative	0.038	0.191	0.034	0.182	0.032	0.177
Stepparent	0.002	0.050	0.002	0.047	0.002	0.046
<u>Schooling</u>						

Take algebra	0.826	0.379	0.866	0.341	0.952	0.213
Take geometry	0.099	0.299	0.092	0.288	0.022	0.146
Take computer science	0.003	0.054	0.003	0.057	0.003	0.057
Take calculus	0.0001	0.011	0.0001	0.010	0.0001	0.008
Fraction honors math classes	0.082	0.269	0.094	0.286	0.102	0.297
Fraction regular math classes	0.824	0.359	0.827	0.355	0.821	0.360
Fraction essential math classes	0.032	0.173	0.029	0.163	0.032	0.172
Fraction basic math classes	0.001	0.036	0.001	0.031	0.001	0.034
Fraction special ed. math classes	0.014	0.115	0.009	0.093	0.009	0.092
Fraction nonlevel math classes	0.006	0.057	0.006	0.055	0.006	0.058
Fraction level missing math classes	0.041	0.163	0.035	0.142	0.029	0.119
Fraction of math grades that are A	0.084	0.254	0.085	0.254	0.094	0.264
Fraction of math grades that are B	0.131	0.293	0.140	0.299	0.153	0.308
Fraction of math grades that are C	0.202	0.346	0.220	0.353	0.234	0.356
Fraction of math grades that are D	0.234	0.366	0.250	0.372	0.252	0.366
Fraction of math grades that are F	0.309	0.426	0.270	0.406	0.238	0.382
Fraction of math grades missing	0.041	0.163	0.035	0.142	0.029	0.119
Number of math/CS classes taken in 9th grade	2.1	0.4	2.1	0.4	2.1	0.4
Number of times in 9th grade	1.17	0.41	1.13	0.36	1.00	0.00
Changed school within the year	0.034	0.180	0.030	0.170	0.027	0.163
Average class size among 9th grade math classes	22.7	7.5	23.2	7.4	23.6	7.5
Cumulative GPA, Spring	1.71	1.08	1.82	1.04	1.93	1.03
Average absences in 9th grade math	13.7	16.6	11.4	13.5	9.7	11.4
Identified as disabled	0.021	0.143	0.024	0.154	0.022	0.147

Notes: The share of students disabled does not include students identified as learning disabled. Roughly 9 percent of CPS students in our estimation sample are identified as learning disabled.

**Table 2**  
**Means and Standard Deviations for Math Test Score Data**  
**Over Time and for Various Subgroups**

	TAP			ITBS		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.
1993-2000 Sample	180,636	8.77	2.68	204,177	7.65	1.53
1997-99 Sample	69,639	9.07	2.74	75,089	7.75	1.55
Year <sup>a</sup>						
1993	21,893	8.52	2.48	24,889	7.45	1.53
1994	21,761	8.07	2.45	25,388	7.55	1.44
1995	21,884	8.49	2.50	26,120	7.49	1.40
1996	22,600	8.08	2.51	27,093	7.44	1.45
1997	23,850	8.79	2.70	25,866	7.55	1.49
1998	22,570	8.89	2.58	24,329	7.85	1.54
1999	23,219	9.53	2.85	24,894	7.87	1.60
2000	22,859	9.72	2.80	25,598	8.01	1.68
Sex of Student ('97-'99)						
Male	34,083	9.11	2.83	37,661	7.68	1.63
Female	35,556	9.04	2.64	37,428	7.83	1.47
Income <sup>b</sup> ('97-'99)						
Low	56,240	8.84	2.54	60,067	7.68	1.50
High	13,399	10.06	3.25	15,022	8.03	1.71
Race/Ethnicity ('97-'99)						
White	6,923	10.93	3.21	6,578	8.57	1.67
African-American	38,852	8.53	2.40	43,484	7.51	1.49
Asian	2,477	12.04	3.29	2,272	9.25	1.62
Native American	150	10.47	3.20	165	8.27	1.70
Hispanic	21,237	9.11	2.57	22,590	7.82	1.45

Notes: Authors' calculations from the Chicago Public School District administrative student data for students enrolled in 9th grade from the 1992-93 through 1999-2000 academic years. Test scores are reported in terms of grade equivalents. Average TAP scores refer to the math portion of the Test of Achievement and Proficiency administered in the 9th grade. Average ITBS scores refer to 9th graders' 8th grade test scores on the math portion of the Iowa Test of Basic Skills.

<sup>a</sup> Year refers to the Spring of the students' 9th grade school year.

<sup>b</sup> Low income is defined as receiving free or reduced price school lunch.

**Table 3**  
**Mean Variance by Teacher of Lagged Student Test Score Measures**

	8th grade scores	6th to 7th change	7th to 8th change
	Fall 1997		
Observed	1.172	0.466	0.508
Perfect sorting across teachers w/in school	0.099	0.029	0.038
Randomly assigned teachers w/in school	1.603	0.430	0.476
Perfect sorting across teachers	0.000	0.001	0.001
Randomly assigned teachers	2.217	0.424	0.468
	Fall 1998		
Observed	1.338	0.441	0.587
Perfect sorting across teachers w/in school	0.148	0.037	0.075
Randomly assigned teachers w/in school	1.806	0.433	0.581
Perfect sorting across teachers	0.001	0.001	0.002
Randomly assigned teachers	2.404	0.424	0.569
	Fall 1999		
Observed	1.172	0.466	0.508
Perfect sorting across teachers w/in school	0.099	0.028	0.034
Randomly assigned teachers w/in school	1.603	0.430	0.476
Perfect sorting across teachers	0.001	0.001	0.008
Randomly assigned teachers	2.674	0.457	0.717

Notes: In each cell, we report the average variance by teacher for the lagged math test measure reported at the top of the column when students are assigned to teachers based on the row description. Observed calculates the average variance for the observed assignment of students to teachers. Perfect sorting assigns students to teachers either within school or across schools based on the test score measure at the top of the column. Randomly assigned teachers sorts students into teachers either within or across schools based on a randomly generated number from a uniform distribution. The random assignments are repeated 100 times before averaging across all teachers and all random assignments. The top panel reports averages for the Fall of 1997, the middle panel for 1998 and the bottom panel for 1999. Calculations for the spring semesters are very similar.

**Table 4**  
**Descriptive Statistics for the Math Teachers Matched to**  
**Teachers in the Student Data**

	Mean	Std. Dev.
<u>Demographics</u>		
Age	45.0	10.5
Female	0.529	0.500
African-American	0.372	0.484
White	0.467	0.499
Hispanic	0.091	0.289
Asian	0.060	0.239
Native American	0.009	0.096
<u>Human capital</u>		
BA major education	0.186	0.389
BA major all else	0.262	0.440
BA major math	0.474	0.500
BA major science	0.078	0.268
BA university, US News 1	0.088	0.284
BA university, US News 2	0.078	0.268
BA university, US News 3	0.153	0.361
BA university, US News 4	0.078	0.268
BA university, US News 5	0.016	0.124
BA university, US News else	0.566	0.496
BA university missing	0.022	0.146
BA university local	0.597	0.491
Master's degree	0.511	0.500
Ph.D.	0.014	0.119
Certificate, bilingual education	0.117	0.322
Certificate, child	0.019	0.137
Certificate, elementary	0.098	0.298
Certificate, high school	0.839	0.368
Certificate, special education	0.109	0.312
Certificate, substitute	0.380	0.486
Potential experience	20.4	10.3
Tenure at CPS	13.3	10.0
Tenure in position	5.6	6.0
Percent time in position	1.0	0.0
<u>Number of Observations</u>		645

Notes: There are 856 teachers identified from the student estimation sample that have at least 15 student-semesters for math classes over the 1997-1999 sample period. The descriptive statistics above apply to the subset of these teachers that can be matched to the teacher administrative records from Chicago Public Schools.

**Table 5**  
**Distribution of the Estimated Teacher Effects**

Distribution of teacher fixed effects:	Unweighted	Weighted
10th percentile	-0.42	-0.36
25th percentile	-0.26	-0.23
50th percentile	-0.07	-0.07
75th percentile	0.21	0.16
90th percentile	0.67	0.57
90-10 gap	1.09	0.94
75-25 gap	0.47	0.39
Variance	0.208	0.155
Adjusted Variance	0.172	
Adjusted R <sup>2</sup>	0.68	
p-value for the F-test on:		
teacher fixed effects	0.000	
8th grade math score and year dummies	0.000	
Math scores units	Grade Equivalents	
Number of students	52,991	
Number of teachers	857	
# of students threshold	15	

Notes: All results are based on a regression of 9th grade math test score on 8th grade math test score, teacher semester counts, and year indicators.

**Table 6**  
**Quartile Rankings of Estimated Teacher Effects in Years t and t+1:**  
**Percent of Teachers by Row**

		Quartile in year t+1			
		1	2	3	4
Quartile in year t	1	40	28	24	8
	2	25	33	33	9
	3	24	31	26	19
	4	5	11	24	60

$\chi^2$  test of random quartile assignment:  $p < 0.001$

Notes: Quartile rankings are based on teacher effects estimated for each year based on the specification in column 1 of table 5.

**Table 7**  
**Distribution of the Estimated Teacher Effects**

	(1)	(2)	(3)	(4)	(5)
Variance	0.208	0.148	0.092	0.096	0.080
Adjusted Variance	0.172	0.115	0.059	0.054	0.040
Weighted Variance	0.155	0.110	0.061	0.061	0.047
p-value, F-test of teacher effects	0.000	0.000	0.000	0.000	0.000
p-value, F-test of lagged test score and year	0.000				
p,value F-test for basic student covariates		0.000			
p-value, F-test for school effects				0.000	0.000
p-value, F-test for additional student, peer, and neighborhood covariates			0.000		0.000
Included Covariates					
Year fixed effects	yes	yes	yes	yes	yes
Basic student covariates	no	yes	yes	yes	yes
Additional student covariates	no	no	yes	no	yes
Math peer covariates	no	no	yes	no	yes
Neighborhood covariates	no	no	yes	no	yes
School fixed effects	no	no	no	yes	yes
# of students threshold	15	15	15	15	15

Notes: All results are based on a regression of 9th grade math test score on 8th grade math test score, teacher semester counts, year indicators, and other covariates as listed in the table. All test scores are measured in grade equivalents. Student covariates include gender, race, age, guardianship, number of times in 9th grade, free or reduced-price lunch status, whether changed school during school year, and average math class size. Additional student covariates include level and subject of math classes, cumulative GPA, class rank, disability status, and whether school is outside of the student's residential neighborhood. Peer covariates include the 10th, 50th, and 90th percentile of math class absences and 8th grade math test scores in 9th grade math classes. Neighborhood covariates include median family income, median house value, and fraction of adult population that fall into five education categories. All neighborhood measures are based on 1990 census tract data. There are 52,991 students and 857 teachers in each specification.

**Table 8**  
**Quartile Rankings of Estimated Teacher Effects in Years t and t+1:**  
**Percent of Teachers by Row**

		Quartile in year t+1			
		1	2	3	4
Quartile in year t	1	34	34	13	19
	2	32	25	31	12
	3	14	25	38	24
	4	16	21	23	40

$\chi^2$  test of random quartile assignment:  $p < 0.000$

Notes: Quartile rankings are based on teacher effects estimated for each year based on the specification including lagged math test score, year indicators, and all student, peer, and neighborhood covariates.

**Table 9**  
**Distribution of the Estimated Teacher Effects**

	Student Threshold		Test Scores Measured in Percentiles	Trimming top and bottom 3 percent in changes	Ability Level		
	50	100			Low	Middle	High
Dependent variable mean	9.21	9.21	37.88	9.08	7.07	8.71	11.81
Mean test score gain	1.14	1.14	-2.08	1.06	0.54	0.67	2.22
Number of teachers	542	335	857	849	564	513	418
Number of students	52,991	52,991	52,991	50,426	16,892	18,625	17,474
Without school effects							
Variance of teacher effects	0.050	0.049	7.30	0.071	0.050	0.095	0.085
Adjusted variance	0.038	0.042	4.25	0.042	0.020	0.057	0.046
Weighted variance	0.047	0.045	4.90	0.045	0.037	0.069	0.059
p-value, F-test for teacher effects	0.000	0.000	0.000	0.000	0.000	0.000	0.000
With school effects							
Variance of teacher effects	0.034	0.030	7.07	0.064	0.056	0.106	0.078
Adjusted variance	0.017	0.021	3.24	0.027	0.016	0.050	0.018
Weighted variance	0.031	0.028	4.30	0.036	0.043	0.078	0.053
p-value, F-test for teacher effects	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Notes: See notes to table 7. Ability level is assigned in thirds based on the 8th grade test score distribution. High ability students have scores above 8.7, middle ability students score between 7.5 and 8.7, and low ability students have scores of less than 7.5. All regressions include the student, peer, and neighborhood covariates included in the Table 7, column 3 and 5 specifications.

**Table 10**  
**Distribution of the Estimated Teacher Effects**

	Teacher Quality Estimates		
	Math Only	Math and English	English Only
<b>Math Teachers</b>			
Variance	0.091	0.092	
Adjusted Variance	0.059	0.042	
Weighted Variance	0.061	0.053	
Number of math teachers	857	857	
<b>English Teachers</b>			
Variance		0.067	0.067
Adjusted Variance		0.012	0.027
Weighted Variance		0.042	0.051
Number of English teachers		1044	1044
F-statistic for math teacher effects	4.6	2.0	
F-statistic for English teacher effects		1.7	3.8

Notes: See notes to Table 7. There are 52,991 students in each specification. Column (1) is the same as column (3) of Table 7. Column (2) additionally includes controls for the English teachers, while column (3) only controls for English teachers

**Table 11**  
**Impact of Observable Characteristics on Teacher Fixed Effects**

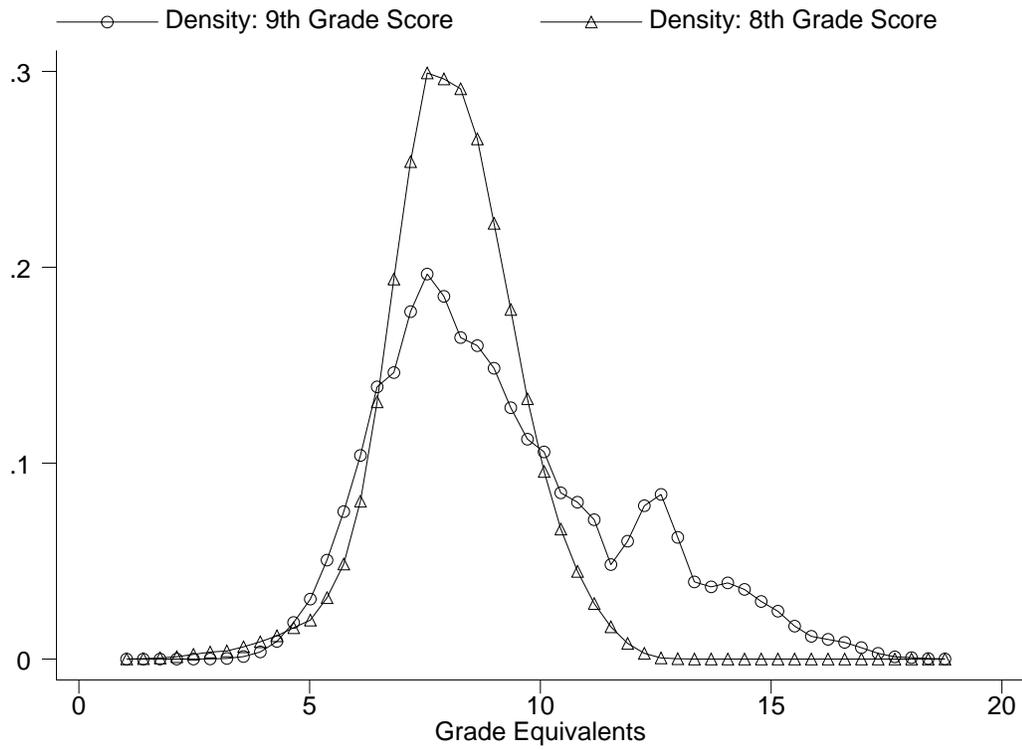
	(1)	(2)	(3)
Female		0.045 *	0.041 *
		(0.023)	(0.023)
Asian		0.048	0.043
		(0.052)	(0.052)
Black		0.024	0.019
		(0.026)	(0.026)
Hispanic		-0.069	-0.066
		(0.046)	(0.046)
Native American		0.024	0.022
		(0.115)	(0.114)
Potential experience		0.012	
		(0.010)	
squared		-0.001	
		(0.001)	
cubed (divided by 1000)		0.011	
		(0.009)	
Potential experience < 1			-0.031
			(0.052)
Potential experience = 1			0.278
			(0.186)
Masters	0.001	-0.015	-0.009
	(0.022)	(0.023)	(0.023)
PhD	-0.077	-0.064	-0.057
	(0.095)	(0.096)	(0.096)
BA major: education	0.096 *	0.076 *	0.079 *
	(0.033)	(0.039)	(0.039)
BA major: math	0.047 *	0.061 *	0.061 *
	(0.026)	(0.029)	(0.029)
BA major: science	0.064	0.076 *	0.083 *
	(0.043)	(0.045)	(0.045)
bilingual certificate		-0.071	-0.073
		(0.044)	(0.044)
child certificate		0.097	0.098
		(0.091)	(0.091)
elementary certificate		0.068	0.065
		(0.046)	(0.045)
high school certificate		0.010	0.008
		(0.038)	(0.038)
special ed certificate		0.023	0.024
		(0.043)	(0.043)
substitute certificate		0.028	0.035
		(0.029)	(0.029)
Tenure at CPS	-0.005	-0.009	-0.003

	(0.007)	(0.012)	(0.011)
squared	0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)
cubed (divided by 1000)	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
BA univ = level 1		0.018	0.017
		(0.043)	(0.043)
BA univ = level 2		0.049	0.041
		(0.044)	(0.044)
BA univ = level 3		0.014	0.014
		(0.033)	(0.033)
BA univ = level 4		0.025	0.023
		(0.045)	(0.045)
BA univ = level 5		0.098	0.108
		(0.091)	(0.090)
BA univ local		-0.002	-0.004
		(0.026)	(0.026)
adjusted R <sup>2</sup>	0.013	0.044	0.044
# of teachers with observables	645	645	645

---

Notes: \* = significant at 10 percent level. The dependent variable is teacher quality estimated using the table 7, column 3 specification. Each specification also includes a constant. Potential experience is calculated as age-education-6 and is the teacher's average over the 3 years.

Figure 1  
Kernel Density Estimates of 8<sup>th</sup> and 9<sup>th</sup> Grade Test Scores



Notes: Test scores are measured in grade equivalents. Estimates are calculated using the Epanechnikov kernel. For the 8<sup>th</sup> grade test score a bin width of approximately 0.14 is used, while for the 9<sup>th</sup> grade test a bin width of approximately 0.26 is used.

**Figure 2**  
Teacher Effect Estimates Versus Student Counts

