Teachers and Student Achievement in the Chicago Public High Schools

Daniel Aaronson, *Federal Reserve Bank of Chicago*

Lisa Barrow, *Federal Reserve Bank of Chicago*

William Sander, *DePaul University*

We estimate the importance of teachers in Chicago public high schools using matched student-teacher administrative data. A one standard deviation, one semester improvement in math teacher quality raises student math scores by 0.13 grade equivalents or, over 1 year, roughly one-fifth of average yearly gains. Estimates are relatively stable over time, reasonably impervious to a variety of conditioning variables, and do not appear to be driven by classroom sorting or selective score reporting. Also, teacher quality is particularly important for lower-ability students. Finally, traditional human capital measures—including those determining compensation—explain little of the variation in estimated 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, Kristin Butcher, Dave Card, Rajeev Dehejia, Tom DiCiccio, Eric French, Brian Jacob, Jeff Kling, Steve Rivkin, Doug Staiger, Dan Sullivan, Chris Taber, and seminar participants at many universities and conferences for helpful comments and discussions. The views expressed in this article are ours and are not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. Contact the corresponding author, Lisa Barrow, at lbarrow@frbchi.org.

[*Journal of Labor Economics*, 2007, vol. 25, no. 1]

0734-306X/2007/2501-0004$10.00

I. Introduction

The Coleman Report (Coleman et al. 1966) broke new ground in the estimation of education production functions, concluding that family background and peers were more important than schools and teachers in educational outcomes such as test scores and graduation rates. While research since Coleman 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.[1] However, how much they matter, the degree to which they vary across subpopulations, how robust quality rankings are to specification choices, 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 from the Chicago public high schools to estimate the importance of teachers on student mathematics test score gains and then relate our measures of individual teacher effectiveness to observable characteristics of the instructors. Our measure of teacher quality is the effect on ninth-grade math scores of a semester of instruction with a given teacher, controlling for eighth-grade math scores and student characteristics. Our data provide us with a key advantage in generating this estimate: the ability to link teachers and students in specific classrooms. In contrast, many other studies can only match students to the average teacher in a grade or school. In addition, because teachers are observed in multiple classroom settings, our teacher effect estimates are less likely to be driven by idiosyncratic class effects. Finally, 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 ninth-grade math achievement. Namely, after controlling for initial ability (as measured by test scores) and other student characteristics, teacher effects are statistically important in explaining ninth-grade math test score achievement, and the variation in teacher effect estimates is large

[1] Literature reviews include Greenwald, Hedges, and Laine (1996) and Hanushek (1996, 1997, 2002). A brief sampling of other work on teacher effects includes Murnane (1975), Goldhaber and Brewer (1997), Angrist and Lavy (2001), Jepsen and Rivkin (2002), Rivers and Sanders (2002), Jacob and Lefgren (2004), Rockoff (2004), Kane and Staiger (2005), Rivkin, Hanushek, and Kain (2005), and Kane, Rockoff, and Staiger (2006). The earliest studies on teacher quality were hampered by data availability and thus often relied on state- or school-level variation. Aggregation and measurement error compounded by proxies such as student-teacher ratios and average teacher experience can introduce significant bias. More recent studies, such as Rockoff (2004), Kane and Staiger (2005), Rivkin et al. (2005), and Kane et al. (2006), use administrative data like ours to minimize these concerns.

enough such that the expected difference in math achievement between having an average teacher and one that is one standard deviation above average is educationally important. However, a certain degree of caution must be exercised in estimating teacher quality using teacher fixed effects as biases related to measurement, particularly due to small populations of students used to identify certain teachers, can critically influence results. Sampling variation overstates our measures of teacher quality dispersion by amounts roughly similar to Kane and Staiger's (2002, 2005) evaluations of North Carolina schools and Los Angeles teachers. Correcting for sampling error, we find that the standard deviation in teacher quality in the Chicago public high schools is at least 0.13 grade equivalents per semester. Thus, over two semesters, a one standard deviation improvement in math teacher quality translates into an increase in math achievement equal to 22% of the average annual gain. This estimate is a bit higher than, but statistically indistinguishable from, those reported in Rockoff (2004) and Rivkin et al. (2005).[2]

Furthermore, we show that our results are unlikely to be driven by classroom sorting or selective use of test scores and, perhaps most importantly, the individual teacher ratings are relatively stable over time and reasonably impervious to a wide variety of conditioning variables. The latter result suggests that test score value-added measures for teacher productivity are not overly sensitive to reasonable statistical modeling decisions, and thus incentive schemes in teacher accountability systems that rely on similar estimates of productivity are not necessarily weakened by large measurement error in teacher productivity.

We also show how estimates vary by initial (eighth-grade) test scores, race, and sex and find that the biggest impact of a higher quality teacher, relative to the mean gain of that group, is among African American students and those with low or middle range eighth-grade test scores. We find no difference between boys and girls.

Finally, the vast majority of the variation in teacher effects is unexplained by easily observable teacher characteristics, including those used for compensation. While some teacher attributes are consistently related to our quality measure, together they explain at most 10% of the total variation in estimated teacher quality. Most troubling, the variables that determine compensation in Chicago—tenure, advanced degrees, and teaching certifications—explain roughly 1% of the total variation in es-

[2] Rivkin et al.'s (2005) lower bound estimates suggest that a one standard deviation increase in teacher quality increases student achievement by at least 0.11 standard deviations. Rockoff (2004) reports a 0.1 standard deviation gain from a one standard deviation increase in teacher quality from two New Jersey suburban school districts. In our results, a one standard deviation increase in teacher quality over a full year implies about a 0.15 standard deviation increase in math test score gains.

timated teacher quality. These results highlight the lack of a close relationship between teacher pay and productivity and the difficulty in developing compensation schedules that reward teachers for good work based solely on certifications, degrees, and other standard administrative data. That is not to say such schemes are not viable. Here, the economically and statistically important persistence of teacher quality over time should be underscored. By using past performance, administrators can predict teacher quality. Of course, such a history might not exist when recruiting, especially for rookie teachers, or may be overwhelmed by sampling variation for new hires, a key hurdle in prescribing recruitment, retention, and compensation strategies at the beginning of the work cycle. Nevertheless, there is clearly scope for using test score data among other evaluation tools for tenure, compensation, and classroom organization decisions.

While our study focuses on only one school district over a 3-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% are Hispanic, and roughly 80% are eligible for 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%–90% nonwhite and roughly 70%–80% eligible for free or reduced-price school lunch (U.S. Department of Education 2003). Therefore, on these dimensions Chicago is quite representative of the school systems that generate the most concern in education policy discussions.

II. Background and Data

The unique detail and scope of our data are major strengths 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.[3] We concentrate on the performance of ninth graders in this article.

The key advantage to using administrative records is being able to work with the population of students, a trait of several other recent studies, including Rockoff (2004), Kane and Staiger (2005), Rivkin et al. (2005), and Kane et al. (2006). Apart from offering a large sample of urban schoolchildren, the CPS administrative records provide several other useful fea-

[3] Of the 88 schools, six 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.

tures that rarely appear together in other studies. First, this is the first study that we are aware of that examines high school teachers. Clearly, it is important to understand teacher effects at all points in the education process. Studying high schools has the additional advantage that classrooms are subject specific, and our data provide enough school scheduling detail to construct actual classrooms. Thus, we can examine student-teacher matches at a level that plausibly corresponds with what we think of as a teacher effect. This allows us to isolate the impact of math teachers on math achievement gains. However, we can go even further by, say, looking at the impact of English teachers on math gains. In this study, we report such exercises as robustness checks, but data like these offer some potential for exploring externalities or complementarities between teachers.

The teacher records also include specifics about human capital and demographics. These data allow us to decompose the teacher effect variation into shares driven by unobservable and observable factors, including those on which compensation is based. Finally, the student and teacher records are longitudinal. This has several advantages. Although our data are limited to high school students, they include a history of pre–high school test scores that can be used as controls for past (latent) inputs. Furthermore, each teacher is evaluated based on multiple classrooms over (potentially) multiple years, thus mitigating the influence of unobserved idiosyncratic class effects.

A. Student Records

There are three general components of the student data: test scores, school and scheduling variables, and family and student background measures. Like most administrative data sets, the latter is somewhat limited. Table 1 includes descriptive statistics for some of the variables available, including sex, race, age, eligibility for the free or reduced school lunch program, and guardian (mom, dad, grandparent, etc.). Residential location is also provided, allowing us to incorporate census tract information on education, income, and house values. We concentrate our discussion below on the test score and scheduling measures that are less standard.

1. *Test Scores*

In order to measure student achievement, we rely on student test scores from two standardized tests administered by the Chicago Public Schools—the Iowa Test of Basic Skills (ITBS) administered in the spring of grades 3–8 and the Test of Achievement and Proficiency (TAP) administered during the spring for grades 9 and 11.[4] We limit the study to

[4] TAP testing was mandatory for grades 9 and 11 through 1998. The year 1999 was a transition year in which ninth, tenth, and eleventh graders were tested. Starting in 2000, TAP testing is mandatory for grades 9 and 10.

Table 1
Descriptive Statistics for the Student Data

	All Students (1)		Students with Eighth- and Ninth-Grade Math Test Scores (2)		Students with Eighth- and Ninth-Grade Math Test Scores 1 Year Apart (3)	
Sample size:						
Total	84,154		64,423		52,957	
1997	29,301		21,992		17,941	
1998	27,340		20,905		16,936	
1999	27,513		21,526		18,080	
	Mean	SD	Mean	SD	Mean	SD
Test scores (grade equivalents):						
Math, ninth grade	9.07	2.74	9.05	2.71	9.21	2.64
Math, eighth grade	7.75	1.55	7.90	1.50	8.07	1.41
Math change, eighth to ninth grade	1.15	1.89	1.15	1.89	1.14	1.75
Reading comprehension, ninth grade	8.50	2.94	8.50	2.89	8.63	2.88
Reading comprehension, eighth grade	7.64	1.94	7.82	1.88	8.01	1.80
Reading change, eighth to ninth grade	.66	2.02	.67	2.02	.62	1.95
Demographics:						
Age	14.8	.8	14.7	.8	14.6	.7
Female	.497	.500	.511	.500	.522	.500
Asian	.035	.184	.033	.179	.036	.185
African American	.549	.498	.570	.495	.562	.496
Hispanic	.311	.463	.304	.460	.307	.461
Native American	.002	.047	.002	.046	.002	.046
Eligible for free school lunch	.703	.457	.721	.448	.728	.445
Eligible for reduced-price school lunch	.091	.288	.097	.295	.103	.304
Legal guardian:						
Dad	.241	.428	.244	.429	.253	.435
Mom	.620	.485	.626	.484	.619	.486
Nonrelative	.041	.197	.039	.195	.037	.189
Other relative	.038	.191	.034	.182	.032	.177
Stepparent	.002	.050	.002	.047	.002	.046
Schooling:						
Take algebra	.825	.380	.865	.342	.950	.217
Take geometry	.101	.302	.092	.290	.022	.145
Take computer science	.003	.054	.003	.057	.003	.057
Take calculus	.0001	.011	.0001	.010	.0001	.008
Fraction honors math classes	.081	.269	.093	.286	.101	.297
Fraction regular math classes	.824	.360	.827	.356	.820	.361
Fraction essential math classes	.032	.172	.029	.163	.032	.172
Fraction basic math classes	.001	.036	.001	.031	.001	.034
Fraction special education math classes	.014	.114	.009	.093	.009	.093
Fraction nonlevel math classes	.006	.057	.005	.054	.006	.057
Fraction level missing math classes	.042	.166	.036	.146	.030	.125
Fraction of math grades that are A	.083	.256	.085	.257	.093	.267
Fraction of math grades that are B	.130	.297	.138	.304	.151	.313
Fraction of math grades that are C	.201	.351	.218	.359	.232	.364
Fraction of math grades that are D	.233	.371	.250	.378	.252	.374
Fraction of math grades that are F	.311	.430	.272	.410	.241	.389
Fraction of math grades missing	.042	.166	.036	.146	.030	.125

Table 1 (*Continued*)

	All Students (1)		Students with Eighth- and Ninth-Grade Math Test Scores (2)		Students with Eighth- and Ninth-Grade Math Test Scores 1 Year Apart (3)	
Number of math/computer science classes taken in ninth grade	2.1	.4	2.1	.4	2.1	.4
Number of times in ninth grade	1.10	.31	1.08	.28	1.00	.00
Changed school within the year	.034	.180	.030	.170	.027	.163
Average class size among ninth-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 ninth-grade math	13.9	16.7	11.6	13.7	9.9	11.7
Identified as disabled	.021	.143	.024	.154	.022	.147

NOTE.—The share of students disabled does not include students identified as learning disabled. Roughly 9% of CPS students in our estimation sample are identified as learning disabled.

ninth-grade students and primarily limit our analysis to math test scores. By limiting the study to ninth-grade students, we can also limit the sample to students with test scores from consecutive years in order to ensure that we associate math achievement with the student's teacher exposure in that same year. Although we also have information on reading test scores, we choose to focus on math achievement because the link between math teachers and math test scores is cleaner than for any single subject and reading scores. In addition, math test scores seem to have more, or are often assumed to have more, predictive power than reading scores for future productivity (see, e.g., Murnane et al. 1991; Grogger and Eide 1995; and Hanushek and Kimko 2000).

Multiple test scores are 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, we rely on a general form of the value-added model of education production in which we regress the ninth-grade test score on the variables of interest while controlling for initial achievement as measured by the previous year's eighth-grade test score.

We observe both eighth- and ninth-grade test scores for the majority of ninth-grade students, as shown in table 1. Scores are reported as grade equivalents, a national normalization that assigns grade levels to test score results in order to evaluate whether students have achieved the skills that are appropriate for their grade. For instance, a 9.7 implies that the student is performing at the level of a typical student in the seventh month of ninth grade. Unique student identifiers allow us to match the ninth-grade students to both their ninth-grade TAP score and their eighth-grade ITBS score.

Eighth- and ninth-grade test score data are reported for between 75% and 78% of the ninth-grade students in the CPS, yielding a potential sample of around 64,000 unique students over the 3-year period. Our sample drops to 53,000 when we exclude students without eighth- and ninth-grade test scores in consecutive school years and those with test score gains in the 1st and 99th percentiles.

Since the ninth-grade test is not a high stakes test for either students or teachers, it is less likely to elicit "cheating" in any form compared to the explicit teacher cheating uncovered in Jacob and Levitt (2003). In addition, by eliminating the outlier observations in terms of test score gains, we may drop some students for whom either the eighth- or ninth-grade test score is "too high" due to cheating. That said, there may be reasonable concern that missing test scores reflect some selection about which students take the tests or which test scores are reported.

Approximately 11% of ninth graders do not have an eighth-grade math test score, and 17% do not have a ninth-grade score.[5] 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. 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 potentially exempt from the test). Convincing exclusion restrictions are not available to adequately assess the importance of selection of this type.[6] However, later in the article we show that our quality measure is not correlated with missing test scores, suggesting that this type of selection or gaming of the system does not unduly influence our measure of teacher quality.

Finally, the raw data suggest that racial and income test score gaps rise dramatically between the eighth and ninth grade. While we expect that higher-ability students may gain more in 1 year of education than lower-ability students, we also suspect that the rising gap may be a function of the different exams. In figure 1, kernel density estimates of the eighth- and ninth-grade math test scores are plotted. The ninth-grade scores are skewed right while the eighth-grade test score distribution is more sym-

[5] Eighty-six percent of the students took the TAP (ninth-grade test), and, of this group, we observe scores for 98%.

[6] If selection is based on potential test score improvements because schools and teachers are gaming test score outcomes by reporting scores only for students with the largest gains, we could overstate the impact of teacher quality. 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. While there is no obvious candidate, we tried several, including absences, distance to school, and distance to school interacted with whether the student is in their neighborhood school. With the caveat that none of these instruments are ideal, our primary conclusions are unaffected by a selection correction that uses them.

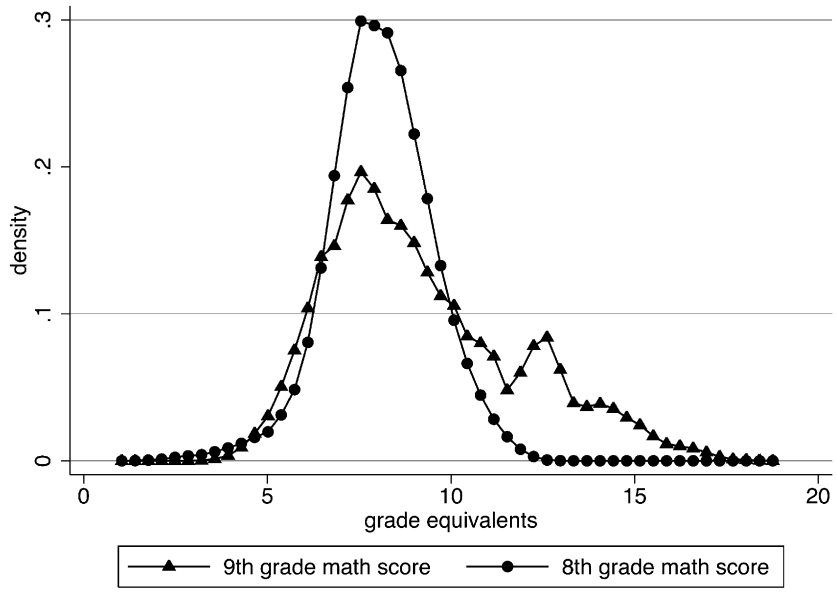

FIG. 1.—Kernel density estimates of eighth- and ninth-grade math test scores. Test scores are measured in grade equivalents. Estimates are calculated using the Epanechnikov kernel. For the eighth-grade test score a bin width of approximately 0.14 is used, while for the ninth-grade test a bin width of approximately 0.26 is used.

metric. As a consequence, controlling for eighth-grade test scores in the regression of ninth-grade test scores on teacher indicators and other student characteristics may not adequately control for the initial quality of a particular teacher's students and may thus lead us to conclude that teachers with better than average students are superior instructors. We drop the top and bottom 1% of the students by change in test scores to partly account for this problem. We also discuss additional strategies, including using alternative test score measures that are immune to differences in scaling of the test, accounting for student attributes, and analyzing groups of students by initial ability.

2. *Classroom Scheduling*

A second important feature of the student data is the detailed scheduling information 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) the class met, the teacher assigned, the title of the class, and the course level (i.e., advanced placement, 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 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 classroom peers.

One natural concern in estimating teacher quality is whether there are lingering influences from the classroom sorting process. That is, students may be purposely placed with certain instructors based on their learning potential. The most likely scenario 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.[7]

To assess the extent to which students may be sorted based on expected test score gains, we calculate test score dispersion for the observed teacher assignments and for several counterfactual teacher assignments. In table 2, we report the degree to which the observed within-teacher standard deviation in students' pre-ninth-grade performance differs from simulated classrooms that are either assigned randomly or sorted based on test score rank. We use three lagged test score measures for assignment: eighth-grade test scores, sixth- to seventh-grade test score gains, and seventh- to eighth-grade test score gains. Each panel reports results for the three fall semesters in our data.[8] The top row of each panel, labeled "Observed," displays the observed average within-teacher standard deviation of these measures. This is the baseline to which we compare the simulations. Each of the four subsequent rows assigns students to teachers either randomly or based on pre-ninth-grade performance.

Row 2 displays the average within-teacher standard deviation when students are perfectly sorted across teachers within their home school.[9] Such a within-school sorting mechanism reduces the within-teacher standard deviation to roughly 20% 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 standard deviation is very close to the within-teacher standard deviation that is observed in the data. Strikingly, there is no evidence that sorting occurs on past gains; the

[7] Informal discussions with a representative of the Chicago public school system suggest that parents have little influence on teacher selection and, conditional on course level, the process is not based on student characteristics. Moreover, our use of first-year high school students may alleviate concern since it is likely more difficult for schools to evaluate new students, particularly on unobservable characteristics.

[8] The estimates for the spring semester are very similar and available from the authors on request.

[9] For example, within an individual school, say there are three classrooms with 15, 20, and 25 students. In the simulation, the top 25 students, based on our pre-ninth-grade measures, would be placed together, the next 20 in the second classroom, and the remainder in the last. The number of schools, teachers, and class sizes are set equal to that observed in the data.

Table 2
Mean Standard Deviation by Teacher of Lagged Student Test Score Measures

	Eighth-Grade Scores (1)	Sixth to Seventh Change (2)	Seventh to Eighth Change (3)
Fall 1997:			
Observed	1.042	.659	.690
Perfect sorting across teachers within school	.214	.132	.136
Randomly assigned teachers within school	1.211	.635	.665
Perfect sorting across teachers	.006	.004	.004
Randomly assigned teachers	1.445	.636	.662
Fall 1998:			
Observed	1.095	.653	.731
Perfect sorting across teachers within school	.252	.151	.175
Randomly assigned teachers within school	1.279	.635	.721
Perfect sorting across teachers	.007	.005	.008
Randomly assigned teachers	1.500	.633	.720
Fall 1999:			
Observed	1.142	.662	.792
Perfect sorting across teachers within school	.274	.168	.217
Randomly assigned teachers within school	1.320	.647	.766
Perfect sorting across teachers	.007	.005	.009
Randomly assigned teachers	1.551	.652	.780

NOTE.—In each cell, we report the average standard deviation 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 standard deviation 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" sort 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.

observed standard deviations are even slightly larger than the simulations. Using eighth-grade test scores, the randomly assigned matches tend to have within-teacher standard deviations that are roughly 15% higher than the observed assignments. But clearly, the observed teacher dispersion 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 performance.[10]

Finally, rows 4 and 5 show simulations of perfectly sorted and randomly assigned classrooms across the entire school district. Here, the exercise disregards which school the student actually attends. Consequently, this

[10] 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.

example highlights the extent to which classroom composition varies across versus within school. We find that the randomly assigned simulation (row 5) is about 18% above the equivalent simulation based solely on within-school assignment and roughly 37% above the observed baseline. Furthermore, there is virtually no variation within randomly assigned classrooms across the whole district. Thus, observed teacher assignment is clearly closer to random than sorted, especially with regard to previous achievement gains, but some residual sorting in levels remains. About half of that is due to within-school classroom assignment and half to across-school variation. School fixed effects provide a simple way to eliminate the latter (Clotfelter, Ladd, and Vigdor 2004).

B. Teacher Records

Finally, we match student administrative records to teacher administrative records using school identifiers and eight-character teacher codes from the student data.[11] The teacher file contains 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 identify 1,132 possible teachers of ninth-grade mathematics classes (these are classes with a "math" course number, although some have course titles suggesting they are computer science). This list is further pared by grouping all teachers who do not have at least 15 student-semesters during our period into a single "other" teacher code for estimation purposes.[12] Ultimately, we identify teacher effects for 783 math 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 administrative files, they allow for reasonably precise estimation.

Matching student and teacher records allows us to take advantage of a

[11] Additional details about the matching are available in the appendix.

[12] The larger list of teachers incorporates anyone instructing a math class with at least one ninth-grade student over our sample period, including instructors who normally teach another subject or grade. The number of student-semesters for each teacher over 3 years may be smaller than expected for several reasons (this is particularly evident in fig. 2 below). Most obviously, some teacher codes may represent errors in the administrative data. Also, some codes may represent temporary vacancies. More importantly, Chicago Public Schools high school teachers teach students in multiple grades as well as in subjects other than math. In fact, most teachers of math classes in our analysis sample (89%) teach students of multiple grade levels. For the average teacher, 58% of her students are in the ninth grade. In addition, roughly 40% of the teachers in the analysis sample also teach classes that are not math classes. Without excluding students for any reason, the teachers in our sample have an average of 189 unique students in all grades and all subjects. Limiting the classes to math courses drops the average number of students to 169. When we further limit the students to ninth graders, the average number of students is 80.

third feature of the data: the detailed demographic and human capital information supplied in the teacher administrative files. In particular, we can use a teacher's sex, race/ethnicity, experience, tenure, university attended, college major, advanced degree achievement, and teaching certification to decompose total teacher effects into those related to common observable traits of teachers and those that are unobserved, such as drive, passion, and connection with students.

In order to match the teacher data to the student data we have to construct an alphanumeric code in the teacher data similar to the one provided in the student data. The teacher identifier in the student data is a combination of the teacher's position number and letters from the teacher's name, most often the first three letters of his or her last name. We make adjustments to the identifiers in cases for which the teacher codes in the student files do not match our constructed codes in the teacher data due to discrepancies that arise for obvious reasons such as hyphenated last names, use of the first initial plus the first two letters of the last name, or transposed digits in the position number. Ultimately we are unable to resolve all of the mismatches between the student and teacher data but are able to match teacher characteristics to 75% of the teacher codes for which we estimate teacher quality (589 teachers). Table 3 provides descriptive statistics for the teachers we can match to the student administrative records. The average teacher is 45 years old and has been in the CPS for 13 years. Minority math and computer science teachers are underrepresented relative to the student population, as 36% are African American and 10% Hispanic, but they compare more favorably to the overall population of Chicago, which is 37% black or African American and 26% Hispanic or Latino (Census 2000 Fact Sheet for Chicago, U.S. Census Bureau). Eighty-two percent are certified to teach high school, 37% are certified to be a substitute, and 10%–12% 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 (48%) or education (18%).[13]

III. 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

[13] Nationally, 55% of high school teachers have a master's degree, 66% have an academic degree (e.g., mathematics major), and 29% have a subject area education degree (U.S. Department of Education 2000).

Table 3
Descriptive Statistics for the Teachers Matched to Math Teachers in the Student Data

	Mean	Standard Deviation
Demographics:		
Age	45.15	10.54
Female	.518	.500
African American	.360	.480
White	.469	.499
Hispanic	.100	.300
Asian	.063	.243
Native American	.007	.082
Human capital:		
BA major: education	.182	.386
BA major: all else	.261	.440
BA major: math	.484	.500
BA major: science	.073	.260
BA university, *US News* 1	.092	.289
BA university, *US News* 2	.081	.274
BA university, *US News* 3	.151	.358
BA university, *US News* 4	.076	.266
BA university, *US News* 5	.019	.135
BA university, *US News* else	.560	.497
BA university missing	.020	.141
BA university local	.587	.493
Master's degree	.521	.500
PhD	.015	.123
Certificate, bilingual education	.119	.324
Certificate, child	.015	.123
Certificate, elementary	.100	.300
Certificate, high school	.823	.382
Certificate, special education	.107	.309
Certificate, substitute	.365	.482
Potential experience	19.12	11.30
Tenure at CPS	13.31	10.00
Tenure in position	5.96	6.11
Number of observations	589	

NOTE.—There are 783 teachers identified from the student estimation sample that have at least 15 student-semesters for math classes over the 1997–99 sample period. The descriptive statistics above apply to the subset of these teachers that can be matched to the teacher administrative records from the Chicago Public Schools. *US News* rankings are from U.S. News & World Report (1995): level 1 = top tier universities (top 25 national universities + tier 1 national universities) + (top 25 national liberal arts colleges + tier 1 national liberal arts colleges); level 2 = second tier national universities + second tier national liberal arts colleges; level 3 = third tier national universities + third tier national liberal arts colleges; level 4 = fourth tier national universities + fourth tier national liberal arts colleges; and level 5 = top regional colleges and universities.

age, as well as cumulative teacher and school inputs, S, from grades kindergarten through the current grade:

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

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 as-

sumptions about the covariance structure of observables and unobservables are maintained. Thus, alternative identification strategies are typically applied.

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 eighth-grade test scores as a covariate in explaining ninth-grade test scores. 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 these data, as information on the family and pre-ninth-grade schooling is sparse.

We estimate an education production model of the general form

$$Y^9_{ikt} = \alpha Y^8_{it-1} + \beta X_i + \tau T_{it} + \theta_i + \rho_k + \varepsilon_{ijkt}, \quad (2)$$

where Y^9_{ikt} refers to the ninth-grade test score of student i, who is enrolled in ninth grade at school k in year t; Y^8_{it-1} is the eighth-grade test score for student i, who is enrolled in ninth grade in year t; and θ_i, ρ_k, and ε_{ijk} measure the unobserved impact of individuals, schools, and white noise, respectively.[14] Each element of matrix T_{it} records the number of semesters spent in a math course with teacher j. To be clear, this is a cross-sectional regression estimated using ordinary least squares with a slight deviation from the standard teacher fixed effect specification.[15] Therefore, τ_j is the jth element of the vector τ, representing the effect of one semester spent with math teacher j. Relative to equation (1), the impact of lagged schooling and other characteristics is now captured by the lagged test score measure.

While the value-added specification helps control for the fact that teachers may be assigned students with different initial ability on average, this strategy may still mismeasure teacher quality. 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, N_j. In this case, estimates of τ_j may be biased by $\rho_k + \frac{1}{N_j}\sum_{i=1}^{N_j}\theta_i + \frac{1}{N_j}\sum_{i=1}^{N_j}\varepsilon_{ijkt}$.

The school term ρ_k is typically removed by including measures of school quality, a general form of which is school fixed effects. School fixed effects 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, 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 3-year period. For short time periods, such as

[14] All regressions include year indicators to control for any secular changes in test performance or reporting.

[15] For repeaters, we use their first ninth-grade year so as to allow only a 1-year gap between eighth- and ninth-grade test results.

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 potential identifying variation. In addition, to the extent that a principal is good because she is able to identify and hire high quality teachers, some of the teacher quality may be attributed to the school. For these reasons, we show many results both with and without allowing for school fixed effects.

Factors influencing test scores are often attributed to a student's family background. In the context of gains, many researchers argue that time-invariant qualities are differenced out, leaving only time-varying influences, such as parental divorce or a student's introduction to drugs, in $\frac{1}{N_j}\sum_{i=1}^{N_j}\theta_i$. While undoubtedly working in gains lessens the omitted variables problem, we want to be careful not to claim that value-added frameworks completely eliminate it. In fact, it is quite reasonable to conjecture that student gains vary with time-varying unobservables. But given our statistical model, bias is only introduced to the teacher quality rankings if students are assigned to teachers based on these unobservable changes.[16] Furthermore, we include a substantial list 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 trajectories.

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 problematic when teacher 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 standard deviation of the distribution of estimated τ_j is most likely inflated.

This problem is illustrated by figure 2, in which we plot our estimates $\hat{\tau}_j$ (conditional on eighth-grade math score, year indicators, and student, family, and peer attributes, as described below) against the number of student-semesters on which the estimate is based. What is notable is that the lowest and highest performing teachers are those with the fewest student-semesters. The expression $\sum_i T_{ij}$ represents the number of student-semesters taught by teacher j over the 3-year period examined (see n. 12 for a discussion of the distribution of $\sum_i T_{ij}$). As more student-semesters are used to estimate the fixed effect, the importance of sampling variation declines and reliability improves. Regressing $|\hat{\tau}_j|$ on $\sum_i T_{ij}$ summarizes this association. Such an exercise has a coefficient estimate of -0.00045 with a standard error of 0.000076, suggesting that number of student-semesters is a critical correlate of the magnitude of estimated teacher quality. The

[16] 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.

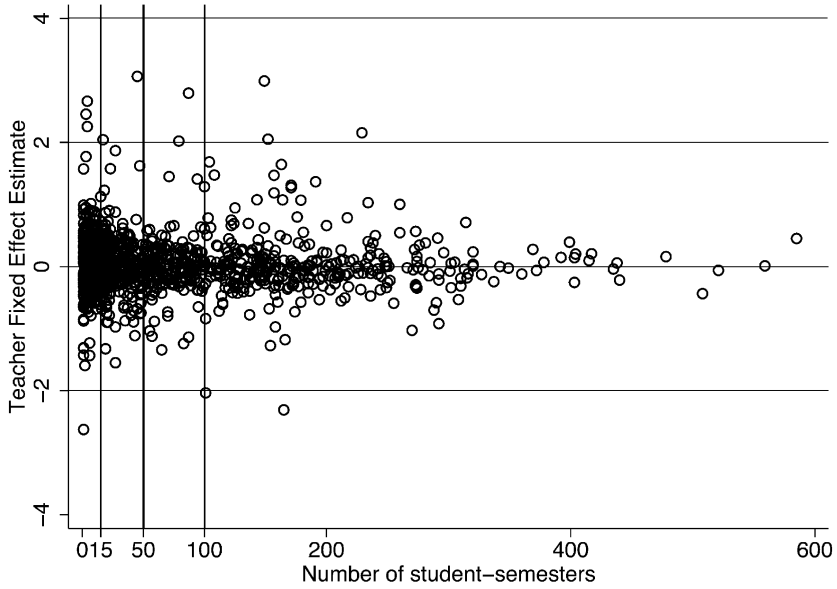

FIG. 2.—Teacher effect estimates versus student counts

association declines as we raise the minimum threshold on $\sum_i T_{ij}$ and completely disappears when $\sum_i T_{ij} \geq 250$.[17]

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 true teacher effect, τ_j, plus some error, ε_j, where ε_j is uncorrelated with τ_j. While we would like to estimate σ_τ^2, the variance of the estimated teacher effects is $\sigma_{\hat{\tau}}^2 = \sigma_\tau^2 + N^{-1}\varepsilon\varepsilon$. That is, the variance of the estimated teacher effects has two components—the true variance of the teacher effects and average sampling variance. We use the mean of the square of the standard error estimates of $\hat{\tau}_j$ as an estimate of the sampling error variance and subtract this from the observed variance of $\hat{\tau}_j$ to get an adjusted variance, σ_τ^2. We report the associated standard deviations, σ_τ and $\sigma_{\hat{\tau}}$, in subsequent tables. We also show how these values vary as we increase the minimum evaluation threshold, $\sum_i T_{ij}$. For statistical models that include school fixed effects, we estimate that roughly 30% of the standard deviation in estimated teacher quality is due to sampling error. If we raise the minimum number

[17] When $\sum_i T_{ij} \geq 250$, the point estimate and standard error are −0.0000367 (−0.0001597). While the standard error doubles due to the smaller sample of teachers as we move from the student threshold from 15 to 250, the point estimate declines substantially as well.

of student-semesters to identify an individual teacher to 200, only 14% of the standard deviation in teacher quality is due to sampling error.[18]

In the section to follow, we present our baseline estimates that ignore the existence of most of these potential biases. 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 ordinary least squares (OLS) estimates in the presence of school-specific fixed effects and because students are assigned multiple teachers per year, we do not include the teacher characteristics directly in equation (2). Rather, we employ a generalized least squares (GLS) estimator outlined in Borjas (1987) and Borjas and Sueyoshi (1994). This estimator regresses $\hat{\tau}_j$ on teacher characteristics Z:

$$\hat{\tau}_j = \phi Z_j + u_j. \tag{3}$$

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.

$$\Omega = \sigma_u^2 I_J + V. \tag{4}$$

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

IV. Results

A. The Distribution of Teacher Quality

Our naive baseline estimates of teacher quality are presented in table 4. In column 1 we present details on the distribution of $\hat{\tau}_j$, specifically the standard deviation and the 10th, 25th, 50th, 75th, and 90th percentiles. We also list the p-value for an F-test of the joint significance of the teacher effects (i.e., $\hat{\tau}_j = 0$ for all j) and the p-value for an F-test of the other regressors. In this parsimonious specification, the list of regressors is limited to eighth-grade math scores, year dummies, and indicators of the test

[18] Note, however, that excluding teachers with small numbers of students is limiting because new teachers, particularly those for whom tenure decisions are being considered, may not be examined. This would be particularly troubling for elementary school teachers with fewer students per year.

Table 4
Distribution of the Estimated Teacher Effects

	Distribution of Teacher Fixed Effects	
	Unweighted (1)	Weighted (2)
10th percentile	−.38	−.33
25th percentile	−.24	−.19
50th percentile	−.08	−.05
75th percentile	.17	.17
90th percentile	.53	.53
90–10 gap	.91	.86
75–25 gap	.41	.36
Standard deviation	.398	.354
Adjusted standard deviation	.355	
Adjusted R^2	.69	
p-value for the *F*-test on:		
Teacher fixed effects	.000	
Eighth-grade math score and year dummies	.000	
Math scores units	Grade equivalents	
Number of students	52,957	
Number of teachers	783	
Number of students threshold	15	

NOTE.—All results are based on a regression of ninth-grade math test score on eighth-grade math test score, ninth-grade test score level, eighth-grade test score level, an indicator equal to one if the information on eighth-grade test score level is missing, teacher semester counts, and year indicators.

level and format.[19] 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,

[19] Naturally, the key covariate in our production functions, regardless of specification, is the eighth-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 eighth-grade test score is 1.30 (0.01). This suggests that 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 tenth or eleventh grade. Nevertheless, a simple way to deal with nonstationarity is to estimate eq. (2) in differenced form. Such a specification will lead to inconsistent estimates because of the correlation between the error term and the lagged differenced dependent variable, but a common strategy to avoid this problem is to use the twice lagged differenced dependent variable, in our case the difference between seventh- and sixth-grade scores, as an instrument. This instrumental variables estimator reduces our estimates of the dispersion in teacher effects slightly (by less than 0.02 in our preferred specifications) but broadly supports the results presented below. It also suggests that controlling for student fixed effects is not likely to change our results significantly. However, we do not want to stress this result too much since it is based on a potentially nonrepresentative sample, those with test scores in every year between sixth and ninth grade.

Table 5
Quartile Rankings of Estimated Teacher Effects in Years t and $t + 1$: Percent of Teachers by Row

Quartile in year t:	Quartile in Year $t + 1$			
	1	2	3	4
1	36	29	26	10
2	24	31	32	12
3	20	32	23	24
4	8	12	23	57

NOTE.—χ^2 test of random quartile assignment: $p < .000$. Quartile rankings are based on teacher effects estimated for each year based on the specification in col. 1 of table 6.

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 standard deviation of $\hat{\tau}$ is 0.40 with gaps between the 90th percentile and 10th percentile teacher of 0.9 grade equivalents. Furthermore, approximately 0.4 grade equivalents separate average gains between the 75th and 25th percentile teacher. An F-test of the joint significance of $\hat{\tau}$ easily rejects no teacher effects at the highest significance level.

Because we have multiple classrooms per teacher and can follow teachers across years, the robustness of these results can be explored by tracking the stability of individual teacher quality over time. To do so, we simply estimate equation (2) separately by school year and then compare estimates for the same teacher in different school years. The extent to which $\hat{\tau}_{jt}$ is autocorrelated gives a measure of the extent to which signal dominates noise in our quality rankings.

Table 5 displays one such analysis. Here we report a transition matrix linking quartile rankings of $\hat{\tau}_{jt}$ with quartile rankings of $\hat{\tau}_{jt+1}$. Quartile 1 represents the lowest 25% of teachers as ranked by the teacher quality estimate, and quartile 4 represents the highest 25%. The table reports each cell's share of the 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. 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 the 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, 36%

remain in year $t + 1$, 29% move into quartile 2, 26% into quartile 3, and 10% into the highest quartile. Of those in the highest quartile in year t (row 4), 57% remain the following year, 23% move one category down, and only 20% fall into the lowest half of the quality distribution. A chi-square test easily rejects random assignment.[20]

Moreover, we have 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, 56% rank there the following year. This is highly significant relative to the random draw scenario whereby 10% would again appear in the top decile in consecutive years. However, of those teachers in the bottom decile, only 6% 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. To appear in our transition matrix, a teacher must be in the administrative records for two consecutive years. Therefore, if poor performing teachers are more likely to leave the school system, our test is biased; the random draw baseline would no longer be 10%. 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.[21] We find that a teacher ranked at the bottom is 13% less likely (standard error of 6%) than a teacher ranked in the 10th to 90th 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 10th to 90th percentile group. While accounting for the higher turnover rate of bottom decile ranked teachers does not lead us to conclude that there is significant persistence at the very bottom of the quality distribution in this particular specification, it does once we begin to refine the production function specification below.

Regardless, all of these results emphasize that teacher quality evaluated using parsimonious specifications with little attention to measurement issues still has an important persistent component. 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.

[20] Similarly, regressing contemporaneous teacher quality on lagged teacher quality results in a point estimate of 0.47 (0.04) for 1998 and 0.62 (0.07) for 1999. Limiting it to teachers in all 3 years, the coefficients (and standard errors) on lagged and twice lagged teacher quality are 0.49 (0.10) and 0.25 (0.09).

[21] Unfortunately, we cannot distinguish quits from layoffs or exits out of teaching from exits into other school systems.

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 test score changes in the extreme tails and by restricting identified teachers to those with more than 15 student-semesters. However, Kane and Staiger (2002) show that more than one-half of the variance in score gains from small North Carolina schools (typically smaller than our counts of student-semesters, $\sum_i T_{ij}$) and one-third of the variance in test score gains from larger North Carolina schools are due to sampling variation. Figure 2 emphasizes the susceptibility of our results to these concerns as well.

The row labeled "Adjusted Standard Deviation" in table 4 presents an estimate of σ_τ, the true standard deviation of the teacher effects after adjusting for sampling variation as described earlier. This modification reduces the standard deviation from 0.40 to 0.36. 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 standard deviation of the teacher effects drops to 0.35, virtually identical to the adjusted standard deviation reported in column 1. In either case, we conclude that dispersion in teacher quality is wide and educationally significant.

C. Family, Student, and Peer Characteristics

The teacher quality results reported thus far are based on parsimonious specifications. They do not fully capture heterogeneity in student, family, and peer background that could be correlated with particular teachers. In table 6 we report results in which available student, family, and peer group characteristics are included. For comparison purposes, column 1 repeats the findings from table 4. In each column we report unadjusted, adjusted, and weighted standard deviation 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.

In column 2 we incorporate student characteristics including sex, 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 ninth-grade math class size, as is standard in educational production analysis, and controls for whether the student changed high school or repeated ninth grade.[22]

[22] Jointly these background measures are quite significant; individually, the sex and race measures are the primary reason. The ninth-grade scores for female students are 0.16 (0.01) less than males, and African American and Hispanic students score 0.50 (0.03) and 0.31 (0.03) less than non–African American, non-Hispanic students. Accounting for additional student characteristics such as dis-

Table 6
Distribution of the Estimated Teacher Effects

	(1)	(2)	(3)	(4)	(5)
Standard deviation	.398	.384	.298	.303	.273
Adjusted standard deviation	.355	.341	.242	.230	.193
Weighted standard deviation	.354	.335	.246	.248	.213
p-value, *F*-test of teacher effects	.000	.000	.000	.000	.000
p-value, *F*-test of lagged test score and year	.000				
p-value, *F*-test for basic student covariates		.000			
p-value, *F*-test for school effects				.000	.000
p-value, *F*-test for additional student, peer, and neighborhood covariates			.000		.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
Number of students threshold	15	15	15	15	15

NOTE.—All results are based on a regression of ninth-grade math test score on eighth-grade math test score, teacher student-semester counts, year indicators, ninth-grade test level, eighth-grade test level, an indicator equal to one if the information on eighth-grade test score level is missing, and other covariates as listed in the table. All test scores are measured in grade equivalents. Basic student covariates include gender, race, age, guardianship, number of times in ninth 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 eighth-grade math test scores in ninth-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 52,957 students and 783 teachers in each specification.

These controls reduce the size of the adjusted standard deviation by a small amount, but the estimates remain large and highly statistically significant.

In column 3 we introduce additional student controls, primarily related to performance, school choice, peers, 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, and disability status, and whether the school is outside of her residential neigh-

ability status and average grades, neighborhood characteristics, and peer controls reduces the racial gaps markedly, but the female gap nearly doubles. Students whose designated guardian is the father have, on average, 0.10–0.20 higher test scores than do students with other guardians, but these gaps decline substantially with other controls. Math class size has a positive and significant relationship with test scores that becomes negative and statistically significant once we include the col. 3 controls.

borhood.[23] The neighborhood measures are based on Census data for a student's residential census tract and include median family income, median house value, and the fraction of adults that fall into five education categories. These controls are meant to proxy for unobserved parental influences. Again, like many of the student controls, the value-added framework should, for example, account for permanent income gaps but not for differences in student growth rates by parental income or education. Finally, the math class peer characteristics are the 10th, 50th, and 90th percentiles of absences, as a measure of disruption in the classroom, and the same percentiles of eighth-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.[24]

Adding student, peer, and neighborhood covariates reduces the adjusted standard deviation to 0.24, roughly two-thirds the size of the naive estimates reported in column 1.[25] 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 large and statistically significant.

Once again, transition matrices for the full control specification clearly reject random quality draws. The quartile-transition matrix is reported in

[23] We also experiment with additional controls for student ability, including eighth-grade reading scores, sixth- and seventh-grade math scores, higher-order terms (square and cube) and splines in the eighth-grade math score, and the variance in sixth to eighth-grade math scores. Compared to the col. 3 baseline, the largest impact is from adding the higher-order terms in eighth-grade scores. This reduces the adjusted standard deviation by just under 0.03. When school fixed effects are also included, the largest impact of any of these specification adjustments is half that size.

[24] 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 10th, 50th, and 90th percentile of peer absences are 0.009 (0.005), −0.002 (0.002), and −0.002 (0.0007). Thus it appears that the main statistically significant association between own performance and peer absences is from the most absent of students. The point estimates on the 10th, 50th, and 90th percentile of eighth-grade math scores are 0.028 (0.013), 0.140 (0.025), and 0.125 (0.019). These peer measures reduce the student's own eighth-grade math test score influence by 17% and suggest that high performers are most associated with a student's own performance.

[25] Arguably, part of the reduction in variance is excessive, as teachers may affect academic performance through an effect on absences or GPA. About half of the reduction in teacher dispersion between cols. 2 and 3 (adding peer and own student performance and schooling measures) is due to peer measures. That said, when we identify teacher effects within-school, peer measures have little additional power in explaining teacher quality dispersion.

Table 7
Quartile Rankings of Estimated Teacher Effects in Years *t* and *t* + 1: Percent of Teachers by Row

Quartile in year *t*:	Quartile in Year *t* + 1			
	1	2	3	4
1	33	32	16	19
2	32	25	31	13
3	17	25	33	26
4	15	21	23	41

NOTE.—χ^2 test of random quartile assignment: $p < .001$. 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 (col. 3 of table 6).

table 7. Forty-one percent of teachers ranking in the top 25% in one year rank in the top 25% in the following year. Another 23% slip down one category, 21% two categories, and 15% to the bottom category.[26]

D. Within-School Estimates

Within-school variation in teacher quality is often preferred to the between-school variety as it potentially eliminates time-invariant school-level factors. In our case, since we are looking over a relatively short window (3 years), this might include the principal, curriculum, school size or composition, quality of other teachers in the school, and latent family or neighborhood-level characteristics that can influence school choice. Because our results are based on achievement gains, we are generally concerned only with changes in these factors. Therefore, restricting the source of teacher variation to within-school differences will result in a more consistent, but less precisely estimated, measure of the contribution of teachers.

Our primary method of controlling for school-level influences is school fixed effects. As mentioned above, identification depends on differences in the intensity of students' exposure to different teachers within schools, as well as teachers switching schools during the sample period.[27] We report these results in columns 4 and 5 of table 6. Relative to the analogous columns without school fixed effects, the dispersion in teacher quality and precision of the estimates decline. For example, with the full set of student controls, the adjusted standard deviation drops from 0.24 (col. 3)

[26] Twenty-six percent and 19% of those in the top and bottom deciles remain the next year. Nineteen percent and 14% rated in the top and bottom deciles in 1997 are still there in 1999. Again, turnover is 15% higher among the lowest performing teachers. Adjusting for this extra turnover , the *p*-value on the bottom decile transition *F*-test drops from 0.14 to 0.06.

[27] Of the teachers with at least 15 student-semester observations, 69% appear in one school over the 3 years and 18% appear in two schools. Additionally, 13%–17% of teachers in each year show up in multiple schools.

Table 8
Correlation between Teacher Quality Estimates across Specifications

	Minimum Number of Student-Semesters Required to Identify a Teacher		
Specification Relative to Baseline	15 (1)	100 (2)	200 (3)
(0) Baseline	1.00	1.00	1.00
(1) Drop neighborhood covariates	1.00	1.00	1.00
(2) Drop peer covariates	.97	.98	.99
(3) Drop additional student covariates	.92	.93	.94
(4) Drop basic student covariates	.99	1.00	1.00
(5) Drop basic and additional student, peer, and neighborhood characteristics	.88	.85	.87
(6) Drop school fixed effects	.86	.68	.65
(7) Drop school fixed effects and basic and additional student, peer, and neighborhood characteristics	.62	.44	.45
Number of teachers	783	317	122

NOTE.—The col. 1 baseline corresponds to the results presented in col. 5 of table 6. Columns 2 and 3 correspond to the results presented in table 9, cols. 2 and 3, respectively. All specifications include the eighth-grade math test score, teacher student-semester counts, year indicators, the ninth-grade test level, the eighth-grade test level, an indicator equal to one if the information on eighth-grade test schore level is missing, and a constant. The baseline specification additionally includes basic and additional student characteristics, neighborhood and peer characteristics, and school fixed effects. All other specifications include a subset of these controls as noted in the table. See table 6 for the specific variables in each group.

to 0.19 (col. 5), roughly one-half the impact from the unadjusted value-added model reported in column 1. Again, an *F*-test rejects that the within-school teacher quality estimates jointly equal zero at the 1% level. We have also estimated column 4 and 5 models when allowing for school-specific time effects, to account for changes in principals, curricula, and other policies, and found nearly identical results. The adjusted standard deviations are 0.23 and 0.18, respectively, just 0.01 lower than estimates reported in the table.

Notably, however, once we look within schools, sampling variation accounts for roughly one-third of the unadjusted standard deviation in teacher quality. Furthermore, sampling variation becomes even more problematic when we estimate year-to-year transitions in quality, as in tables 5 and 7, with specifications that control for school fixed effects.

E. Robustness of Teacher Quality Estimates across Specifications

One critique of using test score based measurements to assess teacher effectiveness has been that quality rankings can be sensitive to how they are calculated. We suspect that using measures of teacher effectiveness that differ substantially under alternative, but reasonable, specifications of equation (2) will weaken program incentives to increase teacher effort in order to improve student achievement. To gauge how sensitive our results are to the inclusion of various controls, table 8 reports the robustness of

the teacher rankings to various permutations of our baseline results (col. 5 of table 6). In particular, we present the correlations of our teacher quality estimate based on our preferred statistical model—which controls for school fixed effects as well as student, peer, and neighborhood characteristics—with estimates from alternative specifications.

Because the estimation error is likely to be highly correlated across specifications, we calculate the correlation between estimates using empirical Bayes estimates of teacher effects (e.g., Kane and Staiger 2002). We rescale the OLS estimates using estimates of the true variation in teacher effects and the estimation error as follows:

$$\tau_j^* = \hat{\tau}_j \cdot \frac{\sigma_\tau^2}{\sigma_\tau^2 + \hat{\sigma}_\varepsilon^2}, \tag{5}$$

where $\hat{\tau}_j$ is our OLS estimate of the value added by teacher j, σ_τ^2 is our estimate of the true variation in teacher effects (calculated as described above), and $\hat{\sigma}_\varepsilon^2$ is the noise associated with the estimate of teacher j's effect, namely, the estimation error for $\hat{\tau}_j$. To further minimize concern about sampling variability, we also look at correlations across specifications estimated from samples of teachers that have at least 100 or 200 students during our period.

In rows 1–4, we begin by excluding, in order, residential neighborhood, peer, student background, and student performance covariates. Individually, each of these groups of variables has little impact on the rankings. Teacher rankings are always correlated at least 0.92 with the baseline. Even when we drop all of the right-hand-side covariates, except school fixed effects, row 5 shows that the teacher ranking correlations are still quite high, ranging from 0.85 to 0.88.

Only when school fixed effects are excluded is there a notable drop in the correlation with the baseline. In row 6, we exclude school fixed effects but leave the other covariates in place. The teacher quality correlation falls to between 0.65 and 0.86. Excluding the other right-hand-side covariates causes the correlation to fall to between 0.44 and 0.62. That is, without controlling for school fixed effects, rankings become quite sensitive to the statistical model. But as long as we consider within-school teacher quality rankings using a value-added specification, the estimates are highly correlated across specifications, regardless of the other controls included.

Importantly, our results imply that teacher rankings based on test score gains are quite robust to the modeling choices that are required for an individual principal to rank her own teachers. But a principal may have more difficulty evaluating teachers outside her home school. More generally, value-added estimates that do not account for differences across

schools may vary widely based on specification choices which in turn may weaken teacher performance incentives.

F. Additional Robustness Checks

Thus far, we have found that teacher quality varies substantially across teachers, even within the same school, and is fairly robust across reasonable value-added regression specifications. This section provides additional evidence on the robustness of our results to strategic test score reporting, sampling variability, test normalization, and the inclusion of other teachers in the math score production function.

1. *Cream Skimming*

One concern is that teachers or schools discourage some students from taking exams because they are expected to perform poorly. If such cream skimming is taking place, we might expect to see a positive correlation between our teacher quality measures τ_j and the share of teacher j's students that are missing ninth-grade test scores. In fact, we find that this correlation is small (-0.02), opposite in sign to this cream-skimming prediction, and not statistically different from zero.

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 only 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.[28] In this case, evidence is consistent with gaming of scores; the correlation is positive (0.07) and statistically different from zero at the 6% 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 6,361 students (12% of the full sample) but had no substantive impact on our results.

2. *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, 100, or 200 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

[28] The student test file includes an indicator for whether the student's test score may be excluded from reported school or citywide test score statistics because, e.g., the student is a special or bilingual education student.

Table 9
Further Evidence on the Distribution of the Estimated Teacher Effects

	Student Threshold			Test Scores Measured in Percentiles	Trimming Top and Bottom 3% in Changes
	50 (1)	100 (2)	200 (3)	(4)	(5)
Dependent variable mean	9.21	9.21	9.21	37.88	9.08
Mean test score gain	1.14	1.14	1.14	−2.08	1.06
Number of teachers	508	317	122	783	773
Number of students	52,957	52,957	52,957	52,957	50,392
Without school effects:					
Standard deviation of teacher effects	.233	.227	.193	2.66	.262
Adjusted standard deviation	.205	.211	.180	2.06	.203
Weighted standard deviation	.223	.216	.188	2.22	.211
p-value, *F*-test for teacher effects	.000	.000	.000	.000	.000
With school effects:					
Standard deviation of teacher effects	.192	.183	.154	2.57	.244
Adjusted standard deviation	.143	.155	.133	1.75	.161
Weighted standard deviation	.182	.176	.152	2.04	.188
p-value, *F*-test for teacher effects	.000	.000	.000	.000	.000

NOTE.—See notes to table 6. All regressions include the student, peer, and neighborhood covariates included in the table 6, cols. 3 and 5, specifications.

the teachers are, particularly since we overlook an important sample of teachers—new instructors with upcoming tenure decisions—in addition to teachers who teach multiple grades or nonmath classes. Finally, sampling variation exists with large numbers of students as well, so we would not expect to completely offset concerns about sampling error by simply setting a high minimum count of student-semesters.

Panel A of table 9 includes all covariates from the specification presented in column 3 of table 6. Panel B adds school fixed effects (i.e., col. 5 of table 6). Using a 50, 100, or 200 student-semester threshold, we find that the adjusted standard deviation is roughly 0.18–0.21 without school fixed effects and 0.13–0.15 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 200 increases the average number of student-semesters per teacher from 109 to 284. Consequently, sampling variability drops substantially, from an adjustment of 0.081 (0.273 − 0.192) for the 15-student threshold to 0.021 (0.155 − 0.134) for the 200-student threshold.

3. *More on Test Score Normalization and the Undue Influence of Outliers*

The remaining columns of table 9 include attempts to minimize the influence of outlier observations. Column 4 reports findings using national percentile rankings that are impervious to the normalization problem inherent in grade-equivalent scores.[29] We find that the adjusted standard deviation of $\hat{\tau}_j$ is 1.75 percentile points, a result that is statistically and educationally significant and broadly consistent with the grade-equivalent results.

In the next column, we simply trim the top and bottom 3% of the distribution of eighth- to ninth-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 score distribution. Still, the adjusted teacher standard deviation remains large in magnitude and statistically significant at 0.16 grade equivalents.[30]

4. *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 that 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 τ. Moreover, the English teachers may help us gauge the importance of teacher externalities, that is, the proverbial superstar teacher who inspires students to do well not just in her 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,

[29] 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. Less than 2% of the sample is censored, of which over 98% 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 standard deviation of the $\hat{\tau}_j$.

[30] We have also tried 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.

Table 10
The Distribution of the Estimated Math Teacher Effects When English Teachers Are Included

	Teacher Quality Estimates		
	Math Only (1)	Math and English (2)	English Only (3)
Math teachers:			
Standard deviation	.273	.278	
Adjusted standard deviation	.193	.170	
Weighted standard deviation	.213	.208	
Number of math teachers	783	783	
English teachers:			
Standard deviation		.257	.254
Adjusted standard deviation		.075	.113
Weighted standard deviation		.208	.209
Number of English teachers		1,049	1,049
p-value, *F*-statistic for math teacher effects	.000	.000	
p-value, *F*-statistic for English teacher effects		.000	.000

NOTE.—See notes to table 6. There are 52,957 students in each specification. Column 1 is the same as col. 5 of table 6. Column 2 additionally includes controls for the English teachers, while col. 3 only controls for English teachers.

we are primarily interested in testing the robustness of our math teacher effects to such controls.

We report estimates that condition on English teachers in table 10. For additional comparison, we also report standard deviations of the English teacher effect estimates both with and without controls for the math teachers. Controlling for English teachers, the math teacher adjusted standard deviation is roughly 0.02 grade equivalents smaller and less precisely estimated. Yet 88% of the math teacher impact remains. However, the size of the English teacher effect is noteworthy on its own. While it is less than half the size (0.075 vs. 0.170) of the dispersion in math teacher quality, it appears to be educationally important. Analogously, when we redo the analysis on reading scores (not reported), the adjusted standard deviation for English teachers is again only slightly smaller, 0.17 versus 0.15 grade equivalents, when we control for other (in this case, math) teachers. Furthermore, the size of the adjusted standard deviation for math teachers is quite notable, roughly 0.12 grade equivalents. Arguably, reading tests are less directly tied to an individual subject. Nevertheless, these results suggest two general conclusions. First, our quality measures, both for math and English teachers, are generally robust to controls for additional teachers. But, second, future research could explore why there are such large achievement effects estimated for teachers whom one would not expect to be the main contributors to a subject area's learning. Can

this be explained by sorting or does a teacher's influence reach beyond his or her own classroom?[31]

G. Heterogeneity by Ability Level, Race, and Sex

Table 11 explores the importance of teachers for different student groups. In columns 1–3, we look at teacher dispersion for students of different "ability." We stratify the sample into ability groups based on the eighth-grade math test score and reestimate the teacher effects within ability group. Low-ability students are defined as those in the bottom one-third of the Chicago public school eighth-grade score distribution, at or below 7.5 grade equivalents. Low-ability students have a mean test score gain of 0.54 grade equivalents. High-ability students are in the top one-third of the eighth-grade test score distribution, with scores above 8.7 (i.e., performing at or above national norms). These students have mean test score gains of 2.2 grade equivalents. All other students are classified as "middle" ability. The middle group has an average gain of 0.67 grade equivalents. 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 test format and 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 one-third to one-half of the standard deviation in estimated teacher effects to sampling variability. That said, a one standard deviation improvement in teacher quality is still worth a sizable gain in average test score growth: 0.13, 0.20, and 0.13 grade equivalents for low-, middle-, and high-ability students. These outcomes are 24%, 29%, and 6% of average test score gains between eighth and ninth grade for each group, respectively.[32] In relative terms, the largest impact of teachers is felt at the lower end of the initial ability distribution. These results are not sensitive to refinements in the way previous test

[31] As one informal test, we controlled for own student absences to assess whether the mechanism by which English teachers might influence math test scores is to encourage (or discourage) students from attending school. However, we found that own absences have no impact on the dispersion of the English teacher fixed effects.

[32] 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) relative to middle- and high-ability students (1.3 and 1.4).

Table 11
Distribution of the Estimated Teacher Effects for Selected Student Subgroups

	Ability Level			Race/Ethnicity			Sex	
	Low (1)	Middle (2)	High (3)	Non–African American, Non-Hispanic (4)	African American (5)	Hispanic (6)	Male (7)	Female (8)
Mean gain	.54	.67	2.22	2.19	.86	1.19	1.22	1.06
Standard deviation	.236	.304	.274	.259	.293	.248	.303	.264
Adjusted standard deviation	.129	.196	.132	.105	.201	.132	.201	.160
p-value, *F*-statistic for teacher effects	.000	.000	.000	.003	.000	.000	.000	.000
Number of teachers	518	478	390	204	579	353	627	620
Number of students	16,880	18,616	17,461	6,940	29,750	16,271	25,299	27,658

NOTE.—See notes to table 6. Ability level is assigned in thirds based on the eighth-grade test score distribution. High-ability students have scores above 8.7, middle-ability students have scores between 7.5 and 8.7, and low-ability students have scores of less than 7.5. All regressions include school fixed effects and the student, peer, and neighborhood covariates included in the table 6, cols. 3 and 5, specifications.

score results are controlled, including allowing for nonlinearities in the eighth-grade score or controlling for sixth- and seventh-grade scores.

By race, teachers are relatively more important for African American and, to a lesser extent, Hispanic students. A one standard deviation, one semester increase in teacher quality raises ninth-grade test score performance by 0.20 grade equivalents (23% of the average annual gain) for African American students and 0.13 grade equivalents (11% of the average annual gain) for Hispanic students. The difference is less important for non–African American, non-Hispanic students both because their mean test score gain is higher and because the estimated variance in teacher effects is somewhat smaller.

There is very little difference in the estimated importance of teachers when we look at boys and girls separately. The adjusted standard deviation of teacher effects equals 0.20 for boys and 0.16 for girls. For both girls and boys, a one standard deviation improvement in teacher quality translates into a test score gain equal to 15%–16% of their respective average annual gains.

Finally, we examined whether quality varies within teacher depending on the initial ability of the student. That is, are teachers that are most successful with low-ability students also more successful with their high-ability peers? To examine this issue, we use the 382 math teachers in our sample that have at least 15 students in both the top half and bottom half of the eighth-grade math test score distribution. We then explored whether teachers ranked in the bottom (or top) half of the quality rankings when using low-ability students are also ranked in the bottom (or top) half of the ability distribution when using high-ability students. We find that 67% of low-ranking teachers for low-ability students are low-ranking teachers for high-ability students. Sixty-one percent of those teachers ranked in the top half using low-ability students are ranked similarly for high-ability students. The correlation between the teacher quality estimates derived from low- and high-ability teachers is a highly statistically significant 0.39, despite small sample sizes that accentuate sampling error. Therefore, there is some evidence that teacher value added is not specific to certain student types; a good teacher performs well, for example, among both low- and high-ability students.

V. Predicting Teacher Quality Based on Resume Characteristics

This final section relates our estimates of τ_j to measurable characteristics of the instructors available in the CPS administrative records. Observable teacher characteristics include demographic and human capital measures such as sex, race, potential experience, tenure at the CPS, advanced degrees (master's or PhD), undergraduate major, undergraduate college attended,

and teaching certifications.[33] We report select results in table 12. All are based on the full control specification reported in column 5 of table 6. We discuss common themes below.

First and foremost, the vast majority of the total variation in teacher quality is unexplained by observable teacher characteristics. For example, a polynomial in tenure and indicators for advanced degrees and teaching certifications explain at most 1% of the total variation, adjusting for the share of total variation due to sampling error.[34] That is, the characteristics on which compensation is based have extremely little power in explaining teacher 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 result. In all cases, the R^2 never exceeds 0.08.

Given a lack of compelling explanatory power, it is of little surprise that few human capital regressors are associated with teacher quality.[35] Standard education background characteristics, including certification, advanced degrees, quality of college attended, and undergraduate major, are loosely, if at all, related to estimated teacher quality. Experience and tenure

[33] Potential experience is defined as age − education − 6 and is averaged over the 3 years of the sample.

[34] The R^2 is an understatement of the explanatory power since a significant fraction, perhaps up to a third, of the variation in $\hat{\tau}_j$ is due to sampling error. If we simply multiply the total sum of squares by a rather conservative 50% to account for sampling variation, the R^2 will double. However, in all cases it is never higher than about 15%. By comparison, the R^2 from a wage regression with education, experience, gender, and race using the 1996–99 Current Population Survey is about 0.2, without any corrections for sampling variation. Furthermore, firm-specific data or modeling unobserved person heterogeneity causes the R^2 on productivity and wage regressions to be quite a bit higher (e.g., Abowd, Kramarz, and Margolis 1999; Lazear 1999).

[35] 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 et al. (2005) also find no link between education level and teacher quality, although they find a small positive relationship between the first 2 years of teacher experience and teacher quality. Kane et al. (2006) find a positive experience effect in the first few years as well. 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 that 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, but Kane et al.'s (2006) results suggest otherwise. Other work on teacher training programs is likewise mixed (e.g., Angrist and Lavy 2001; Jacob and Lefgren 2004).

Table 12
Impact of Observable Characteristics on Teacher Fixed Effects

	(1)	(2)	(3)
Female		.073*	.069*
		(.020)	(.020)
Asian		.007	.008
		(.041)	(.041)
Black		.050*	.048*
		(.023)	(.023)
Hispanic		−.057	−.056
		(.039)	(.039)
Potential experience		.004	
		(.008)	
Squared		.000	
		(.000)	
Cubed (divided by 1,000)		.004	
		(.007)	
Potential experience<= 1:			.021
			(.042)
Master's	.002	.004	.007
	(.020)	(.020)	(.020)
PhD	−.103	−.077	−.068
	(.077)	(.076)	(.076)
BA major: education	.003	−.012	−.016
	(.030)	(.034)	(.033)
BA major: math	.003	.022	.021
	(.024)	(.025)	(.025)
BA major: science	.001	.029	.035
	(.040)	(.040)	(.040)
Certificate, bilingual education		−.067*	−.069*
		(.037)	(.037)
Certificate, child		.121	.120
		(.082)	(.082)
Certificate, elementary		.004	.006
		(.038)	(.038)
Certificate, high school		−.033	−.033
		(.033)	(.032)
Certificate, special education		.007	.008
		(.037)	(.036)
Certificate, substitute		−.004	−.005
		(.026)	(.026)
Tenure at CPS	−.001	−.001	.003
	(.008)	(.010)	(.009)
Squared	.000	.000	.000
	(.001)	(.001)	(.001)
Cubed (divided by 1,000)	.004	.005	.009
	(.011)	(.012)	(.011)
BA university, *US News* 1		−.010	−.014
		(.037)	(.037)
BA university, *US News* 2		.013	.012
		(.037)	(.037)
BA university, *US News* 3		.004	.002
		(.029)	(.029)
BA university, *US News* 4		.003	.003
		(.038)	(.038)
BA university, *US News* 5		−.003	.002
		(.072)	(.072)
BA university, local		.008	.005
		(.023)	(.022)
Adjusted R^2	.005	.077	.074
Number of teachers with observables	589	589	589

NOTE.—The dependent variable is teacher quality estimated using the table 6, col. 5, 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.

* Significant at 10% level.

have little relation to τ_j when introduced in levels (unreported), higher order polynomials (col. 2), or as a discontinuous effect of rookie teachers (col. 3). We have also tried identifying experience and/or tenure effects from a specification that includes teacher-year fixed effects (rather than just teacher fixed effects) which allows us to use variation within teacher over time, using various combinations of intervals for experience and tenure (e.g., 0–3, 3–7, 7–10, 10 plus), and capping experience at 10 years. None of these adjustments show a large or statistically important effect for either tenure or experience. Rather, at best, it appears that there is a 0.02 grade-equivalent increase in quality over the first few years of experience that flattens and eventually recedes. Given our sample sizes, such an effect is impossible to precisely estimate.

Female and African American teachers are associated with test scores roughly 0.07 and 0.05 grade equivalents higher than male and white teachers. Some of this influence derives from students with similar demographics.[36] In particular, African American boys and girls increase math test scores by 0.067 (standard error of 0.037) and 0.042 (standard error of 0.034) grade equivalents in classrooms with an African American teacher rather than a white teacher. However, we do not find an analogous result for Hispanic student-teacher relationships. Across all student race groups, including Hispanics, math test scores are 0.05–0.10 grade equivalents lower in classrooms with Hispanic teachers.

Likewise, female teachers have a larger impact on female students, especially African Americans. African American girls increase math test scores by 0.066 (standard error of 0.032) grade equivalents when in a classroom with a female teacher. This compares to a 0.032 (standard error of 0.033) grade equivalent boost for boys. Because of small sample sizes, we cannot distinguish Hispanic boys from Hispanic girls, but among all Hispanic students, female teachers boost math test scores by 0.060 (standard error of 0.024) grade equivalents. All of these results are similar under simpler specifications that include only the race and/or gender of the teacher.

VI. 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 decade 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

[36] Goldhaber and Brewer (1997) find teacher quality higher among female and lower among African American instructors. Ehrenberg, Goldhaber, and Brewer (1995) and Dee (2004) 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.

makes a good teacher. Our results are consistent with related studies like Hanushek (1992) and Rivkin et al. (2005), who argue that characteristics that are not easily observable in administrative data are driving much of the dispersion in teacher quality. Traditional human capital measures have few robust associations with teacher quality and explain a very small fraction of its wide dispersion. That our teacher quality measure persists over time 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 for whom little or no information is available on the teacher's effect on students' test score achievement. More generally, teacher quality rankings can be quite sensitive in a value-added framework when across-school differences are ignored. Without such controls, naive application of value added may undermine teacher performance incentives. 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 serious attention to implementation problems (Murnane et al. 1991), including, but far from limited to, important measurement issues associated with identifying quality.

Data Appendix

The student administrative records assign an eight-character identification code to teachers. The first three characters are derived from the teacher's name (often the first three 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 and birth date are the same teacher and adjust accordingly. Additionally, for position numbers that appear to change over time in the student data, we made assumptions about whether it was likely to be the same teacher based on the presence of the teacher in that school in a particular year in the teacher administrative data.

Finally, we match students to teachers using a three-letter name code and the position number for the combinations that are unique in the teacher data.[37]

References

Abowd, John M., Francis Kramarz, and David Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67, no. 2:251–333.

Angrist, Joshua D., and Victor Lavy. 2001. Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics* 19, no. 2:343–69.

Borjas, George J. 1987. Self-selection and the earnings of immigrants. *American Economic Review* 77, no. 4:531–53.

Borjas, George J., and Glenn T. Sueyoshi. 1994. A two-stage estimator for probit models with structural group effects. *Journal of Econometrics* 64, no. 1–2:165–82.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2004. Teacher sorting, teacher shopping, and the assessment of teacher effectiveness. Unpublished manuscript, Public Policy Studies, Duke University.

Coleman, James S., et al. 1966. *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Dee, Thomas S. 2004. Teachers, race, and student achievement in a randomized experiment. *Review of Economics and Statistics* 86, no. 1: 195–210.

Ehrenberg, Ronald G., Daniel D. Goldhaber, and Dominic J. Brewer. 1995. Do teachers' race, gender, and ethnicity matter? *Industrial and Labor Relations Review* 48, no. 3:547–61.

Ferguson, Ronald. 1998. Paying for public education. *Harvard Journal of Legislation* 28:465–98.

Goldhaber, Dan D., and Dominic J. Brewer. 1997. Why don't school and teachers seem to matter? *Journal of Human Resources* 32, no. 3:505–23.

Greenwald, Rob, Larry Hedges, and Richard Laine. 1996. The effect of school resources on student achievement. *Review of Educational Research* 66:361–96.

Grogger, Jeff, and Eric Eide. 1995. Changes in college skills and the rise in the college wage premium. *Journal of Human Resources* 30, no. 2: 280–310.

Hanushek, Eric A. 1971. Teacher characteristics and gains in student achievement. *American Economic Review* 61, no. 2:280–88.

———. 1992. The trade-off between child quantity and quality. *Journal of Political Economy* 100, no. 1:84–117.

[37] Note that we assigned some three-letter teacher codes for cases in which the teacher code did not correspond to the first three letters of the teacher's last name.

———. 1996. Measuring investment in education. *Journal of Economic Perspectives* 10, no. 4:9–30.

———. 1997. Assessing the effects of school resources on student performance: An update. *Education Evaluation and Policy Analysis* 19: 141–64.

———. 2002. Publicly provided education. In *Handbook of public finance*, vol. 4, ed. Alan Auerbach and Martin Feldstein. Amsterdam: North-Holland Press.

Hanushek, Eric A., and Dennis D. Kimko. 2000. Schooling, labor-force quality, and the growth of nations. *American Economic Review* 90, no. 5:1184–1208.

Hoxby, Caroline. 2000. Peer effects in the classroom: Learning from gender and race variation. Working paper no. 7867, National Bureau of Economic Research, Cambridge, MA.

Jacob, Brian A., and Lars Lefgren. 2004. The impact of teacher training on student achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources* 39, no. 1: 50–79.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118, no. 3:843–77.

Jepsen, Christopher, and Steven Rivkin. 2002. What is the tradeoff between smaller classes and teacher quality? Working paper no. 9205, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2006. What does certification tell us about teacher effectiveness? Evidence from New York City. Working paper no. 12155, National Bureau of Economic Research, Cambridge, MA.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promises and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16, no. 4:91–114.

———. 2005. Using imperfect information to identify effective teachers. Working paper, Department of Economics, Dartmouth University.

Lazear, Edward. 1999. Personnel economics: Past lessons and future directions: Presidential address to the Society of Labor Economists, San Francisco, May 1, 1998. *Journal of Labor Economics* 17, no. 2:199–236.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *Review of Economic Studies* 60, no. 3:531–42.

Moulton, Brent. 1986. Random group effects and the precision of regression estimates. *Journal of Econometrics* 32:385–97.

Murnane, Richard. 1975. *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger.

Murnane, Richard, Judith Singer, John Willett, James Kemple, and Randall Olsen. 1991. *Who will teach? Policies that matter*. Cambridge, MA: Harvard University Press.

Rivers, June, and William Sanders. 2002. Teacher quality and equity in educational opportunity: Findings and policy implications. In *Teacher quality*, ed. Lance T. Izumi and Williamson M. Evers. Stanford, CA: Hoover Institution Press.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73, no. 2:417–58.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94, no. 2:247–52.

Sacerdote, Bruce. 2001. Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics* 116, no. 2: 681–704.

Summers, Anita A., and Barbara L. Wolfe. 1977. Do schools make a difference? *American Economic Review* 67, no. 4:639–52.

U.S. Census Bureau. Census 2000, summary file 1. Generated by authors using American FactFinder, http://factfinder.census.gov, accessed October 23, 2006.

U.S. Department of Education. National Center for Education Statistics. 2000. *The condition of education 2000*. NCES publication no. 2000–062. Washington, DC: U.S. Government Printing Office.

U.S. Department of Education. National Center for Education Statistics. 2003. *Characteristics of the 100 largest public elementary and secondary school districts in the United States: 2001–02*. NCES 2003-353, Jennifer Sable and Beth Aronstamm Young. Washington, DC: U.S. Government Printing Office.

U.S. News & World Report. 1995. *America's best colleges*. Washington, DC: U.S. News & World Report.