NBER WORKING PAPER SERIES

WHAT DOES CERTIFICATION TELL US ABOUT TEACHER EFFECTIVENESS? EVIDENCE FROM NEW YORK CITY

Thomas J. Kane
Jonah E. Rockoff
Douglas O. Staiger

Working Paper 12155
http://www.nber.org/papers/w12155

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2006

We are grateful to a number of persons whose efforts made this project possible. Vicki Bernstein, Bruce Groneveldt, Dominic Mancello, Lori Mei, Cathi Schoenberger, Henry Solomon, and Harvey Weinstein at the New York City Department of Education were generous with their time in putting together data on our behalf. Timothy Daly, Michelle Rhee, and Andrew Sokatch of The New Teacher Project and Abigail Smith and Cynthia Skinner of Teach for America were extremely helpful in answering our questions regarding these programs. We thank Caroline Hoxby and Andrew Leigh for providing us with data on college median SAT scores. Dan Obus provided excellent research assistance. We would also like to thank seminar participants at the New York City Department of Education, Columbia University and the NBER summer institute for helpful comments and suggestions. E-mail: jonah.rockoff@columbia.edu, kaneto@gse.harvard.edu, douglas.o.staiger@dartmouth.edu. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City
Thomas J. Kane, Jonah E. Rockoff and Douglas O. Staiger
NBER Working Paper No. 12155
April 2006, Revised July 2006
JEL No. I2, J0

ABSTRACT

We use six years of data on student test performance to evaluate the effectiveness of certified, uncertified, and alternatively certified teachers in the New York City public schools. On average, the certification status of a teacher has at most small impacts on student test performance. However, among those with the same certification status, there are large and persistent differences in teacher effectiveness. This evidence suggests that classroom performance during the first two years, rather than certification status, is a more reliable indicator of a teacher's future effectiveness. We also evaluate turnover among teachers with different certification status, and the impact on student achievement of hiring teachers with predictably high turnover. Given relatively modest estimates of experience differentials, even high turnover groups (such as Teach for America participants) would have to be only slightly more effective in their first year to offset the negative effects of their high exit rates.

Thomas J. Kane
Harvard University
Graduate School of Education
Gutman Hall 455
Appain Way
Cambridge, MA 02138
and NBER
kaneto@gse.harvard.edu

Jonah E. Rockoff
Columbia Business School
3022 Broadway, Uris Hall 603
New York, NY 10027
jr2331@columbia.edu

Douglas O. Staiger
Dartmouth College
Department of Economics
HB6106, 301 Rockefeller Hall
Hanover, NH 03755-3514
and NBER
douglas.o.staiger@dartmouth.edu

1. Introduction

Federal and state governments have traditionally regulated teacher quality with *ex ante* certification requirements. To gain legal permission to teach, individuals are generally required to study full-time for one or two years in an approved education program. However, recruiting difficulties have forced many districts to hire large numbers of uncertified or alternatively certified teachers. Despite the ubiquity of alternative teacher certification (AC) programs across the country, there is little research on the relative quality of certified, uncertified, and AC teachers. This is particularly regrettable given that AC teachers are more likely to work in urban areas with low-income and low-achieving students. In this paper, we aim to fill this gap in the literature by providing new evidence on how the certification of teachers in a large, urban school district relates to teachers' impacts on student achievement.

Forty-seven (47) states and the District of Columbia currently have AC programs, the first of which were created in the early 1980s.[1] An estimated 35,000 AC teachers are hired each year, about one third of all new teachers nationwide. Although AC programs often differ in other ways, participants are generally required to possess a bachelor's degree, pass state licensing exams, participate in a special training or mentoring program, and enroll in a teacher education program as they begin teaching.

There exist only a few high-quality studies of AC teachers, most notably Decker et al. (2004), who evaluate the Teach for America (TFA) program, a non-profit entity that recruits and sends AC teachers to districts throughout the nation. They find that teachers recruited through

TFA (TFA corps members) are significantly more effective than both uncertified and certified teachers at math instruction and statistically indistinguishable in reading instruction. This study is particularly notable because the schools they examine agreed to randomly allocate students and teachers across classrooms.

Still, one must be cautious in using the results from this and other studies of TFA to draw conclusions regarding AC teachers and programs more generally. Although TFA is a highly visible program, it is also unique in many respects. It is highly selective, has a national scope, and recruits individuals who commit to teach for only two years. Indeed, the stated mission of TFA is "to build the movement to eliminate educational inequity" by recruiting individuals who will "become lifelong leaders in the effort to expand educational opportunity" (see www.teachforamerica.org). In addition, existing studies of TFA focus on a very small number of schools and teachers.[2] Though generalizations are limited in all studies—ours included—we use a much larger sample of schools and teachers in a district that employs teachers with many different types of certification, including TFA corps members.

We examine the relationship between teacher impacts on student achievement and certification status among teachers in New York City. Besides being the largest and one of the

[1] The National Center for Educational Information has assembled a detailed accounting of AC programs in the United States (Feistritzer (2005)). The term emergency certification is often used synonymously with alternative certification, and both refer to special programs to facilitate the employment of individuals who would otherwise be uncertified. However, Feistritzer suggests that they generally differ in educational requirements and support given to teachers. For example, most alternative certification programs provide training or mentoring to their teachers, whereas emergency certification programs generally do not.

most diverse school districts in the country, New York is a major employer of certified, uncertified, and alternatively certified teachers. More than 50,000 new teachers were hired in New York schools during the school years 1999-2000 to 2004-2005. Certified, uncertified and AC teachers accounted for, respectively, 46 percent, 34 percent, and 20 percent of these new hires. AC teachers in New York are recruited through multiple sources, but the vast majority comes from the New York City Teaching Fellows program. Teaching fellows formed about 28 percent of all new teachers in New York during the school years 2002-2003 through 2004-2005. A small but significant number of teachers come from other special programs, such as Teach for America. The city also hires certified teachers through a series of international recruitment initiatives. Together, Teach for America and international recruitment supplied New York with roughly 10 percent of its new teachers over this same period.

We measure the relative effectiveness of teachers using panel data on students' reading and math test scores that also identifies students' reading and math teachers. Specifically, we use students' test scores in grades four through eight to estimate the value-added of their teachers, controlling for students' prior-year test scores and a number of student, classroom, grade, and school related factors, and controlling for teachers' experience levels.

Boyd et al. (2005a) use similar data to evaluate differences in teacher effectiveness by initial certification status. They focus on between-group differences in student performance and

[2] Additional evidence on the effectiveness of TFA teachers also comes from studies by Raymond et al. (2001) and Darling-Hammond et al. (2004). The Decker et al. study includes 44 TFA teachers. The Raymond et al. study includes 117, and, though they do not report the number of TFA teachers, the number studied by Darling-Hammond et al. is likely to be considerably lower. They examine teachers in the same city over roughly the same time period, but they restrict their analysis to grades four and five, whereas Raymond et al. examine grades four to eight. These additional studies both use data from Houston, Texas, yet Raymond et al. conclude that TFA teachers are as good as or better than other teachers, while Darling-Hammond et al. find they are less effective than certified teachers, and perform about as well as other uncertified teachers. It is difficult to reconcile the two sets of results because of major distinctions between their empirical specifications and the dependent variables they examine.

differential effects of experience for Teaching Fellows and other groups of teachers. Our work differs from theirs in several important ways: First, we use an additional year of data from the school year 2004-05. Because of recent expansions in the Teaching Fellows and Teach for America programs, this additional year of data provides us with greater precision.[3] Second, we use application data from the Teaching Fellow program to study the role of various applicant traits on the likelihood of being interviewed, offered a job and accepting a position in the NYC public schools. Third, we estimate variation in teacher effectiveness within each certification group and interpret between-group differences in effectiveness in light of these within-group differences. Although there are statistically significant differences between groups, such differences are generally dwarfed by the differences in effectiveness within groups of teachers.[4] Fourth, since exiting teachers are generally replaced by novice teachers who still have a lot to learn on the job, we explore the implications of differing turnover rates by certification status for steady state differences in effectiveness between groups of teachers.

We find no difference between teaching fellows and certified teachers or between uncertified and certified teachers in their impact on math achievement. Classrooms of students assigned to internationally recruited teachers scored .02 standard deviations lower in math than similar classrooms assigned to certified teachers, while classrooms of students assigned to Teach for America corps members scored .02 standard deviations higher relative to certified teachers. (We measure teacher effectiveness in terms of test scores among New York City students, where test scores have been normalized by year and grade level to have a mean of zero and standard deviation of one.) In reading, students assigned to teaching fellows underperformed students

[3] In an earlier version of this paper Kane et al. (2005), we report a similar analysis using data through 2003-04.

assigned to certified teachers by .01 standard deviations. These are the only instances in which we find that a teacher's initial certification status has statistically significant implications for student achievement.

Consistent with other studies, we find that teachers' effectiveness improves with the first few years of experience. We estimate that achievement level of students assigned to teachers in their first year of teaching is .06 and .03 standard deviations lower in math and reading, respectively, as students assigned to those same teachers after they have gained two years of experience. Our results suggest that the returns to experience may be somewhat higher for teaching fellows than for certified teachers, at least in promoting reading achievement scores. Although teaching fellows underperform relative to regularly certified teachers in reading in their first year of teaching, they seem to close the gap by their third year.

Because they have not made the same pre-career investment in teacher training that traditionally certified teachers have made, one might fear that AC teachers would have a higher turnover rate. This could be costly in terms of student achievement (high turnover rates imply more novice teachers in steady state) as well as the increased direct costs of hiring. However, teaching fellows and traditionally certified teachers had very similar retention rates. Only Teach for America corps members had discernibly higher exit rates. Many TFA corps members leave after their two-year commitment is up. We estimate that the higher turnover among TFA corps members implies that roughly 45 percent of them will be in their first or second year of teaching in the steady state, relative to about 20 percent of traditionally certified teachers and teaching fellows. The high exit rate of Teach for America participants has been a primary target of the

[4] In a revised draft of their paper, Boyd et al. (2005b) come to a similar conclusion. However, they reach such a conclusion without ever estimating within-group variation in teacher effectiveness.

program's critics. However, the impact on student achievement of this higher turnover rate, while negative, is rather modest. We estimate that TFA corps members would have to produce about .02 standard deviations additional achievement in math and reading per year to offset the impact of their higher turnover rates in steady state. This is roughly the size of the differential we estimate for Teach for America participants (and considerably smaller than the positive impact estimated by Decker et. al.

Although the differences in teacher effectiveness between groups of teachers with different certification are small, there are large differences in teacher effectiveness within all of these groups. This latter finding is consistent with other research.[5] We estimate that the average value-added among the top quarter of elementary school math teachers is .33 standard deviations greater than the value-added among the bottom 25 percent—almost ten times the magnitude of any between-group difference! Thus, although shifting the mix of teachers with different types of certification does not appear to be a useful tool for improving student achievement, there is great potential for school districts to improve student achievement by selectively retaining only those who were estimated to be most effective during their first years of teaching.

2. Alternative Teacher Certification in New York City

Recruiting a sufficient number of certified teachers has been a long-standing problem for the New York City Department of Education (the DOE). In the school year 1999-2000, approximately 60 percent of all new teachers hired were uncertified. Recruiting difficulties were more severe in schools with low average achievement levels. Grouping schools by the fraction of students who passed end-of-year math examinations in 1999-2000, we find that 73 percent of new hires were uncertified in both elementary and high schools in the lowest deciles of pass-rates (Figure 1). Since that time, the DOE has taken a number of steps to decrease its use of uncertified personnel, one of which has been to expand its recruitment of alternatively certified teachers. Over the school years 1999-2000 to 2004-2005, the fraction of uncertified hires fell from 60 percent to 7 percent, while the fraction of AC hires rose from 2 percent to 36 percent. It is unlikely that this shift was just a re-labeling of individuals who would have otherwise become uncertified teachers. The populations of uncertified teachers and AC teachers differ on a number of characteristics (see Table 1).

The major source of growth of AC teachers has been the New York City Teaching Fellows Program (NYCTF), which was created in the summer of the year 2000.[6] The NYCTF program was created as a response to pressure from the state government to hire only certified teachers in the city's lowest performing schools. After the DOE failed to hire only certified teachers for these schools in the school year 1999-2000, it was sued by State Education Commissioner Richard Mills (Mills v. Levy, et al., Superior Court, Kings County, Index No. 26196/00). As a result, the number of teaching fellows hired grew from 350 in the school year 2000-2001 (less than 5 percent of new hires) to 2,500 in the school year 2003-2004 (more than

[5] Recent research has yielded remarkably consistent estimates of the heterogeneity in teacher impacts. For example, using data from two school districts in New Jersey, Rockoff (2004) reports that one standard deviation in teacher effects is associated with a .1 student-level standard deviation in achievement. Using data from Texas, Rivkin et al. (2005) report very similar estimates—suggesting that a standard deviation in teacher effectiveness is associated with .11 student-level standard deviations in math and .095 standard deviations in reading. Using data on high school students in Chicago Public Schools, Aaronson et al. (2003) find that a standard deviation in teacher effectiveness is associated with a .09 to .16 student-level standard deviation difference in performance.

30 percent of new hires) and 2,000 in the school year 2004-2005 (more than 25 percent of new hires). The expansion of the program was motivated by changes in New York State law in the year 2000 that required all teachers in the state to be certified in their subject of instruction by the school year 2003-04.[7]

New York recruits AC teachers through several other sources: Teach for America, the Peace Core Fellows Program, and the Teaching Opportunity Program Scholars. Teach for America, as mentioned above, is a national program, founded in 1990, that recruits individuals for AC teacher positions in numerous districts across the country. The Peace Corps Fellows program, created in 1985, is a partnership between the NYC Department of Education and Columbia Teachers College, and recruits individuals who have returned from Peace Corps volunteer service overseas. The Teaching Opportunity Program (TOP) is operated in partnership between the DOE and the City University of New York to recruit teachers of math, science, and Spanish. Finally, DOE also recruits teachers from other countries. International teachers have been certified in their home country and are not AC teachers. However, given their non-traditional recruitment, we consider them separately from certified teachers in our analysis.

Teaching fellows represent the vast majority of all AC teachers hired since 1999-2000, and we focus much of our attention on them. The number of TFA corps members and international teachers is substantial enough that we present results for them in our basic analysis. In contrast, there are only a handful of Peace Corps Fellows and TOP Scholars in our data, and we therefore do not discuss them further. We classify teachers based on their certification in the

[6] NYCTF is a partnership between the New York City Department of Education and The New Teacher Project (TNTP). TNTP is a national non-profit organization, founded in 1997, that helps school districts recruit AC teachers. TNTP has also worked with school districts in Miami, New Orleans, Oakland, Philadelphia, Washington, DC, and a number of other urban and rural communities.

[7] More information on certification regulations can be found at www.highered.nysed.gov/tcert/certificate/index.html.

year that they are hired. Thus, uncertified teachers who gain certification are still considered uncertified. This classification allows us to answer the question: what is the expected difference in teacher effectiveness between certified, uncertified, and AC teachers at the time they are hired?

Certified, uncertified, international, and AC teachers are observably different (Table 1). The fraction of teachers who are black or Hispanic is lower among regularly certified teachers and TFA corps members (about 20 percent) than among teaching fellows (30 percent), uncertified teachers (49 percent), or international teachers (48 percent). Certified teachers are also slightly more likely to be female than other groups (80 percent vs. about 70 percent). The median age of certified teachers, uncertified teachers and teaching fellows is roughly the same (about 28), but TFA corps members are considerably younger (median age 23) and international recruits considerably older (median age 36). Not surprisingly, certified teachers and international recruits are much more likely to have graduate education. Lastly, there are differences across groups in the selectivity of undergraduate institution. As measured by median SAT scores, certified and uncertified teachers attended significantly less selective colleges than teaching fellows or TFA corps members.[8]

There are also substantial differences in the characteristics of students taught by different groups of teachers (Table 2).[9] Uncertified teachers, teaching fellows, TFA corps members, and

[8] We have data from the DOE on the undergraduate institution attended for about 25 percent of teachers hired during our sample who are not teaching fellows or international recruits. We have this data for 95 percent of teaching fellows (98 percent of teaching fellows who attended a U.S. institution) because we have data from applications. We have this data for less than 1 percent of international recruits because so few attended U.S. institutions. The distribution of median SAT scores of teaching fellows with and without data from the DOE is very similar, leading us to believe that the availability of this data is not strongly related to the selectivity of undergraduate institution.

[9] We use school level averages to construct this figure. It therefore does not take into account any sorting of students within schools. In addition, we do not have school level data for the school year 2004-2005, and these teachers are assigned the characteristics of the students in their school from the prior year.

international teachers all tend to teach in schools that—relative to those employing certified teachers—have a higher fraction minority students, higher fraction eligible for free lunch (our best measure of household poverty), bilingual education, special education, lower fraction passing elementary and high school tests in reading and math, and lower high school graduation rates. In the results we present below, controlling for these differences among the students assigned to different groups of teachers will be important in drawing conclusions regarding their relative effectiveness.

2.1 Teaching Fellow Selection

In addition to our data on teaching fellows, we have data on the teaching fellow selection process for the school years 2003-2004 and 2004-2005. Though we cannot be certain that selectivity was constant across years, NYCTF officials estimate that the fraction of applicants who became teaching fellows was about 14 percent in the school year 2000-2001. This is about the same as the fraction who became teaching fellows in 2003-04 and 2004-05, even though the program was more than five times as large.

Teaching fellows are selected in a three-step process. First, applicants submit information on their demographic characteristics, their previous work experience, their academic history, a personal essay, and their qualifications for teaching particular subject areas. Approximately 60 percent of applicants are then invited for an interview. The interview process lasts approximately 4-5 hours and has multiple parts: applicants give a five minute lesson on a subject of their choosing, participate in a guided discussion, write an essay on a topic not given out in advance, and sit through a one-on-one interview. Approximately 50 percent of those interviewed have their applications forwarded to the final stage. These applications are reviewed by a committee, and about 85 percent are offered positions. Roughly 75 percent of offers are accepted, and 85 percent of those accepting offers complete the necessary training to gain alternative certification. Overall, less than 15 percent of all applicants become teaching fellows.

Teaching fellows' academic credentials are related to the probability of selection. Figure 2 shows how the selectivity of applicants' undergraduate institution (measured by median SAT scores) and applicants' grade point average relate to how they fare in the selection process. These figures are created using data from applications for the school year 2003-2004. Each panel shows the distribution of academic credentials for three groups: applicants who were screened out before the interview, applicants who were interviewed but rejected, and applicants who were offered a position. The average academic credentials of applicants increases noticeably as they pass through the selection process. Among those applicants who are screened out before the interview, the median quantitative SAT score of their undergraduate institution was at the median of the national distribution of colleges and universities. For the average applicant rejected after the interview or offered a position, the median quantitative SAT score of their undergraduate institution was at the 57th and 69th percentile respectively. The respective numbers for verbal SAT are 55th, 62nd, and 73rd percentiles. The same pattern occurs with applicants' reported undergraduate grade point average (GPA). The average among screened-out applicants is 2.7, while among those rejected at interview it is 3.0 and among those offered a position it is 3.2.

Among those offered a position, we find no relationship between academic credentials and initial acceptance or academic credentials and employment in the district. Thus, it does *not* appear that the selectivity of the teaching fellows program is being undone after offers have been given out.

Teaching fellows are required to attend a seven-week summer training course and to observe and assist in a summer school classroom. The focus of the summer course is on practical teaching methods, not subject material. Before they begin work, teaching fellows must have a bachelor's degree and pass teacher certification examinations.[10] During the summer training, they begin coursework in an approved graduate school of education and must progress towards gaining traditional, permanent certification within three years. Teaching fellows receive a tuition grant to help pay for their courses.[11]

Teaching fellows are assigned a subject area and one of ten geographic regions in which they must look for a teaching position. In practice, a few fellows each year find jobs outside of their assigned subject or region, and the DOE does not prohibit them from taking these positions. The DOE gives teaching fellows help in finding employment through multiple job fairs and, recently, online postings. Teaching fellows are no different than their colleagues with respect to union membership, salary, and general rights and privileges afforded to teachers in New York.

3. Data on Students and their Teachers in New York City

Our data on students consists of information on demographic background, attendance, suspensions from school, test performance, eligibility for free lunch, special education and

[10] For example, to teach mathematics under an alternative certification license, an individual must pass two exams: the New York State Teacher Certification Exam - Liberal Arts & Science Test (LAST) and a Content Specialty Test (CST) in Mathematics. To teach mathematics under a traditional certification license, an individual must pass an additional exam: the New York State Teacher Certification Exam - Secondary Assessment of Teaching Skills (ATS-W).

[11] In the program's initial years, all tuition was paid for by the DOE. In order to defray tuition costs, fellows now have $4,000 deducted from their salary over two years.

bilingual education, and a student identification number.[12] It also contains teacher identification numbers for students' math and reading teachers, which were often the same teacher for elementary school students. All student data is limited to grades three through eight, the grades in which students in New York City take standardized math and reading examinations. This data spans the school years 1998-1999 through 2004-2005.[13]

We match students to teachers using information from the DOE payroll system. Specifically, we have a snapshot of all teachers working in the DOE at the end of September, November, and May of each school year from September 1999 through September of 2005. This gives us information on teachers' certification, recruitment through various programs, and their position on a salary schedule.[14] We use the salary schedule variables to construct measures of teachers' education and experience. In addition, we have application information for all teaching fellows (including all applicants, not just those subsequently hired) for the school years 2003-2004 and 2004-2005. This provides us with information on applicants' undergraduate institution

[12] One minor flaw in the student data is that the number of students with missing free lunch status grows considerably at the end of our sample, from about 2 percent to 20 percent. This is due to schools with very high percentages of free lunch students enrolling in a Universal School Meals program. These schools, though predominantly populated by students on free lunch, often fail to report student data to the district because their funding for meals is unaffected by these reports. The growth in the number of students with missing data is mirrored by a drop in the number of students coded as eligible for free lunch. We do not drop students with missing free lunch status from our analysis. Instead, our free lunch status variables include dummies for eligibility and for missing data. Under an alternative specification, coding all students with missing data as eligible, we find almost exactly the same results as those presented below.

[13] We do not match third graders to their teachers because we use prior test scores as a control variable in our analysis, and this is only available for third graders who are repeating grades. We also do not examine reading test outcomes for grade seven in the school year 2001-2002 and grade eight in the school year 2002-2003, because no reading test scores are available for grade seven in the school year 2001-2002.

and GPA, and their progress through the application process.

Because our testing data for students in New York is limited to reading and math tests taken in grades three to eight, and because our empirical strategy requires a prior year score for every teacher, our analysis focuses on those in grades four through eight. (In grades four and five, most students have one teacher for both reading and math. In higher grades, we use the instructors listed for mathematics or reading courses.) Table 3 shows how teachers are selected into our analysis based on a number of criteria. The first row shows the overall number of teachers hired during our sample period by certification and program. The second row shows the number of teachers who worked in a school serving students in grades four to eight. We use school level data because information on grade level and subject for all teachers is not collected by the DOE. Roughly 20 percent of certified teachers, uncertified teachers, and teaching fellows were not teaching in one of these schools. In contrast, only 7 percent of TFA corps members and almost 35 percent of international teachers were not teaching in a school serving any children in grades four to eight.

In the third row of Table 3, we show the number of teachers who were matched to at least one student in grades four through eight in the reading and math test data. Overall, we were able to match 32 percent of all teachers to students with achievement data. Though this match rate may appear low, it is due principally to the number of teachers giving reading and math

[14] There are a small but significant number of coding errors in the salary schedule information from the payroll data. Most are a result of late updating of information regarding prior experience of incoming teachers. For example, there are a number of cases where a teacher is classified as having no experience in September of their first year, but is classified as having multiple years of experience by November. Teachers wishing to get credit for prior teaching experience must submit a form, and this may create delays in accounting for prior experience in the payroll records. We base our experience measures on the premise that the DOE does not overpay, but may underpay, its teachers based on experience. Thus, if a teacher has a discontinuous jump in their experience profile, we use their highest experience level to infer teaching experience in other years. We use the same method to impute experience in years when it is missing, so long as experience is non-missing for the same teacher at some point during our sample period.

instruction to students in grades four through eight. 95 percent of elementary school students and 82 percent of middle school students are matched to their math and reading teacher(s). The lower match rate for middle school students is due to our needing course identification information in order to match them with teachers, and roughly 20% of schools do not use the administrative system that reports this data.[15] We investigate the relationship between schools' reporting of course information and their average student characteristics, and we find no statistically significant relationship with race, free lunch eligibility, ELL and special education status, or prior test scores. (Results available upon request.)

At the bottom of the Table 3, we show the number of teachers that are included in our regression analysis. Overall, about 20 percent of teachers from each group are included in our regression sample; only TFA corps members have a slightly higher rate of inclusion (27 percent). We drop classrooms in which fewer than seven or more than 45 students are tested (corresponding to the 1st and 99th percentile), because we are concerned that classes with extremely low or high numbers of students may be incorrectly identified in our data. We drop classrooms where the teacher did not work in the school during the entire year and classrooms taught by teachers listed as working in more than one school per year. We also drop classrooms in any school-year cell for which less than 75 percent of students were successfully matched to a teacher. Finally, we drop classrooms where 25 percent or more of the students receive special education. In addition, the regression sample does not include students who were not tested in the prior year, since prior test scores are used as a control variable.

The mean characteristics of teachers are quite similar between our regression sample and

[15] This system, called "Automate the Schools" or ATS, is available to all schools but not universally used. It standardizes and automates the collection and reporting of student data at the school level. The DOE provides this (cont'd on next page)

the full sample of teachers (Table 4). The one noticeable difference is that teachers in the regression sample are more likely to be black. This is due mainly to the elimination of high school teachers, who are more likely to be white.

The distribution of students across groups of teachers is similar in our regression sample and the full sample of teachers in terms of race and the proportion receiving free/reduced price lunch (Table 5). However, students in our regression sample are substantially less likely to be ELL students or to receive special education than the overall city averages. The decrease in special education students is due to our dropping classrooms if more than 25 percent of its students receive special education. (Over 95 percent of special education students in New York attend such classrooms, most of which are filled entirely with special education students.) This criterion also is responsible for about half of the decline in ELL students; 20 percent of students in classes with more than 25 percent special education students are classified as English Language Learners. The other factor that drives this change is our requirement of at least seven tested students per classroom. More than 50 percent of ELL students—but only 6 percent of non-ELL students—did not take the reading exam in the years we observe them, and 30 percent of ELL students are observed in a classroom with fewer than seven tested students. Students in our regression sample are also more likely to pass their exams than the overall student population. This is driven almost entirely by the drop in ELL and special education population.[16]

4. Estimation of Teacher Effectiveness

system and training to schools, but does not mandate its use.

[16] A small part of this change is due to two other factors. First, the overall pass rate for New York City students was higher in 2004-2005 than in past school years. As noted above, because we lack school level data—used to calculate the full sample characteristics—from the school year 2004-2005, we assign those teachers the characteristics of students in their school for the prior year. Second, we only use students with prior test scores in our regressions. These students are slightly more likely to pass their exams.

To generate estimates of teachers' effectiveness in raising student achievement, we estimate the following regression with student-level data:

$$(1)\quad A_{it} = \beta_g X_{it} + \gamma_g \overline{X}^c_{it} + \zeta_g \overline{X}^s_{it} + \delta W_{it} + \pi_{gt} + \varepsilon_{it}$$

where A_{it} represents the math or reading test score of student *i* in year *t*, X_{it} represents student characteristics, $\overline{X}^c_{it}$ and $\overline{X}^s_{it}$ are the mean characteristics of the students in student *i*'s classroom and school respectively in year *t*, W_{it} represents characteristics of the teacher to which the student is assigned in year *t*, and π_{gt} is a fixed effect for the grade in which student *i* is enrolled and the year *t* in which we observe him/her. Student characteristics include a cubic polynomial in prior-year math scores, a cubic polynomial in prior-year reading score, gender, six categories for race and ethnicity, an indicator for free/reduced price lunch status, an indicator for special education status, an indicator for English Language Learners, and the number of absences and suspensions for the student in the previous school year.[17] The coefficients on individual, class, and school characteristics (β_g , γ_g, and ζ_g) are allowed to vary by the grade level of the test being taken. The class-level variables also include class size, and school-level variables include average class size in the school. All test scores are normalized within grade and year to have a mean of zero and a standard deviation of one. A standard deviation of student test scores in New York City is comparable to that of students nationally, according to data from the National Assessment of

[17] No reading test scores are available for seventh grade students in 2002-2003, so we do not examine seventh grade reading achievement in that year and eighth grade reading achievement the following year. We do examine eighth grade math achievement in both years. To correct for the fact that these students are missing prior reading scores in 2003-2004, we set their prior reading scores to zero and include an interaction between prior math scores and an indicator variable for this group. We do not add a main effect for this group because grade by year fixed effects are already included in the regression.

Educational Progress (NAEP).[18]

When studying impacts on math (reading) scores, we focus on the teacher identified as teaching the student's math (reading) course in that year. In grades six through eight, these often turn out to be different teachers. Students in grades four and five (and grade six students in elementary schools) typically had the same teacher in both math and reading.

To check the robustness of our results, we also run regressions that remove school average characteristics and replace them with either school fixed effects or fixed effects for permutations of school, grade and year. These fixed effects capture potentially important variation at the school level in factors that we cannot observe (e.g., the effectiveness of the school administration). When school or school-grade-year fixed effects are included, the coefficients on teacher characteristics are identified only from variation among teachers working within the same school or the same school, grade level, and year. Because uncertified and AC teachers are not spread evenly over schools, grades, and years, this significantly reduces the sample providing identification. We therefore prefer the specification given by equation (1), though our results are substantially the same across specifications.

4.1 Baseline Results on Teacher Certification and Teacher Effectiveness

In Table 6, we report differences in student performance in math and reading between groups of teachers with different types of certification. In all specifications, we make these comparisons while adjusting for individual years of teaching experience, and all standard errors are calculated allowing for clustering by school-grade-year cells. For illustrative purposes, in

[18] The standard deviation (SD) of NAEP scores in New York City and the nation as a whole were identical for 4th grade math (SD=28) and 8th grade reading (SD=35). The SD in New York City was slightly higher for 8th grade (cont'd on next page)

columns (1) and (6) we show the results of regressions of test scores on teacher characteristics, with no controls for student, classroom, or school characteristics. There are very large differences in test scores for students assigned to different types of teachers. The math scores of students assigned to teaching fellows and TFA corps members were .20 and .28 standard deviations below those of students assigned to regularly certified teachers. Students assigned to international recruits and uncertified teachers scored .48 and .13 standard deviations, respectively, below students in classrooms taught by regularly certified teachers. These differences were similar for reading test scores.

However, much of the apparent difference in student performance for those assigned to different groups of teachers is simply due to differences in students' prior year test performance and demographics. Uncertified teachers and those participating in an alternative certification program tend to be concentrated in schools with low performing students. When we include test scores from the prior year and other student level covariates in the regressions (columns (2) and (7)), the coefficients change dramatically. In math, students assigned to teaching fellows performed no differently than similar students assigned to traditionally certified teachers, while those assigned to TFA corps members outperformed by .01 standard deviations. Students assigned to international recruits and uncertified teachers continue to underperform relative to those assigned to certified teachers, but the differences are much smaller (-.05 and -.005 standard deviations, respectively). In reading, when controlling for baseline characteristics of students, students assigned to teaching fellows, TFA corps members, international recruits, and uncertified teachers underperformed those assigned to regularly certified teachers, but these differences are a

math (SD=38 versus SD=36), and slightly lower for 4th grade reading (SD=33 versus SD=36). These estimates were calculated using the *NAEP Data Explorer* at http://nces.ed.gov/nationsreportcard/nde/.

small fraction of those reported in columns (1) and (6).

In columns (3) and (8), we include classroom level and school level averages of student characteristics as control variables. The addition of these group level control variables has little impact on the coefficient estimates for teaching fellows and uncertified teachers. Relative to the students of regularly certified teachers, Teach for America corps members' students are now estimated to score .02 standard deviations higher, and international recruits' students score .03 standard deviations lower. In reading, students assigned to teaching fellows scored .01 standard deviations below those assigned to regularly certified teachers. This is the only between-group difference that is statistically significant for reading test scores.

In columns (4) and (9), we include school fixed effects—thereby implicitly holding constant all fixed school characteristics including test scores. The sole source of identification for differences in teacher effectiveness is the variation in performance of students assigned to different types of teachers within the same schools. Therefore, teachers in schools hiring only traditionally certified teachers or only AC teachers play no role in identifying the differences between these groups of teachers. Moreover, to the extent that principals are selecting for specific traits when they hire, the differences between groups within schools may mask important differences in effectiveness of teachers between schools. This is a potential downside to the fixed effects specification. However, we find results that are very similar to those shown in columns (2) and (7). Differences between groups of teachers are small and usually statistically insignificant. In math, there is a positive and statistically significant coefficient for assignment to a TFA corps member (.03) and a negative significant coefficient for assignment to an international teacher (-.03). In reading, there is a negative significant coefficient for assignment to a teaching fellow (-.02).

In columns (5) and (10), we add fixed effects for each school by grade by year permutation. While the results in columns (4) and (9) allowed for comparisons across grades and school years, as long as teachers were working in the same schools, the results in column (5) and (10) focus only on those teachers working in the same school and grade during the same school year. This change in specification has little or no effect on the results. Again, in math, there is a statistically significant positive effect for TFA corps members (.02) and a statistically significant negative effect for international teachers (-.02). In reading, there is a statistically significant negative effect for teaching fellows (-.01). Thus, controlling for school average student characteristics, allowing for comparisons across grades and across years within the same school, or allowing for comparisons only within school, grade, and year does not substantially change our findings.

It is important to note that the estimates we present—even those with school fixed effects—can be identified with comparisons other than direct comparisons between teachers. To see this, suppose that teaching fellows and certified teachers never worked in the same school, but that both sets of teachers worked with members of another group of teachers. Our empirical framework will still estimate a difference in value-added between teaching fellows and certified teachers through comparisons with this third group. In other regressions (not shown), we restrict our sample to (1) certified teachers and teaching fellows working in the same school and grade and (2) uncertified teachers and teaching fellows working in the same school and grade. These “head to head” comparisons produce very similar results to those in Table 6. .

4.2 Additional Estimates of the Relative Effectiveness of Groups of Teachers

In Table 7, we present the results of several additional specifications that examine more closely the relative effectiveness of different groups of teachers. As in columns (3) and (8) of Table 6, these regressions included controls for teacher experience, the full set of student, classroom level, and school level baseline characteristics, and fixed effects by grade and year. The first column of Table 7 redisplays our baseline estimates from Table 6; results for math (reading) are shown in the top (bottom) panel. In math, TFA corps members' students score .02 standard deviations higher, and international recruits' students score .03 standard deviations lower. In reading, students assigned to teaching fellows score .01 standard deviations below those assigned to regularly certified teachers. All other between-group differences are not statistically significant.

The remaining columns in Table 7 provide additional estimates based on various sub-samples. First we separately examine elementary and middle school grades.[19] Then, we separately examine schools above and below the median average test score. In each of these specifications, teachers at all levels of experience are included, and dummy variables are used to control for years of teaching experience. Last, we separately examine teachers with zero, one, and two years of experience. Here the identification of between-group differences comes only from direct comparisons between teachers with the same amount of experience.

When comparing elementary and middle grades or schools with above or below-median test scores, we find few changes in our estimates of the impact of certification status. The only notable exceptions are for international teachers: In math, their students underperform those assigned to traditionally certified teachers by -.03 standard deviations in middle school grades. In reading, their students outperform those assigned to traditionally certified teachers by .04

[19] Elementary grades are 4 and 5, middle grades are 7 and 8. 6th graders are considered elementary if they attend a school where the maximum grade is 6, otherwise they are considered middle.

standard deviations in schools with above median average test scores. When focusing on teachers with zero, one, or two years of experience, we find a pattern of coefficients that indicates teaching fellows, TFA corps members, and uncertified teachers may fare worse as rookie teachers than after they have gained a year or two of experience. For example, estimated achievement impacts for teaching fellows with zero, one and two years of experience are -.009, .005 and .018, respectively, for math and -.018, -.007, and -.003, respectively, for reading. One possibility for this is that AC teachers and uncertified teachers may have higher returns to experience than other teachers on average. Another possibility is that there is negative selective attrition of these teachers, relative to traditionally certified teachers. In section 4.5 we focus directly on the return to experience and explore these potential sources of dynamic heterogeneity across groups of teachers.

4.3 Selection of Teaching Fellows and Teacher Effectiveness

Prior research has found a relationship between teacher effectiveness and the selectivity of the college a teacher attended (Summers and Wolfe (1977)), tests of teachers' verbal ability (Hanushek (1971)), or a teacher's own ACT (American College Testing program) scores when applying to college (Ferguson and Ladd (1996)). However, as Hanushek and Rivkin (2004) argue in summarizing the research on teacher impacts on student achievement, the association between teacher test scores and student outcomes is relatively weak. Moreover, the literature on teacher effectiveness has consistently failed to find that those holding master's degrees are more effective, despite the fact that most teacher pay scales reward higher educational attainment (Murnane (1975), Summers and Wolfe (1977), Ehrenberg and Brewer (1994), Aaronson et al. (2003), Clotfelter et al. (2006)).

Academic credentials are strongly related to teaching fellow applicants' probabilities of selection into the program (Figure 2), and, on average, teaching fellows attended more selective colleges than traditionally certified teachers (Table 1). Yet we find that teaching fellows are, on average, no more effective in the classroom than traditionally certified teachers. These facts contrast with the evidence from other studies that measures of academic skills are positively related to teachers' effectiveness in raising student achievement. Moreover, whether academic credentials are related to teaching effectiveness is an important issue for recruitment policies. If attending a selective college is not a good indicator of teacher effectiveness, alternative certification programs may do well to focus on other characteristics.

We investigate whether value-added is associated with better academic credentials among teaching fellows. We focus on teaching fellows because we have high quality data from their applications on their undergraduate institution and their own undergraduate GPA. Although we do possess data on undergraduate institution for many teachers who are not in the teaching fellows program, this information is missing for a considerable fraction.

If academic skill is a significant predictor of teacher effectiveness, we would expect higher value-added from teaching fellows with better academic credentials. Columns (1) and (4) of Table 8 show the relationship between math and reading value-added and the median math SAT score of teaching fellows' undergraduate institution, measured in deciles. There is a coefficient of .003 standard deviations in math and a coefficient of less than .0005 in reading, neither of which are statistically significant at conventional levels.[20] Thus, there is little evidence that academic credentials are a strong predictor of teacher effectiveness. The point estimates still indicate only small differences in value-added, even among students who went to vastly different undergraduate institutions. For example, the difference in math value-added between teachers whose colleges were three deciles away in the median math SAT distribution (e.g., CUNY Baruch College vs. New York University) is about .009 standard deviations. This is less than 25 percent of the gains to experience in the first year of teaching (see column (2) of Table 10, presented below).

Columns 2 and 5 of Table 8 show the relationship between value-added and teaching fellows' undergraduate GPA. We do not find any statistically significant relationship for either reading or math, and the point estimates are actually negative. Columns 3 and 6 add an interaction between GPA and median math SAT score, to account for the fact that higher GPA may be a greater signal of academic ability at more selective institutions. We do not find a statistically significant interaction between GPA and median math SAT for either reading or math, and the point estimates are negative. In other words, neither certification status nor easily observable academic traits such as college selectivity and undergraduate GPA seem to be associated with teacher effectiveness.

4.4 Validity of Within-School Comparisons

Our estimates above are based on comparisons of teachers with different initial certification status within schools or among schools with similar observable characteristics. However, the certified teachers working in schools hiring large numbers of alternatively certified teachers may be systematically different from certified teachers hired in other schools. In fact,

[20] These regressions control for year-grade fixed effects, student characteristics and classroom- and school-level average student characteristics (as in Table 6 column (3)). Regressions using reading SAT decile show no (cont'd on next page)

given that the teaching fellows are placed in the most hard-to-staff schools, one might fear that the comparison teachers in those schools may be less effective than certified teachers working in other schools. Teachers could be sorting across schools based on observed or unobserved traits. In this section, we test whether the difference in observed traits between teaching fellows and other teachers depends upon whether one is making comparisons within or between schools.

We investigate sorting on observables by estimating regressions of observable teacher characteristics on indicator variables for whether that teacher belongs to a particular group. This specification is shown in equation 2.

$$(2) \qquad T_i^{st} = \alpha_t + \delta W_i + \mu_i , \quad \mu_i = \pi_s + \varepsilon_i$$

Variation in characteristics (*T*) across teachers can be explained partially by the year (*t*) in which the teachers were hired, teachers' certification and program participation at the time of hire (*W*), and an error term (μ) that can be decomposed into a school effect (π_s) and an i.i.d. disturbance (ε). The coefficients on certification and program participation (δ) estimate the average difference in observable characteristics across groups of teachers. If teachers are sorted across schools based on observables, estimates of between-group differences will be biased if school effects are omitted from the regression.

The characteristics we use as dependent variables include three measures of academic skill (whether the teacher has a master's degree and the percentile of their undergraduate median math and verbal SAT scores) and two demographic measures (age and whether the teacher is Black or Hispanic).[21] Estimates of between-group differences in teacher characteristics (δ), with

significant predictive power for either math or reading value-added, so we do not present results for them here.

[21] We include all teachers hired on or after the school year 1999-2000. Time varying characteristics (including the school in which the teacher works) are measured during the teacher's first year. Year fixed effects are always included. Undergraduate information is known for about 30 percent of the sample.

and without controlling for school fixed effects, are shown in Table 9. Coefficient estimates from regressions without school fixed effects display the same facts shown earlier in Table 1. Relative to traditionally certified teachers: (1) teaching fellows, TFA corps members, and uncertified teachers are less likely to have graduate degrees, (2) teaching fellows and, to a greater extent, TFA corps members attended more selective colleges, (3) TFA corps members are younger, and (4) teaching fellows, international teachers, and uncertified teachers are more likely to be Black or Hispanic.

However, the inclusion of school fixed effects in these regressions produces little or no change to the between-group differences in measures of academic credentials. For example, teaching fellows are estimated to have attended colleges with a median math SAT score 8.2 percentile points (standard error .37) above traditionally certified teachers when school fixed effects are omitted. When school fixed effects are included, the estimate difference is 8.6 percentile points (standard error .44).

One instance where teachers' sorting across schools does appear to matter is sorting on ethnicity. While teaching fellows are, overall, far more likely to be ethnic minorities than traditionally certified teachers, they are only slightly more likely to be so when examining differences within schools. Similarly, while TFA corps members are equally likely to be ethnic minorities than traditionally certified teachers, they are less likely to be so when examining differences within schools. We check further that the ethnicity based sorting we find does not lead us to misinterpret the regressions of measures of academic skills. This could happen because ethnic minorities on average have lower measures of academic skill. For example, in a regression of median math SAT percentile that includes school fixed effects, the lack of change in the teaching fellows coefficient might reflect a combined effect of academic skills sorting (pushing the coefficient towards zero) and ethnic sorting (pushing the coefficient away from zero). If this were the case, one should be able to see academic skills sorting when ethnicity is added as a control variable. This turns out not to be true. Adding a control for being an ethnic minority barely changes the coefficient on teaching fellow; it moves from 8.6 percentile points (standard error .44) to 8.45 (standard error .42).

Unfortunately, we cannot test for sorting on unobservables. However, if the certified teachers working in the type of schools hiring large numbers of teaching fellows were worse on unobserved traits, one might have expected to see evidence of sorting on observables as well.

4.5 The Returns to Teaching Experience

Throughout the analysis above, we were controlling for teaching experience in order to make comparisons across different groups of teachers. In Table 10, we report results on the returns to teaching experience in regressions that control for student, classroom average, and school average baseline characteristics, as well as grade-year fixed effects. There are significant returns to experience, most of which occur in the first two years of teaching. Teachers in their second year of teaching have value-added .033 standard deviations higher in math and .023 standard deviations higher in reading than teachers in their first year. Teachers in their third year of teaching have value-added .052 standard deviations higher in math and .034 standard deviations higher in reading than teachers in their first year. However, the estimated returns to additional years of experience are quite small.

The above estimates of returns to experience are identified using two sources of variation: comparisons between teachers with varying amounts of experience and comparisons within teachers (observing changes in outcomes for a given teacher over time). The former source of variation may lead to selection bias in estimating the returns to experience for the average teacher. For example, if less effective teachers are more likely to remain teaching, we will understate the returns to experience; if more effective teachers are more likely to remain, we will overstate the returns to experience. Estimates of the return to experience that are identified using variation within teachers over time are not susceptible to these sources of bias. In order to isolate this variation, we include teacher fixed effects in our regression (column (2) and (8) of Table 10). The inclusion of teacher fixed effects does not cause our estimates of the return experience to change in reading, but it substantially increases the returns to experience in math. When estimated using variation within teachers over time, the returns to experience are significantly higher in math than in reading. On average, teachers in their 2nd through 4th year of teaching achieve gains .019, .031 and .035 standard deviations higher in reading than when they were first-year teachers and .037, .059 and .073 standard deviations higher in math.

The rise in the estimated return to experience in math when teacher fixed effects are added to the regression suggests that variation in experience among math teachers is negatively correlated with variation in other characteristics that raise teacher effectiveness. We consider several potential sources of bias. First, recently hired cohorts of math teachers may be more effective than older cohorts. Second, more effective math teachers may have a higher attrition rate. Third, identification with teacher fixed effects comes almost purely from experience gained while teaching math to elementary or middle school students in New York, and the returns to such experience (for individuals teaching math to elementary or middle school students in New York) may be higher than the returns to experience gained elsewhere. We investigate these potential sources of bias by estimating regressions that omit teacher fixed effects but include additional controls that account for the biases mentioned above. If the source of bias is one of those mentioned, we should find estimated returns to experience are more in line with those from the regression with teacher fixed effects.[22]

First, we control for differences across cohorts by including indicator variables for the year teachers were hired. For simplicity, we include a single variable for teachers hired at or before the school year 1992-1993. This change in specification does not move the estimated returns to teaching experience closer to the teacher fixed effects results (Table 10 column (3)); if anything, estimated returns are lower when controlling for cohort. Indeed, the pattern of coefficients for the cohort dummy variables (shown in appendix Table A1) indicates that older cohorts have slightly *higher* value-added.

Next, we include dummy variables identifying teachers who left employment in the district some time before the fall of 2005. Because selective attrition may vary over teachers' careers, we include separate dummy variables for teachers who left the district at each level of teaching experience. Again, this does not move the estimated returns to teaching experience closer to the teacher fixed effects results; if anything, estimated returns are lower when controlling for selective attrition. The pattern of coefficients for the attrition dummy variables (shown in Table A1) indicates that teachers who leave with zero experience have lower value-added (-.03 standard deviations) than teachers with zero experience who stay. Attrition among more experienced teachers is not significantly related to teacher effectiveness; the point estimates are generally positive but statistically distinguishable from zero.

Next, we control for each teacher's experience level in the first year for which we observe

[22] In theory, these results could also be driven by specification error. Teacher fixed effects may be correlated with omitted variables at the school level. Hence, it is possible that the change in the estimated return to experience is driven by the fact that school fixed effects are omitted from the first regression. However, if we include school fixed effects (or even school-grade-year fixed effects) and then replace them with school by teacher fixed effects we find the same increase in the estimated return to experience.

them in our data. Estimated returns to experience in this specification will not be identified from variation in experience gained prior to observation in our sample and will more closely match the identifying variation used in the teacher fixed effects specification.[23] Including controls for initial observed experience has a dramatic effect on estimated returns to experience. Estimated returns in this specification are *larger* than in the teacher fixed effects regression. Moreover, the pattern of coefficients on initial observed experience indicates that returns to experience are significantly smaller if experience was gained prior to observation in our sample. Earlier experience may come from teaching either inside or outside of New York, and inside or outside of the grades and subjects we observe in our data. The fact that the return to earlier experience is lower than experience gained in our sample suggests that experience gained in other districts or in other grades and subjects is less useful than experience gained teaching math to elementary and middle school students in New York.

The issue of heterogeneous returns to different forms of experience has received little attention in the value added literature, but has important policy implications. For example, if a large part of the return to experience comes from learning district specific skills, then school administrators should be far more concerned with retaining their current staff than recruiting outsiders with experience from other districts. For reasons of space, we do not pursue the issue further here, but plan to do so in future work.

In Table 7, we presented some evidence that teaching fellows, TFA corps members, and uncertified teachers improve, relative to other teachers, over their first few three years teaching in New York. This suggests that the return to experience may be higher for these teachers than for

[23] Controlling for initial observed experience is different than controlling for cohort. If we had one year of data, cohort fixed effects would capture most of the variation in observed experience. In a multi-year sample, the (cont'd on next page)

traditionally certified teachers.   There are several reasons why one might think these groups should have higher returns to experience than traditionally certified teachers. These individuals generally do not have student teaching experience, unlike traditionally certified teachers who usually assist and teach in a public school under the supervision of a more experienced teacher for one or two years while studying. Also, AC teachers are usually studying during their first two years of teaching, and that might also lead to greater improvements when they complete their coursework. However, it is equally plausible that returns to experience are higher for traditionally certified teachers. Traditionally certified teachers, because of their training, may be better equipped to learn and improve their teaching skills. Furthermore, the coursework AC teachers are required to take at night and on weekends in their first two years of teaching may well be a hindrance that stops them from learning as much as possible at work. (Alternatively, AC teachers could have higher returns to experience; when coursework stops, the AC teachers could see a jump in performance.)

We investigate differential returns to experience across teacher certification groups by including interactions between group dummy variables (e.g., whether a teacher is a teaching fellow, TFA corps member, etc.) and experience level dummy variables. We also included teacher fixed effects in these regressions to avoid the sources of selection bias discussed above. We also restrict our sample to teachers with three years of experience or less, since AC teachers are almost all in this population.

In Table 11, we report the coefficients from these regressions and the p-value on the test of the joint significance of the interactions between experience levels and being a teaching fellow, being a TFA corps member, and being an uncertified teacher. (We exclude international

predictive power of cohort fixed effects for initial observed experience is greatly reduced.

teachers from this analysis since most arrive with several years of prior teaching experience.) For math test scores, we cannot reject the hypothesis that the returns to experience are the same for traditionally certified teachers, teaching fellows, and TFA corps members. However, we can reject equal returns to experience between uncertified teachers and regularly certified teachers at the 3 percent confidence level. For reading test scores, we cannot reject equal returns to experience for traditionally certified teachers, uncertified teachers, and TFA corps members. However, we could reject the hypothesis that teaching fellows have the same return to experience as traditionally certified teachers (p-value=7 percent).

For ease of exposition, in Figure 3 we plot point estimates and 95 percent confidence intervals for the estimated returns to experience from Table 11. The top left panel of Figure 3 shows the estimated returns for traditionally certified teachers, and each subsequent panel compares these estimates with estimates from one of the other three groups.[24] For both math and reading test scores, the point estimates for teaching fellows are noticeably higher than those of traditionally certified teachers. In particular, it appears that teaching fellows gain appreciably higher returns in their first two years of teaching. For example, for teachers who start with the same value-added, teaching fellows are expected to have .03 standard deviations higher value-added than traditionally certified teachers in both reading and math test scores in their third year of teaching. For uncertified math teachers, initial returns to experience are also appreciably higher. Given the same initial value-added, uncertified teachers are expected to have .03 standard deviations higher value-added than traditionally certified teachers in both reading and math test scores in their third year of teaching.

[24] We do not plot estimates for TFA corps members after 2 years of experience because of extremely large standard errors.

4.5 Attrition and Steady State Differences in Experience

Holding constant a teacher's baseline effectiveness, the positive payoff to teaching experience implies that there is a cost to hiring a teacher with a higher probability of turnover. Therefore, an analysis of between-group differences in value-added is not complete without considering differences in attrition. Suppose that there were two groups of teachers with identical impacts on student achievement after controlling for experience, but with different retention rates. Given significant returns to experience, a school district would be better off hiring the group of teachers with higher retention rates, since this would lead to fewer novice teachers in the steady state. Lower turnover would also reduce hiring and training costs, though we do not consider this effect in our analysis here.

Suppose that $e_{j1}, e_{j2},...e_{jT}$ represents the proportion of teachers from a given group j that would be in their 1st through Tth year of teaching in steady state, given group-specific retention rates. And suppose δ_j represents the mean value-added of group j teachers in their first year of teaching and that $r_{j2},...r_{jT}$ represent the average returns to experience for the same group of teachers in their 2nd through Tth year of teaching. In steady state, the average value-added of the jth group of teachers in steady state would be equal to $\delta_j + \sum_{t=2}^{T} e_{jt} r_{jt}$.

To study the differences in retention for different groups of teachers, we estimated logistic regressions of hazard rates for certified, uncertified, and AC teachers hired since the school year 1999-2000 with 0 to 7 years of experience, while controlling for age (in 5 year categories) and year dummies.[25] Unlike our analysis of teacher effectiveness, we use the full sample of teachers in the payroll files, and do not limit the analysis to those teaching reading and math in grades three through eight. However, this has little impact on our results. We also do not include international recruits; although they have high turnover, they are generally hired with several years of experience. We included interactions between years of experience and initial certification status at time of hiring (teaching fellow, TFA corps member and uncertified teachers). For uncertified teachers, we also included an interaction with the year dummy in 2003. As mentioned above, starting in the fall of the school year 2003-2004, school districts in New York State were no longer permitted to employ uncertified teachers. As a result, many of them left after the school year 2002-2003.

Figure 4 reports the cumulative retention rates adjusting for age and year for the different groups of teachers. Teaching fellows have very similar retention rates to regular certified teachers (with teaching fellows having slightly higher retention rates in the first two years). By their fifth year in teaching (with four years of experience), approximately 50 percent of both groups are still with the district. Uncertified teachers have somewhat lower retention rates, with 45 percent remaining with the district in their fifth year.

In contrast, Teach for America corps members have much lower cumulative retention rates. By the fifth year, only about 18 percent of corps members remain with the district. Presumably, this reflects the fact that TFA corps members sign up for a two-year teaching commitment.

Assuming a constant hazard rate between the 5th year and the 30th year of experience,

[25] We found no substantive impact on the results using logistic regression versus OLS, nor adding school fixed effects.

Figure 5 reports the steady state proportions of all four groups by year of experience implied by the above analysis. The only large difference is between TFA corps members and the other three groups. In steady state, TFA corps members would be roughly twice as likely to be in their first year of teaching ($e_{1TFA} = .256$) than certified teachers ($e_{1Cert} = .111$).

The above framework provides a means for resolving a longstanding debate about the TFA program. Many supporters of the TFA program argue that corps members have larger impacts on student achievement than regularly certified teachers, particularly among those certified teachers willing to work in low-income schools (that is, they argue that the corps members have a high δ_{TFA}). Critics of the program argue that because of high turnover rates, school districts are constantly having to replace TFA corps members with novice teachers (i.e., they argue that the program has low $e_{j2},...e_{jT}$). To the extent that novice teachers underperform experienced teachers, there would be some cost to hiring high-turnover teachers. But it is a question of magnitude. Both groups could be correct in their assertions, but the net result depends upon the magnitude of the differences in retention rates and experience-adjusted impacts on student performance. Assuming that returns to experience are similar between TFA corps members and certified teachers (which is consistent with our findings), we can calculate the difference between δ_{TFA} and δ_{Cert} that would be required to ensure that TFA corps members had larger steady state impacts than certified teachers:

$$\delta_{TFA} + \sum_{t=2}^{T} e_{TFAt} r_t > \delta_{Cert} + \sum_{t=2}^{T} e_{Certt} r_t$$

$$\Rightarrow (\delta_{TFA} - \delta_{Cert}) > \sum_{t=2}^{T} (e_{Certt} - e_{TFAt}) r_t$$

We do this using the point estimates from Table 10, columns (2) and (8), which are not affected by common selection biases such as correlation between attrition and teacher effectiveness. We estimate that $\delta_{TFA} - \delta_{Cert}$ would have to be greater than .019 in math and .012 in reading in order for TFA corps members to have greater steady state value-added.[26] This is quite a modest difference. In other words, despite large differences in retention rates, the differences in returns to experience and retention rates are not enough to generate large differences in steady-state impacts between TFA and other groups of teachers.[27] The results of our analysis suggest that even the small positive difference in value-added for TFA corps members' teaching math is enough to compensate for their higher turnover.

5. Variation in Value-Added Within Groups of Teachers

The evidence presented above suggests that there are little or no differences in average value-added between groups of teachers with different types of certification. It is important to distinguish this lack of variation from variation in value added among teachers within each of these groups. Variation in teacher effectiveness within groups provides a measure of the potential benefits of policies that enable districts to selectively retain only the highest performing teachers.

In order to measure variation in value-added among teachers, we first estimate each teacher's value-added in each class and year controlling for student, classroom, and school characteristics, as well as grade by year fixed effects. Specifically, we first take estimates of equation (1) and calculate classroom average residuals (ε_{jct})

[26] If we limited our analysis only to teachers included in our regression analysis, we estimate that $\delta_{TFA} - \delta_{Cert}$ would have to be greater than .014 in math and .008 in reading in order for TFA corps members to have greater steady state value-added. These differences are approximately two-thirds of those for the full sample of teachers.

[27] Ballou (1996) makes this point more generally using a simulation. Large differences in turnover and modest experience effects do not generate large costs to hiring high turnover teachers in terms of value added.

$$(1)\ A_{it} = \beta_g X_{it} + \gamma_g \overline{X}_{it}^{c} + \zeta_g \overline{X}_{it}^{s} + \delta W_{it} + \pi_{gt} + \varepsilon_{it}\ ,\ \varepsilon_{jct} = \tfrac{1}{N_{jct}} \sum_{i,t \in j,c,t} \varepsilon_{it}$$

ε_{jct} is the average student level residual for classroom *c*, taught by teacher *j*, in year *t*. Our sample restrictions are the same as in the analysis in Section 4. The residual (ε_{jct}) represents each teacher's individual contribution to value-added in that class—a teacher's performance residual in class c and year t , after controlling for the teacher's observable characteristics and the observable characteristics of his/her students, classroom, and school.

There is considerable variation in this teacher performance residual in any given year. As can be seen in Table 12, the standard deviation of ε_{jct} is 0.21 student standard deviations for elementary school math, and 0.20 for elementary school reading, with only slightly less variation in middle school. In other words, even within grade-year cells and after controlling for observable differences among classrooms, the standard deviation in our estimates of teacher value-added is about one-fifth of a student-level standard deviation. The finding of large variation in value-added is common to other studies of teachers, e.g., Aaronson et al. (2003), Rockoff (2004), Rivkin et al. (2005), and Hanushek et al. (2005).

Of course, not all of this variation is the result of persistent differences in ability across teachers: some of the variation may reflect random sampling error in estimating the teacher fixed effect, while some may reflect other non-persistent factors such as a particularly disruptive student or a dog barking on the day of the test.

A simple method of determining the proportion of the variation that is persistent from classroom to classroom is to estimate correlations in the teacher performance residuals when teaching in different years. Suppose we decompose ε_{jct} into two components: a persistent component (μ_{jct}) that represents teacher effectiveness and a non-persistent component (ξ_{jct}) that represents sampling variation, unobserved classroom characteristics, and other idiosyncratic shocks to classroom performance. The non-persistent component is independent from class to class. If the persistent component is fixed across years (i.e., unchanging teacher effectiveness with $\mu_{jct} = \mu_j$), then it is straightforward to show that the covariance in ε_{jct} between any two classrooms is equal to the variance of the persistent component.[28]

In Table 12, we also report estimates of the standard deviation in the signal variance based on the covariance in the teacher performance residual (ε_{jct}) between classrooms. The estimates suggest that there is a large persistent component in the teacher residuals, particularly in elementary schools. In elementary schools, the standard deviation of the persistent component of teacher performance (μ_j) is estimated to be 0.13 in math and 0.10 in reading. There is slightly less variation (0.12 and 0.08) in teacher performance among novice teachers —those in their first 3 years of teaching —with similar amounts of variation found among certified, uncertified, and teaching fellows. In results not reported in the table, we found little variation across cohorts of teachers in the estimated signal variance, while within cohorts of teachers the estimated signal variance was always lower for novice teachers. Thus, less variation in performance among new teachers reflects an experience effect (as teachers gain experience, the variation in performance grows) rather than a cohort effect (recent cohorts are not more uniform in performance).

In middle school, we see considerably less variation, with the standard deviation of the

[28] The assumption that teacher effects are the same across classrooms and time may be too strong. If the persistent component follows a time series process such as an AR(1) or a martingale (e.g., teacher ability evolves over time with $\mu_{jt} = \rho_{jt-1} + v_{jt}$), then the correlation in ε between any two years will be strongest in adjacent years. Similarly, if a teacher's skills are specific to particular subject matter, then the correlation in ε between any two classes will be stronger when the subject matter is more closely related (e.g. 2 sections of the same course). We did not see any consistent evidence that the correlations were stronger in adjacent years, but there was evidence from middle school that correlations were stronger between sections of the same course. We discuss this issue in more detail when we present our results.

persistent component of performance being about two-thirds as large. Part of the explanation for this result is that performance is less persistent when a middle school teacher is teaching different courses (e.g., advanced versus remedial math): when we focus on teachers teaching the same course in multiple classrooms (the bottom panel of Table 12), the persistent component grows. To the extent that teachers in middle school often change the courses they teach from year to year, this will make it more difficult to reliably identify performance differences for middle school teachers.

For both math and reading in elementary and middle school, the estimates in Table 12 imply that there are large and persistent performance differences across teachers. If we were able to rank teachers by their value added, the difference in average value added between the top 25 percent and the bottom 25 percent of teachers would be approximately 2.5 times the standard deviation across teachers (assuming that teacher performance is normally distributed). Thus, the estimates in Table 12 imply that the average value added among the top quarter of elementary school math teachers is 0.33 standard deviations greater than the value added among the bottom 25 percent of teachers. For middle school teachers, this difference is somewhat smaller, but there is still at least a 0.20 standard deviation difference in value added between the top and bottom 25 percent of teachers. In other words, the impact of assigning a student to a bottom quarter teacher rather than a top quarter teacher is roughly three times the impact of being assigned to a novice teacher rather than an experienced teacher, and more than ten times the impact of being assigned to a teacher with a particular kind of certification or from a particular program!

Figure 6 shows the variation in teacher effectiveness (value-added) within and among four groups of teachers—certified, uncertified, Teaching Fellows, and Teach for America. This figure plots kernel density estimates of the distribution of the persistent component (μ_j) of teacher effectiveness (value-added). For each teacher, we estimated their persistent component using an empirical Bayes (shrinkage) estimator of the form (ignoring the *c* subscript, and assuming that every classroom has the same number of students for simplicity):

$$E(\mu_j \mid \varepsilon_{j,1},\ldots,\varepsilon_{j,t}) = \bar{\varepsilon}\left(\frac{t}{t+\sigma_{\xi}^{2}\big/\sigma_{\mu}^{2}}\right),\ where\ \bar{\varepsilon} = \tfrac{1}{t}\sum_{s=1}^{t}\varepsilon_{j,s}$$

In other words, the estimate of each teacher's performance residual is simply their average teacher performance residual multiplied by a scaling factor. The scaling factor lies between zero and one, and depends positively on the number of years the teacher has been observed, and depends negatively on the ratio of noise (non-persistent) variance to signal (persistent) variance. Thus, these estimators account for noise in the teacher estimates (which overstates the dispersion across teachers) by shrinking noisier estimates back toward zero. The estimates of noise and signal variance come from Table 12. Finally, we add the mean difference between groups (from Table 6, columns 3 and 8) to the each teacher's empirical Bayes estimate of their performance residual.

While the differences *between* the four groups observed in Figure 6 are small, the differences *within* the four groups are quite dramatic. In other words, there is not much difference between certified, uncertified, and alternatively certified teachers overall, but effectiveness varies substantially among each group of teachers. To put it simply, teachers vary considerably in the extent to which they promote student learning, but whether a teacher is certified or not is largely irrelevant to predicting their effectiveness.

7. Conclusion

State and federal efforts to regulate teacher effectiveness focus almost entirely on *ex ante* qualifications of prospective teachers. For example, under the federal No Child Left Behind act, states and districts are required to hire certified teachers or those enrolled in an alternative certification program. However, our results suggest the emphasis on certification status may be misplaced. We find little difference in the average academic achievement impacts of certified, uncertified and alternatively certified teachers.

On average, the students assigned to teaching fellows performed similarly to students assigned to certified teachers in math, and slightly lower (-.01 standard deviations) in reading. This average difference belies somewhat larger gaps among novice teachers (-.02 standard deviations) and no differences between teaching fellows and certified teachers with multiple years of experience. This is because teaching fellows have somewhat higher average returns to experience than certified teachers in generating reading gains at the beginning of their careers.

We find evidence that Teach for America corps members have slightly higher value-added (.02 standard deviations) for math test scores than traditionally certified teachers, but we find no difference in reading. (Although the magnitudes of our estimates are smaller, Decker et al. find the same pattern of positive effects on math but not reading in their evaluation of TFA.) The other notable difference between TFA corps members and other teachers is that they have much higher turnover after two years, reflecting the two-year commitment that is part of the TFA program. We estimate that the steady-state difference in experience between TFA corps members and other teachers implies a small loss in value-added of about -.01 or -.02 standard deviations. Thus, the cost of higher TFA turnover (i.e., more of them will be novice teachers) in terms of value-added is quite small, and, at least in math, appears to be compensated by higher average teacher effectiveness.[29]

The fact that we find little or no differences in the average teacher effectiveness of certified, uncertified, and AC teachers does not imply that selection of teachers is unimportant. The standard deviation in value-added among teachers is roughly .10 student level standard deviations. Moreover, variation in effectiveness is roughly equivalent within each group of teachers (e.g., certified teachers, teaching fellows, etc.). To put this value in context, raising the effectiveness of novice teachers in New York by one standard deviation would have a similar impact on student achievement as the expected improvement of novices who spend 8 years teaching in the district! Thus, policies that enable districts to attract and retain high quality teachers (or screen-out less effective teachers) have potentially large benefits for student achievement.

Our work suggests that selecting high quality teachers at the time of hire may be difficult if districts rely on information related to teachers' academic background. We find that neither selectivity of undergraduate institution nor undergraduate GPA have predictive power for a teaching fellow's value-added, yet these characteristics are highly predictive of whether or not an applicant is selected to become a teaching fellow.

The large observable differences in teacher effectiveness *ex post* suggest that districts should use performance on the job, rather than initial certification status to improve average teacher effectiveness. Recent research on teachers certified by the National Board for Professional Teaching Standards (NBPTS) suggests that *ex post* measurement of teacher quality

[29] It is also worth noting that TFA corps members leave teaching for other positions in education. TFA alumni surveys indicate that more than 50 percent of its corps members are working in education even ten years after they enter the program, 18 alumni are currently working as principals or assistant principals in New York City's schools, and another ten alumni are working for the DOE in other district level jobs.

has large potential benefits.[30] Goldhaber and Anthony (2004), Cavalluzzo (2004) and Vandevoort, Amrein-Beardsley and Berliner (2004) all find that student achievement is higher in classrooms taught by NBPTS certified teachers. Admittedly, some states and districts, such as the Los Angeles Unified School District, offer salary bonuses to those with certification by the NBPTS. However, most of those bonus programs preceded the recent evidence that these teachers are more effective in promoting student achievement.

While traditional measures of teacher effectiveness, e.g. certification, have come up short, more research is needed on alternative, non-traditional *ex ante* measures. It is also important to investigate the potential use of value-added measures as an *ex post* signal of teacher effectiveness. Value-added measures have drawbacks, such as their limited scope and their potential malleability (Figlio 2005 and Figlio and Winicki 2002), but they also have appealing characteristics, such as objectivity, and the data necessary to construct them is already being collected by most school districts. We intend to pursue these issues in future work.

[30] The NBPTS, a non-profit organization, was created in 1987 to provide an objective means for recognizing and rewarding effective teaching. When applying for certification by NBPTS, teachers provide a videotape of their work in front of class, submit examples of written assignments and the feedback they provided to students, and answer a number of essay questions in a testing center.

References:

Aaronson, Daniel, Lisa Barrow and William Sander. (2003) "Teachers and Student Achievement in the Chicago Public Schools," Federal Reserve Bank of Chicago WP-2002-28.

Ballou, Dale (1996) "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics*, February 1996

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, James Wyckoff (2005a) "How Reduced Barriers to Entry into Teaching Changes the Teacher Workforce and Affects Student Achievement" Manuscript, July 2005.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, James Wyckoff (2005b) "How Reduced Barriers to Entry into Teaching Changes the Teacher Workforce and Affects Student Achievement," NBER Working Paper No. 11844, December 2005.

Cavalluzzo, Linda C. (2004) "Is National Board Certification an Effective Signal of Teacher Quality?" CNA Corporation Working Paper, November, 2004.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006) "Teacher-Student Matching and the Assessment of Teacher Effectiveness," NBER Working Paper No. 11936, January 2006.

Darling-Hammond, Linda, Deborah J. Holtzman, Su Jin Gatlin and Julian Vasquez Heilig. (2005) "Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness," Stanford University Working Paper, April 2005.

Decker, Paul T., Daniel P. Mayer and Steven Glazerman. (2004) "The Effects of Teach For America on Students: Findings from a National Evaluation," Mathematica Policy Research Report No. 8792-750, June 9, 2004.

Ehrenberg, Ronald and Dominic Brewer. (1994) "Do School and Teacher Characteristics Matter?: Evidence from High School and Beyond," *Economics of Education Review* Vol. 13, No. 1, pp. 1-17.

Ferguson, Ronald and Helen Ladd. (1996) "How and Why Money Matters: An Analysis of Alabama Schools," in Helen Ladd (ed.) *Holding Schools Accountable* (Washington, DC: Brookings Institution).

Feistritzer, Emily C. (2005) *Alternative Teacher Certification: A State-by-State Analysis: 2005*, (Washington, DC: National Center for Education Information, 2005).

Figlio, David N. (2005) "Testing, Crime and Punishment," NBER Working Paper 11194, March 2005.

Figlio, David N. and Joshua Winicki. (2002) "Food for Thought: The Effects of School Accountability Plans on School Nutrition" NBER Working Paper No. 9319, November, 2002.

Goldhaber, Dale and Emily Anthony. (2004) "Can Teacher Quality Be Effectively Assessed?" University of Washington and Urban Institute Working Paper, April 27, 2004.

Hanushek, Eric. (1971) "Teacher Characteristics and Gains in Student Achievement: Estimation using Micro Data," *American Economic Review* Vol. 61, No. 2, pp. 280-288.

Hanushek, Eric and Steven G. Rivkin. (2004) "How to Improve the Supply of High-Quality Teachers," in Diane Ravitch (Ed.) *Brookings Papers on Education Policy 2004* (Washington DC: Brookings Institution)

Hanushek, Eric, John F. Kain, Daniel O'Brien and Steven G. Rivkin. (2005) "The Market for Teacher Quality," NBER Working Paper 11154, February 2005.

Kane, Thomas J., Jonah E. Rockoff and Douglas O. Staiger. (2005) "Identifying Effective Teachers in New York City" Manuscript, July 2005.

Murnane, Richard. (1975) "The Impact of School Resources on the Learning of Inner City Children," (Cambridge, MA: Ballinger).

Raymond, Margaret, Stephen H. Fletcher and Javier Luque. (2001) "Teach For America: An Evaluation of Teacher Differences and Student Outcomes in Houston, Texas," (Stanford, CA: The Hoover Institution, Center for Research on Education Outcomes).

Rivkin, Steven G., Eric Hanushek, and John Kain. (2005) "Teachers, Schools and Academic Achievement," *Econometrica* Vol. 73, No. 2.

Rockoff, Jonah E. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, May Papers and Proceedings.

Summers, Anita and Barbara Wolfe. (1977) "Do Schools Make a Difference?" *American Economic Review*, Vol. 67, No. 4, pp. 639-652.

Vandevoort, Leslie G., Audrey Amrein-Beardsley and David Berliner. (2004) "National Board Certified Teachers and Their Students' Achievement," *Education Policy Analysis Archives* Vol. 12, No. 46.

Appendix: Additional Specification Checks

Analyses of teacher quality have generally used one of four different empirical specifications, which we will refer to as "quasi-gains," "gains," "levels with fixed effects," and "gains with fixed effects." Our specification falls under the "quasi-gains" category: the dependent variable in our regressions is the level of current student achievement, and prior achievement levels are used as control variables. In a "gains" specification, the dependent variable would be the change in student test scores from the current period to some base period, usually the end of the prior school year (see Rivkin et al. 2005). This is essentially a quasi-gains specification where past achievement is restricted to have a one-to-one relationship with current achievement. In a "levels with fixed effects" specification, the level of current student achievement is the dependent variable and student fixed effects—not prior achievement levels—are used as control variables (see Rockoff (2004)). A "gains with fixed effects" specification, as the name implies, combines both of these features.

The results shown in this paper do not vary when we use any of these three other approaches. In columns (1) and (5) of Table A2, we display the main results of our paper, replicated from columns (3) and (8) of table 6. In columns (2) and (6) we show results using a "gains" specification, in columns (3) and (7) we show results using a "levels with fixed effects" specification, and in columns (4) and (8) we show results using a "gains with fixed effects" specification. While the point estimates change slightly across specifications, the qualitative findings of our analysis are unchanged. Teaching Fellows perform similarly to certified teachers in math instruction but slightly worse in reading instruction, a difference of about .01 standard deviations. In math, TFA corps members perform slightly better (.02 standard deviations) and international teachers perform slightly worse (-.03 standard deviations). Overall, there is little evidence of substantial variation in value-added between groups of teachers.

Table 1: Characteristics of Teachers by Certification and Program

	Regular Certified	*Regular Uncertified*	*Teaching Fellow*	*Teach For America*	*Internat'l Teacher*
Number of Teachers	23,306	15,910	8,976	1,544	2,052
Black	11.1%	30.9%	18.9%	8.9%	40.1%
Hispanic	8.9%	17.7%	11.2%	9.1%	7.9%
Female	79.8%	66.7%	66.7%	72.6%	72.9%
Median Age at Hire	27	29	27	23	36
Graduate Education at Hire	35.5%	15.0%	13.9%	3.6%	60.1%
College SAT Math Pctile	59	55	68	74	*n/a*
College SAT Verbal Pctile	63	59	73	79	*n/a*

Note: This table includes data for teachers hired during the 1999-2000 to 2004-2005 school years. Age at hire is calculated as the difference between school year and birth year. College data is unavailable for over 99% of international teachers, so we do not report means of SAT percentiles for this group.

Table 2: School Avg. Student Characteristics, by Certification and Program

	Regular Certified	*Regular Uncertified*	*Teaching Fellow*	*Teach For America*	*Internat'l Program*
Black or Hispanic	72.8%	78.4%	88.1%	96.2%	89.1%
Free Lunch	70.5%	72.4%	80.3%	88.9%	77.4%
English Language Learner	13.4%	13.6%	15.3%	19.0%	14.2%
Special Education	10.3%	10.3%	10.9%	13.2%	10.2%
Math Pass (ES)	78.1%	69.8%	69.2%	68.6%	66.3%
English Pass (ES)	83.7%	79.3%	76.5%	75.9%	77.4%
Math Pass (HS)	68.9%	66.8%	65.6%	*n/a*	63.5%
English Pass (HS)	75.3%	75.3%	72.5%	*n/a*	71.6%
Graduation Rate	58.8%	57.2%	55.4%	*n/a*	53.5%

Note: "Math/English Pass (ES)" denotes the fraction of students scoring above level 1 on city and state exams for grades 3-8. "Math/English Pass (HS)" denotes the fraction of students scoring above a 55 on the state Regents examinations. High school information is not given for Teach for America corps members because very few of these teachers work in high schools; less than 3 percent of all TFA corps members teach in schools that do not serve grades three to eight.

Table 3: Sample Selection of Teachers Hired Since 1999, by Certification and Program

	All Teachers	*Regular Certified*	*Regular Uncertified*	*Teaching Fellow*	*Teach For America*	*Internat'l Program*
Total	51,977	23,306	15,910	8,976	1,544	2,052
Worked in School Serving Grades 4·	42,075	19,160	12,979	7,060	1,431	1,356
Percent of Total	*81%*	*82%*	*82%*	*79%*	*93%*	*66%*
Matched with Student Data	16,542	6,447	5,939	2,933	592	602
Percent of Total	*32%*	*28%*	*37%*	*33%*	*38%*	*29%*
Regression Sample	10,040	4,465	3,037	1,749	416	366
Percent of Total	*19%*	*19%*	*19%*	*19%*	*27%*	*18%*

Note: This includes only teachers hired from the 1999-2000 through 2004-2005 school years. Regression sample refers to teachers successfully matched with students included in our analysis. See the text for additional details on sample selection.

Table 4: Teacher Characteristics by Certification, Program, and Sample Selection

	Regular Certified		*Regular Uncertified*		*Teach For America*		*Internat'l Program*		*Teaching Fellow*	
	Full Sample	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*
Number of Teachers	23,306	4,027	15,910	3,113	1,544	445	2,052	419	8,976	1,845
Black	11.1%	13.4%	30.9%	37.9%	8.9%	11.2%	40.1%	64.9%	18.9%	24.6%
Hispanic	8.9%	8.0%	17.7%	16.2%	9.1%	9.4%	7.9%	6.4%	11.2%	10.0%
Female	79.8%	82.5%	66.7%	71.2%	72.6%	68.3%	72.9%	81.1%	66.7%	66.6%
Median Age at Hire	27	27	29	28	23	23	36	38	27	27
College SAT Math Pctile	59	58	55	55	74	75	*n/a*	*n/a*	68	67
College SAT Verbal Pctile	63	61	59	59	79	81	*n/a*	*n/a*	73	73
Graduate Education	35.5%	36.9%	15.0%	12.9%	3.6%	4.3%	60.1%	66.3%	13.9%	15.8%

Note: This table includes data for teachers hired during the 1999-2000 to 2004-2005 school years. Regression sample refers to teachers successfully matched with students included in our analysis. See the text for additional details on sample selection.

Table 5: Student Characteristics in Full and Regression Samples, by Certification and Program

	Regular Certified		*Regular Uncertified*		*Teaching Fellow*		*Teach For America*		*Internat'l Program*	
	Full Sample	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*	*Full Sample*	*Regression Sample*
Black or Hispanic	72.8%	69.8%	78.4%	78.2%	88.1%	90.4%	96.2%	96.5%	89.1%	92.7%
Free Lunch	70.5%	67.2%	72.4%	74.3%	80.3%	80.9%	80.3%	84.6%	88.9%	79.8%
English Language Learner	13.4%	2.5%	13.6%	3.5%	15.3%	5.2%	15.3%	7.6%	19.0%	4.7%
Special Education	10.3%	0.3%	10.3%	0.4%	10.9%	0.2%	10.9%	0.2%	13.2%	0.2%
Math Pass (ES)	78.1%	83.7%	69.8%	76.3%	69.2%	75.8%	69.2%	70.7%	68.6%	69.0%
English Pass (ES)	83.7%	90.4%	79.3%	86.6%	76.5%	84.8%	76.5%	81.3%	75.9%	81.5%

Note: Characteristics in the full sample represent the school average characteristics of students in the school in which teachers work. Characteristics in the regression sample represent the characteristics of students in teachers' classrooms. The regression sample does not include classrooms: in which fewer than seven or more than 45 students are tested; where the teacher did not work in the school during the entire year; where the teacher is listed as working in more than one school per year; where less than 75% of students in the school-year cell were successfully matched to a teacher; where where 25% or more of the students receive special education. In addition, the regression sample does not include students who were not tested in the prior year, since that is used as a control variable. "Math/English Pass (ES)" denotes the fraction of students scoring above level 1 on city and state exams for grades 3-8.

Table 6: Differences Between Teacher Certification Groups in Math and Reading Value-Added

	Math					Reading				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Teaching Fellow	-0.199	0.004	0.007	0.004	0.000	-0.238	-0.017	-0.012	-0.016	-0.012
	(0.014)	(0.005)	(0.005)	(0.005)	(0.004)	(0.014)	(0.005)	(0.004)	(0.004)	(0.004)
Teach for America	-0.280	0.012	0.019	0.031	0.024	-0.311	-0.018	-0.003	-0.000	0.005
	(0.022)	(0.010)	(0.009)	(0.009)	(0.009)	(0.023)	(0.008)	(0.007)	(0.007)	(0.008)
Internat'l Programs	-0.476	-0.054	-0.027	-0.029	-0.023	-0.445	-0.030	0.004	0.002	0.004
	(0.025)	(0.008)	(0.008)	(0.008)	(0.007)	(0.026)	(0.009)	(0.008)	(0.008)	(0.007)
Other Uncertified	-0.126	-0.005	0.002	0.001	-0.000	-0.128	-0.010	-0.000	0.002	0.005
	(0.011)	(0.004)	(0.003)	(0.003)	(0.003)	(0.011)	(0.003)	(0.003)	(0.003)	(0.003)
Student Covariates		√	√	√	√		√	√	√	√
Class Average Covariates			√	√	√			√	√	√
School Average Covariates			√					√		
School FE				√					√	
School*Grade*Year FE					√					√
Grade*Year FE	√	√	√	√		√	√	√	√	
Sample Size	1,462,100	1,462,100	1,462,100	1,462,100	1,462,100	1,366,479	1,366,479	1,366,479	1,366,479	1,366,479
R^2	0.03	0.67	0.68	0.68	0.70	0.03	0.63	0.64	0.64	0.66

Note: Coefficients on indicator variables for TOP scholars and Peace Corps fellows were also estimated but were not reported given large standard errors. All specifications include dummy variables for teacher experience, a dummy for those missing experience and a dummy for those hired before the 1999-2000 school year. Student-level covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, and an indicator for English Language Learners. Each of these were also interacted with grade level. The school average and class average covariates included the school-level and classroom-level means of the student-level covariates and class size, each also interacted with grade level. Standard errors (in parentheses) allow for clustering within school, grade level and year.

Table 7: Additional Estimates of Between Group Differences in Teacher Value-Added

Math Test Scores	*Baseline Estimates from Table 6 Column 3*	*Students in Elementary Grades*	*Students in Middle Grades*	*Schools w/ Above Median Avg Test Scores*	*Schools w/ Below Median Avg Test Scores*	*Teachers with Zero Years of Experience*	*Teachers with One Year of Experience*	*Teachers with Two Years of Experience*
Teaching Fellow	0.007	0.001	0.003	0.006	0.001	-0.009	0.005	0.018
	(0.005)	(0.007)	(0.007)	(0.008)	(0.006)	(0.008)	(0.009)	(0.012)
	[1,261]	[784]	[477]	[336]	[925]	[505]	[410]	[203]
Teach for America	0.019	0.015	0.027	0.039	0.016	-0.006	0.016	0.021
	(0.009)	(0.015)	(0.011)	(0.026)	(0.010)	(0.013)	(0.017)	(0.031)
	[260]	[145]	[115]	[41]	[219]	[127]	[96]	[20]
Internat'l Programs	-0.027	0.002	-0.033	-0.028	-0.036	-0.005	-0.034	-0.024
	(0.008)	(0.017)	(0.009)	(0.017)	(0.009)	(0.029)	(0.028)	(0.024)
	[238]	[76]	[162]	[49]	[189]	[16]	[15]	[16]
Uncertified Teachers	0.002	-0.005	-0.001	0.000	0.001	-0.017	-0.015	-0.000
	(0.003)	(0.005)	(0.005)	(0.005)	(0.005)	(0.007)	(0.008)	(0.008)
	[2,210]	[1,243]	[967]	[883]	[1,327]	[539]	[498]	[430]
Sample Size	1,462,100	749,552	712,548	796,388	665,712	129,000	163,612	139,941
R^2	0.68	0.66	0.70	0.66	0.60	0.65	0.65	0.66

Reading	*Baseline Estimates from Table 6 Column 3*	*Students in Elementary Grades*	*Students in Middle Grades*	*Schools w/ Above Median Avg Test Scores*	*Schools w/ Below Median Avg Test Scores*	*Teachers with Zero Years of Experience*	*Teachers with One Year of Experience*	*Teachers with Two Years of Experience*
Teaching Fellow	-0.012	-0.018	-0.009	-0.009	-0.009	-0.018	-0.007	-0.003
	(0.004)	(0.006)	(0.006)	(0.009)	(0.005)	(0.007)	(0.008)	(0.011)
	[1,246]	[756]	[490]	[289]	[957]	[546]	[397]	[205]
Teach for America	-0.003	-0.012	-0.001	0.002	0.001	-0.015	0.005	-0.025
	(0.007)	(0.011)	(0.009)	(0.017)	(0.008)	(0.011)	(0.012)	(0.021)
	[301]	[140]	[161]	[43]	[258]	[167]	[100]	[18]
Internat'l Programs	0.004	-0.008	0.006	0.045	-0.011	0.062	-0.018	-0.025
	(0.008)	(0.015)	(0.009)	(0.017)	(0.009)	(0.027)	(0.025)	(0.031)
	[219]	[78]	[141]	[51]	[168]	[12]	[15]	[11]
Uncertified Teachers	-0.000	0.002	-0.006	-0.002	0.004	-0.013	-0.010	-0.012
	(0.003)	(0.005)	(0.005)	(0.005)	(0.004)	(0.007)	(0.007)	(0.007)
	[2,100]	[1,214]	[886]	[832]	[1,268]	[450]	[500]	[427]
Sample Size	1,366,479	746,465	620,014	768,767	597,712	125,857	159,126	140,292
R^2	0.64	0.63	0.66	0.62	0.57	0.61	0.62	0.63

Note: The number of teachers from each group included in the regression is shown in brackets. All specifications (excluding those where experience is used to restrict the sample) include dummy variables for years of teaching experience and a dummy for those missing experience. They also include student-level covariates, and class and school average student covariates, each interacted with grade level. Student covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, an indicator for English Language Learners. Elementary grades are 4 and 5, middle grades are 7 and 8. 6th graders are considered elementary if they attend a school where the maximum grade is 6, otherwise they are considered middle. Standard errors (in parentheses) allow for clustering within school, grade level and year.

Table 8: Teaching Fellows' Academic Credentials and Value-Added

	Math			Reading		
	(1)	(2)	(3)	(4)	(5)	(6)
Teaching Fellow	0.005	0.004	0.004	-0.013	-0.012	-0.012
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
College SAT Math Decil	0.003		0.003	0.000		0.000
	(0.002)		(0.002)	(0.002)		(0.002)
College GPA		-0.008	-0.009		-0.007	-0.008
		(0.011)	(0.011)		(0.011)	(0.011)
SAT Math * GPA			-0.006			-0.001
			(0.006)			(0.006)
Sample Size	1,462,100	1,462,100	1,462,100	1,366,479	1,366,479	1,366,479
R^2	0.68	0.68	0.68	0.64	0.64	0.64

Note: Standard errors (in paretheses) allow for clustering within school, grade level and year. All specifications include dummy variables for participation in other recruitment programs, being an uncertified teacher, teacher experience, a dummy for those missing experience, a dummy for those hired before the school year 1999-2000, and student, classroom, and school level covariates. Student-level covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, an indicator for English Language Learners. Each of these were also interacted with grade level. The classroom and school covariates included the classroom-level and school-level means of all the student-level covariates, each also interacted with grade level. Controls are also included for a teaching fellow missing either GPA or SAT scores when those variables are included in the regressions.

Table 9: Sorting on Observables of Teachers Among Schools

	Has a Graduate Degree		Percentile Math SAT of Undergaduate Inst.		Percentile Verbal SAT of Undergaduate Inst.		Age (in Years)		Black or Hispanic Race/Ethnicity	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Teaching Fellow	-0.211	-0.190	8.217	8.632	9.704	9.097	1.830	0.749	0.111	0.013
	(0.005)	(0.005)	(0.369)	(0.442)	(0.418)	(0.499)	(1.074)	(1.189)	(0.006)	(0.006)
Teach for America	-0.279	-0.236	16.460	17.413	17.090	16.213	-4.368	-5.155	-0.010	-0.177
	(0.011)	(0.012)	(2.297)	(2.363)	(2.599)	(2.668)	(2.258)	(2.541)	(0.012)	(0.012)
Internat'l Programs	0.108	0.126	0.614	-2.961	5.120	2.569	7.489	6.238	0.273	0.164
	(0.009)	(0.010)	(7.543)	(7.547)	(8.534)	(8.522)	(1.991)	(2.128)	(0.010)	(0.010)
Uncertified Teachers	-0.219	-0.211	-2.138	-2.258	-1.928	-1.458	1.661	1.071	0.249	0.209
	(0.005)	(0.005)	(0.415)	(0.457)	(0.469)	(0.516)	(0.984)	(1.065)	(0.005)	(0.005)
School Fixed Effects		√		√		√		√		√
Sample Size	52,977	52,977	15,147	15,147	15,147	15,147	52,977	52,977	52,977	52,977

Note: The sample is limited to all teachers hired before the 1999-2000 school year, observed in their first year of working in New York City. Differences were also estimated for TOP scholars and Peace Corps Fellows but were not reported given the small number of teachers in these programs.

Table 10: The Returns to Teaching Experience for Math and Reading Value-Added

	Math						Reading	
(Relative to 0 Years Experience)	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1 Year Experience	0.033	0.037	0.027	0.030	0.048	0.039	0.023	0.019
	(0.004)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
2 Years Experience	0.052	0.059	0.044	0.049	0.074	0.064	0.034	0.031
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
3 Years Experience	0.055	0.073	0.045	0.051	0.088	0.076	0.035	0.035
	(0.004)	(0.005)	(0.004)	(0.004)	(0.005)	(0.006)	(0.004)	(0.005)
4 Years Experience	0.061	0.080	0.051	0.058	0.102	0.090	0.041	0.041
	(0.005)	(0.006)	(0.005)	(0.005)	(0.006)	(0.007)	(0.005)	(0.006)
5 Years Experience	0.056	0.087	0.046	0.054	0.107	0.094	0.045	0.049
	(0.005)	(0.007)	(0.005)	(0.005)	(0.007)	(0.007)	(0.005)	(0.007)
6 Years Experience	0.058	0.089	0.048	0.056	0.117	0.103	0.035	0.042
	(0.005)	(0.008)	(0.005)	(0.005)	(0.007)	(0.008)	(0.005)	(0.008)
7+ Years Experience	0.051	0.088	0.050	0.050	0.130	0.115	0.044	0.052
	(0.004)	(0.008)	(0.005)	(0.004)	(0.007)	(0.008)	(0.004)	(0.009)
Teacher Fixed Effects		√						√
Cohort Fixed Effects			√			√		
Attrition Interactions				√		√		
Initial Experience Effects					√	√		
Sample Size	1,462,100	1,462,100	1,462,100	1,462,100	1,462,100	1,462,100	1,366,086	1,366,086
R^2	0.68	0.71	0.68	0.68	0.68	0.68	0.64	0.66

Note: Standard errors allow for clustering within school, grade level and year. All specifications include student, classroom, and school covariates, as well as dummy variables for teacher certification status and a dummy for those missing experience. Student-level covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, an indicator for English Language Learners. Each of these were also interacted with grade level. Classroom level and school-level covariates included the classroom-level and school-level means of all the student-level covariates, each also interacted with grade level.

Table 11: Differences in Math and Reading Impacts by Teacher Experience for Regular Certified Teachers and Teaching Fellows

	Math	Reading
2nd Year Teachers	0.033	0.018
(Relative to 1st Yr)	(0.006)	(0.006)
3rd Year	0.047	0.027
	(0.007)	(0.006)
4th Year	0.069	0.037
	(0.009)	(0.008)
5th+ Year	0.082	0.048
	(0.010)	(0.010)
Teaching Fellow	0.011	0.010
*2nd Year	(0.009)	(0.010)
TF*3rd Year	0.030	0.036
	(0.013)	(0.013)
TF*4th Year	0.014	0.015
	(0.018)	(0.017)
TF*5th+ Year	0.013	0.023
	(0.028)	(0.030)
TFA Corps Member	-0.006	0.022
*2nd Year	(0.015)	(0.013)
TFA*3rd Year	-0.001	0.012
	(0.030)	(0.024)
TFA*4th Year	-0.042	0.037
	(0.041)	(0.027)
TFA*5th+ Year	-0.136	0.023
	(0.091)	(0.048)
Uncertified Teacher	0.016	-0.007
*2nd Year	(0.008)	(0.010)
Uncertified*3rd Year	0.029	-0.007
	(0.010)	(0.011)
Uncertified*4th Year	0.015	-0.007
	(0.011)	(0.012)
Uncertified*5th+ Year	0.006	-0.019
	(0.013)	(0.014)
Teaching Fellow Experience Interactions = 0 (p-val)	0.197	0.067
TFA Experience Interactions = 0 (p-val)	0.558	0.462
Uncertified Experience Interactions = 0 (p-val)	0.031	0.736
Sample Size	1,462,100	1,366,479
R^2	0.71	0.66

Note: Standard errors allow for clustering within school, grade level and year. All specifications include dummy variables for teacher certification status and a dummy for those missing experience, as well as student-, classroom-, and school-level covariates. Student-level covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, an indicator for English Language Learners. Each of these were also interacted with grade level. The classroom- and school-level covariates included the classroom- and school-level means of all the student-level covariates also interacted with grade level.

Table 12: Variation of Teacher Quality Within Groups

A. Elementary School

	Math		Reading	
Standard Deviation:	Total	Signal	Total	Signal
Full sample	0.21	0.13	0.20	0.10
Novices				
All Novices	0.21	0.12	0.19	0.08
Traditionally Certified	0.20	0.12	0.19	0.08
Uncertified	0.22	0.12	0.19	0.09
Teaching Fellows	0.22	0.12	0.20	0.07

B. Middle School

	Math		Reading	
	Total	Signal	Total	Signal
Full sample	0.17	0.08	0.17	0.06
Novices				
All Novices	0.16	0.09	0.16	0.07
Traditionally Certified	0.16	0.08	0.16	0.07
Uncertified	0.16	0.09	0.15	0.07
Teaching Fellows	0.16	0.09	0.17	0.08

C. Middle School Teaching Same Course

	Math		Reading	
	Total	Signal	Total	Signal
Full sample	0.17	0.09	0.16	0.07
Novices				
All Novices	0.16	0.10	0.16	0.08
Traditionally Certified	0.16	0.10	0.16	0.09
Uncertified	0.17	0.10	0.16	0.08
Teaching Fellows	0.16	0.10	0.16	0.09

Note: "Total" is the variation among classroom-mean residuals in a regression of student test scores on student, classroom, teacher, and school characteristics. "Signal" is the covariance among mean residuals for classrooms taught by the same teacher. Regressions include dummy variables for teacher certification and program participation (e.g. teaching fellows), teacher experience, a dummy for those missing experience and a dummy for those hired before the 1999-2000 school year. They also include student, classroom, and school level covariates. Student-level covariates include a cubic polynomial in both prior-year math and reading scores, gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, and an indicator for English Language Learners. Each of these were also interacted with grade level. The school average and class average covariates included the school-level and classroom-level means of the student-level covariates and class size, each also interacted with grade level. Standard errors (in parentheses) allow for clustering within school, grade level and year.

Table A1: Additional Coefficent Estimates from Regressions in Table 10, Columns (3) Through (6)

(Relative to 1993 Cohort or Prior)	(3)	(4)	(5)	(6)
1994 Cohort	0.014			0.011
	(0.005)			(0.006)
1995 Cohort	0.017			0.015
	(0.005)			(0.006)
1996 Cohort	0.015			0.014
	(0.006)			(0.007)
1997 Cohort	0.018			0.015
	(0.005)			(0.005)
1998 Cohort	0.009			0.002
	(0.004)			(0.005)
1999 Cohort	0.015			0.006
	(0.004)			(0.005)
2000 Cohort	0.014			0.002
	(0.005)			(0.005)
2001 Cohort	0.011			0.002
	(0.005)			(0.005)
2002 Cohort	0.012			0.006
	(0.005)			(0.005)
2003 Cohort	0.000			-0.003
	(0.006)			(0.006)
2004 Cohort	-0.004			-0.004
	(0.006)			(0.007)
2005 Cohort	-0.041			-0.037
	(0.008)			(0.009)
(Relative to Non-Attriters)				
Attrited with Experience = 0		-0.033		0.005
		(0.009)		(0.006)
Attrited with Experience = 1		0.006		0.005
		(0.006)		(0.007)
Attrited with Experience = 2		0.005		0.004
		(0.007)		(0.006)
Attrited with Experience = 3		0.004		0.016
		(0.006)		(0.010)
Attrited with Experience = 4		0.017		-0.007
		(0.010)		(0.012)
Attrited with Experience = 5		-0.008		0.005
		(0.012)		(0.012)
Attrited with Experience = 6		0.006		-0.002
		(0.012)		(0.003)
Attrited with Experience = 7+		-0.005		0.000
		(0.003)		(0.000)
(Relative to Initial Experience = 0)				
Initial Experience = 1			-0.026	-0.032
			(0.004)	(0.005)
Initial Experience = 2			-0.033	-0.051
			(0.005)	(0.006)
Initial Experience = 3			-0.051	-0.059
			(0.005)	(0.007)
Initial Experience = 4			-0.060	-0.065
			(0.006)	(0.007)
Initial Experience = 5			-0.066	-0.065
			(0.007)	(0.008)
Initial Experience = 6			-0.067	-0.073
			(0.007)	(0.008)
Initial Experience = 7+			-0.083	-0.076
			(0.007)	(0.074)
Cohort Fixed Effects	√			√
Attrition Interactions		√		√
Initial Experience Effects			√	√
Sample Size	1,462,100	1,462,100	1,462,100	1,462,100
R^2	0.68	0.68	0.68	0.68

Note: Dependent variable is math test scores. Standard errors allow for clustering within school, grade level and year. See Table 10 for other covariates in the regressions.

Table A2: Specification Checks on Use of Quasi-Gains vs. Gains and Student FE

	Math				Reading			
	Levels	Gains	Levels	Gains	Levels	Gains	Levels	Gains
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teaching Fellow	0.007	0.009	0.000	0.005	-0.012	-0.011	-0.006	-0.009
	(0.005)	(0.005)	(0.005)	(0.006)	(0.004)	(0.004)	(0.004)	(0.006)
Teach for America	0.019	0.020	0.024	0.035	-0.003	-0.002	-0.009	-0.006
	(0.009)	(0.009)	(0.009)	(0.012)	(0.007)	(0.007)	(0.007)	(0.009)
Internat'l Programs	-0.027	-0.020	-0.034	-0.043	0.004	0.008	-0.003	-0.004
	(0.008)	(0.008)	(0.008)	(0.010)	(0.008)	(0.008)	(0.009)	(0.010)
Other Uncertified	0.002	0.005	0.004	0.000	-0.000	0.000	0.005	0.000
	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Prior Test Scores	√				√			
Student Fixed Effects			√	√			√	√
Student Covariates	√	√			√	√		
Sample Size	1,462,100	1,462,100	1,462,100	1,462,100	1,366,479	1,366,479	1,366,479	1,366,479

Note: Standard errors allow for clustering within school, grade level and year. All specifications include dummy variables for teacher experience, a dummy for those missing experience and a dummy for those hired before the school year 1999-2000. Differences were also estimated for TOP scholars and Peace Corps Fellows but were not reported given the small number of teachers in those groups. Prior test score controls include a cubic polynomial in both prior-year math and reading scores. All specifications include class-, and school-level covariates. Student-level covariates include gender, six categories for race and ethnicity, an indicator for Free/Reduced Price Lunch status, an indicator for special education status, an indicator for English Language Learners. The classroom level and school-level covariates included the classroom-level means of all the student-level covariates. All of these are interacted with grade level.

Figure 1
Percent of Uncertified Hires in 1999-2000, by Decile of Pass Rate on Math Examinations

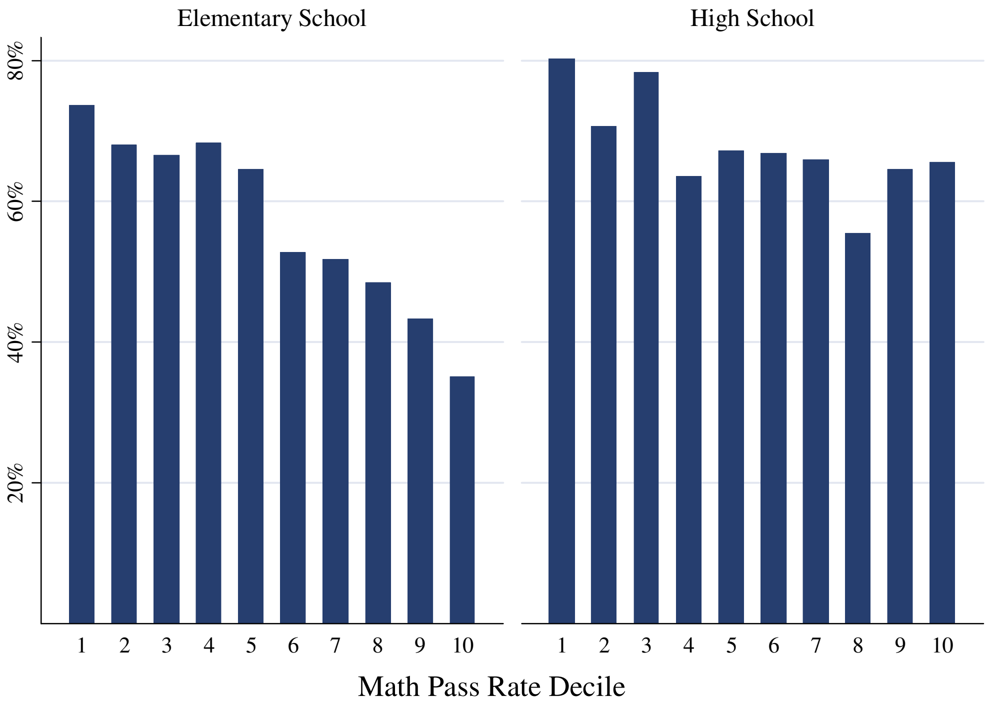
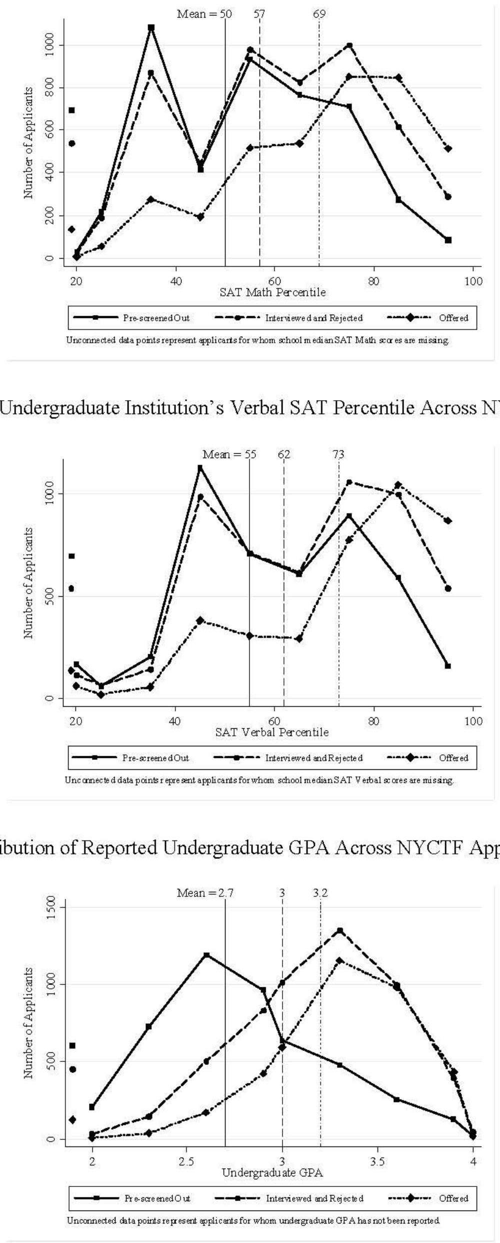


Figure 2: Academic Credentials and NYCTF Applicant Status



Panel A: Distribution of Undergraduate Institution's Math SAT Percentile Across NYCTF Applicant Groups

Unconnected data points represent applicants for whom school median SAT Math scores are missing.

Panel B: Distribution of Undergraduate Institution's Verbal SAT Percentile Across NYCTF Applicant Groups

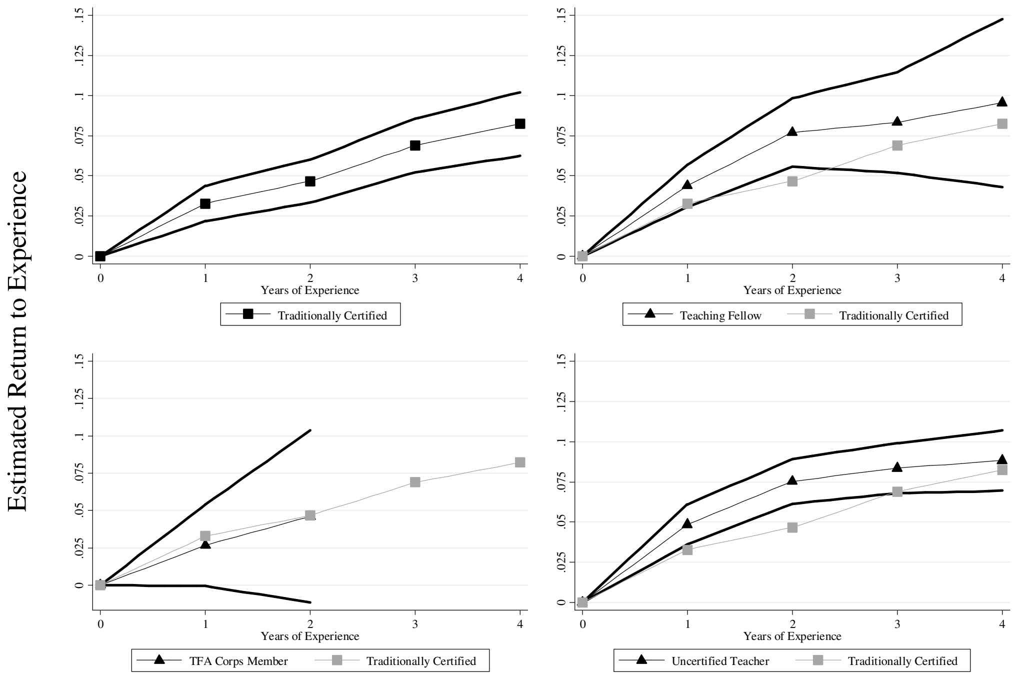
Unconnected data points represent applicants for whom school median SAT Verbal scores are missing.

Panel C: Distribution of Reported Undergraduate GPA Across NYCTF Applicant Groups

Unconnected data points represent applicants for whom undergraduate GPA has not been reported.

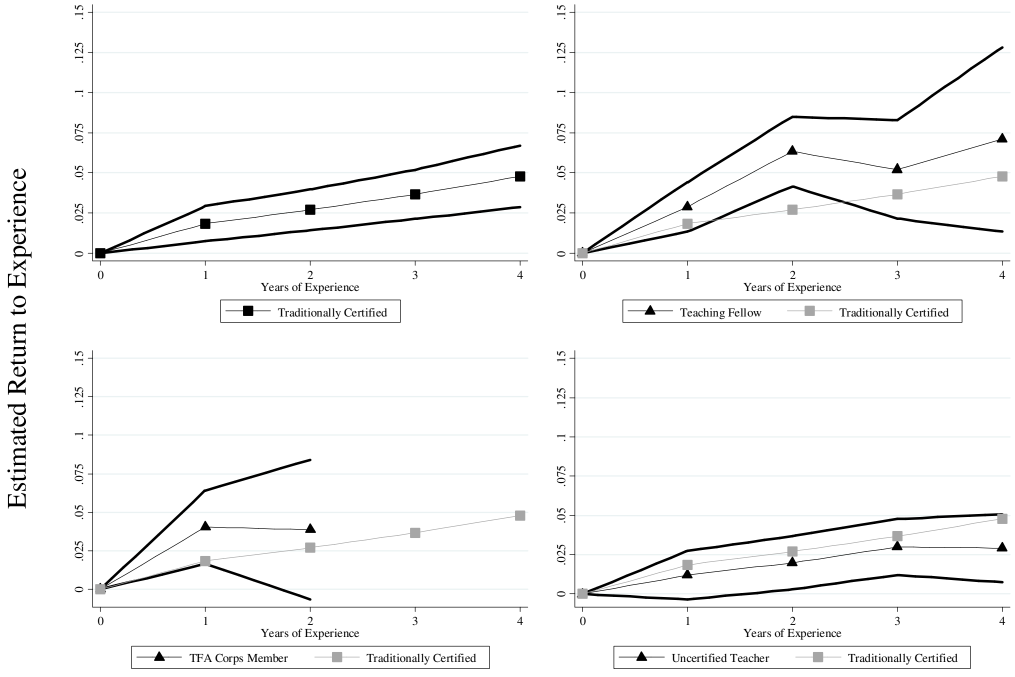
Figure 3, Heterogeneous Returns to Experience Across Groups of Teachers:

Math Test Scores



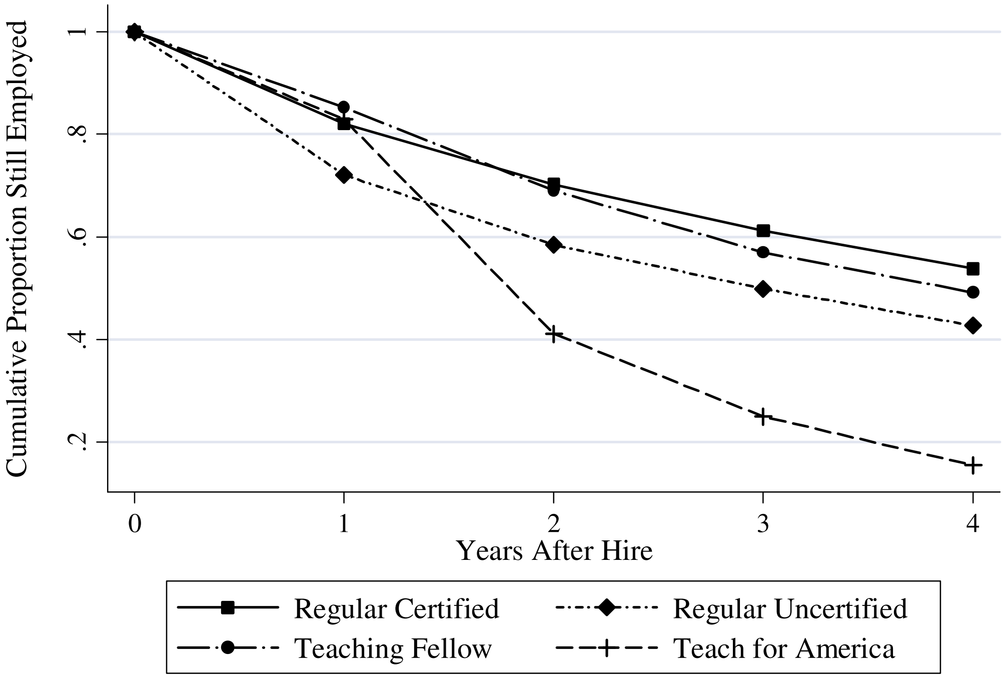
Note: Plotted are point estimates and 95% confidence intervals of experience effects for different groups. Estimates are taken from regressions that include teacher fixed effects.

Reading Test Scores



Note: Plotted are point estimates and 95% confidence intervals of experience effects for different groups. Estimates are taken from regressions that include teacher fixed effects.

Figure 4
Cumulative Retention Estimates for Certification and Program Groups



Estimates from Logit Regression on Teacher Leaving NYC

Figure 5
Steady State Distributions of Teacher Experience by Group

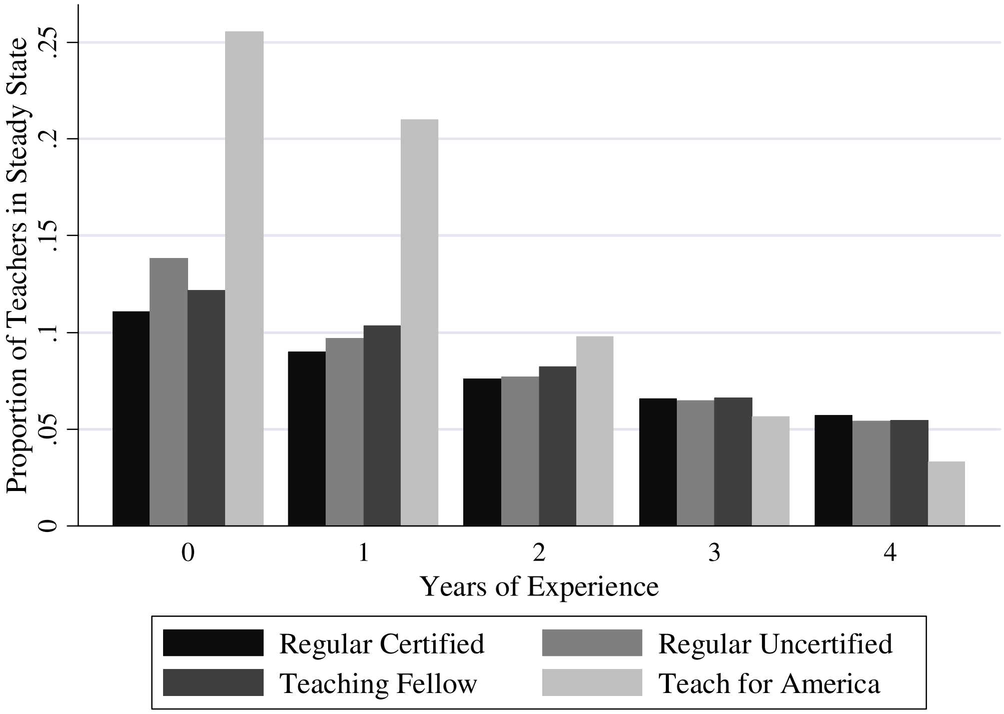
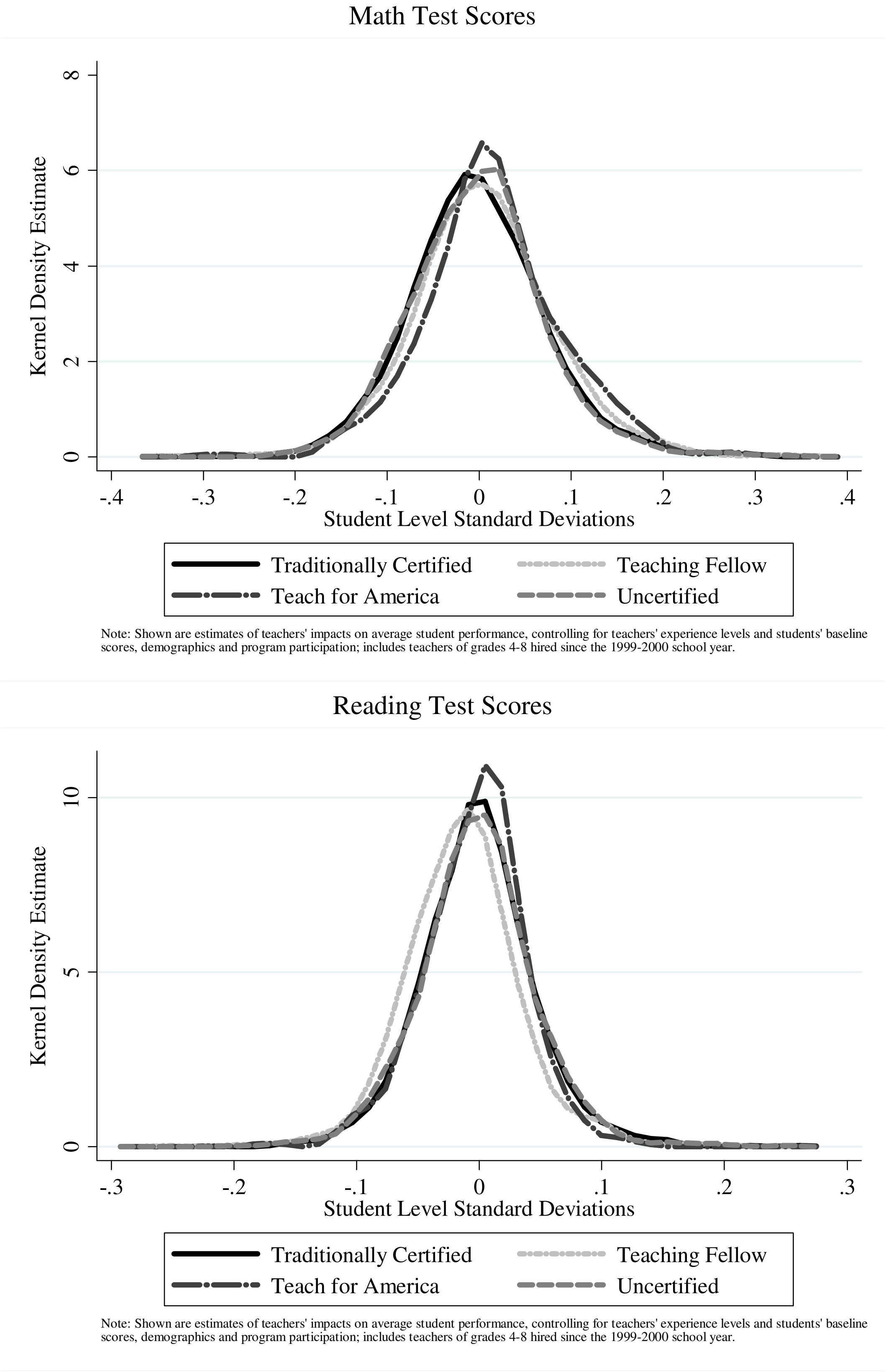


Figure 6, Variation in Value-Added Within and Between Groups of Teachers



Math Test Scores

Note: Shown are estimates of teachers' impacts on average student performance, controlling for teachers' experience levels and students' baseline scores, demographics and program participation; includes teachers of grades 4-8 hired since the 1999-2000 school year.

Reading Test Scores

Note: Shown are estimates of teachers' impacts on average student performance, controlling for teachers' experience levels and students' baseline scores, demographics and program participation; includes teachers of grades 4-8 hired since the 1999-2000 school year.