Using Imperfect Information to Identify Effective Teachers

by

Thomas J. Kane UCLA and NBER	Douglas O. Staiger Dartmouth College and NBER

April 25, 2005

Preliminary Draft
Please do not cite or circulate without authors' permission.

The authors acknowledge the generous support of the Andrew W. Mellon Foundation and the Spencer Foundation. Eric Taylor provided assistance with the literature review.

I. Introduction

As panel data on students and their teachers become available, it is increasingly apparent that there is considerable heterogeneity in the achievement gains produced by individual teachers (Aaronson, Barrow and Sander (2003), Rivkin, Hanushek and Kain (2005), Sanders and Horn (1994), Rockoff (2004)). According to Hanushek, Kain, O'Brien and Rivkin (2005), about half the variance in classroom level student achievement gains in Texas is attributable to heterogeneity in teacher effects-- persistent differences in the gains achieved by individual teachers. In this paper, we evaluate differences in achievement gains produced by the marginal and average teacher—at the time of recruitment as well as at the point of retention, as some teachers decide to exit the district. We also explore the prospects of using imperfect measures of teaching effectiveness during the first few years of a teacher's career to identify and screen out ineffective teachers at the time of the tenure decision.

Efforts to improve the quality of the teacher labor force have focused on better screening at the time of recruitment. For example, when required by the No Child Left Behind Act to define what it means to be a "highly qualified" teacher, most states based the definition on a short list of qualifications (having a bachelor's degree, completing an approved certification program and passage of a test of basic content knowledge) rather than demonstrated performance on the job.

We begin by evaluating the predictive value of easily observable traits—such as educational attainment and credential status-- in identifying effective teachers. We use panel data on students and teachers in the Los Angeles Unified School District to test the relationship between various teacher characteristics—such as whether or not a teacher had a teaching certificate when they were hired, whether they entered an alternative certification program or whether they held "emergency credentials" (uncertified teachers who did not enter an alternative certification training program)-- and student performance. After controlling for student characteristics and school fixed effects, we find no evidence that those with teaching certificates at the time of hiring are any more effective than those without traditional teaching credentials in raising student achievement.

Of course, the school district is observing more than credential status and educational attainment in its induction process. Recruiters may be able discern differences in effectiveness even if they are not apparent in the limited number of traits we observe. We test this hypothesis, by studying the aftermath of a large spike in hiring in Los Angeles following a statewide classroom reduction initiative. In a single year, LAUSD nearly tripled the number of elementary teachers hired. If the marginal teacher truly were any less effective than the average teacher, we would have expected negative consequences for student achievement. Therefore, we compare the achievement of students taught by the two cohorts of elementary teachers hired before the 1995-96 and 1996-97 school year. With or without controls for baseline characteristics, we find no statistically significant difference in math achievement for the students assigned to the 1996 and 1997 cohorts. Moreover, a small statistically significant difference in reading achievement disappeared once baseline controls were included. Such results imply that it may be difficult to discern differences in teaching effectiveness at intake. Moreover, since a larger share of the 1997 cohort was uncertified, it provides an indirect test of the effect of certification status on student learning. The absence of any impact on student achievement is consistent with our cross-sectional results, suggesting certified teachers were no more effective than the uncertified.

Second, we study the relationship between teacher experience and student achievement in the first few years of a teacher's career. Large samples allow us to identify growth in effectiveness by single year of experience. The effect of experience on student achievement appears to be sharply non-linear in the first few years of teaching. Between the first and second year of teaching, test scores rose by roughly 1.5 points (roughly .075 student level standard deviations). Between the second and third year of teaching, student achievement rise by roughly .5 points (.025 student-level standard deviations). After the third year of teaching, the estimated payoff to experience is not significantly different from zero.

With panel data on teachers, we can also estimate the returns to experience over the course of a typical career ("within-teacher" as opposed to "between-teacher"). We find evidence that teachers leaving the district tend to be less effective than the teachers who remain, leading to some upward bias in the returns to experience in the cross-sectional results. The results suggest that teacher effectiveness rises by only slightly less than 2 points in the first two years of experience and remains flat thereafter.

Third, we use estimates of the signal and noise in our measures of teaching impacts to generate filtered estimates of teaching effectiveness during the first few years of teaching. Despite the measurement error resulting from small sample sizes and other non-persistent shocks to classroom level performance, we estimate that the district could learn a considerable amount about a teachers' effectiveness in the first few years of teaching. After observing one year of teaching, we could predict 57 percent of the signal variance in teachers' future effectiveness; after 2 years, 73 percent; after 3 years, 80 percent. Therefore, while student achievement from a single classroom of students (typically about 20 students per classroom in Los Angeles) is an imperfect measure of a teacher's effectiveness, the marginal value of additional information diminishes considerably after the first several years of teaching.

Finally, we describe a model for using imperfect information on teacher effectiveness during the first years of teaching to screen out ineffective teachers and raise student achievement. The model highlights the important trade-offs implicit in the decision to use performance-based measures when retaining teachers. For example, replacing a second-year teacher with a novice teacher means forfeiting the gain in student achievement that comes from the first year of experience. If a decision-maker has a very noisy measure of teacher effectiveness (and, therefore, has little reason to believe that the marginal teacher is very different from the average teacher), there may be little benefit from screening out low-performing teachers in the early years-- since the district would give up the two point gain in expected student achievement that comes from the two years of teaching. Moreover, although the district gains information by accumulating evidence on a teacher's performance over more than one year, the district also pays a price when it retains an ineffective teacher for another year.

Calibrating the model to our estimates of the payoffs to experience and the measurement error in teacher effectiveness in LAUSD, we find that imperfect measures of teacher effectiveness are sufficiently reliable to justify aggressive action even after only one year. In fact, a strategy that filters out roughly two-thirds of the lowest performing teachers after the first year of teaching would be projected to lead to a 3 to 4 point increase in student achievement in the long run (roughly one-fifth of a student level standard deviation). Although one could improve reliability by waiting until the second or third year of teaching, the payoffs to waiting are insufficient to offset the loss in student achievement from retaining ineffective teachers.

Therefore, our results suggest a very different approach to raising the quality of the teaching force. Most school districts attempt to screen out ineffective teachers at the point of hiring-- and then do little to screen out ineffective teachers afterwards. Our evidence suggests that there may be little point to screening teachers at hiring, since there is little difference in the effectiveness of the marginal and average teacher hired. Moreover, although there is evidence that those who subsequently left LAUSD were less effective than those who remained, the difference in achievement impacts between those leaving and those staying was only 1 point (about .05 student-level standard deviations). Rather, our evidence suggests that one could identify much larger differences between the marginal and average teacher—nearly 6 points—by observing a single year of teaching performance and retaining only the highest-scoring teachers. These estimates of the likely gains are particularly striking, since they incorporate estimates of the measurement error in teacher effectiveness.

II. Literature Review

Partially in response to the No Child Left Behind Act of 2001, a growing number of states and school districts are collecting annual data on students and matching it to teachers.[1] Recent research has yielded remarkably consistent estimates of the

[1] The data requirements for measuring heterogeneity in teaching effectiveness are high. First, one needs longitudinal data on achievement for individual students matched to specific teachers. Second,

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, Hanushek and Kain (2004) report very similar estimates—suggesting that a standard deviation in teacher quality is associated with .11 student-level standard deviations in math and .095 standard deviations in reading. Using data on middle school students in Chicago Public Schools, Aaronson, Barrow and Sander (2003) report that a standard deviation in teacher quality is associated with a .09 to .16 student-level standard deviation difference in performance.[2] (The latter study adjusted for sampling variation, but not for other classroom level sources of error.)

While the evidence of considerable heterogeneity in teacher effectiveness has been remarkably robust across school districts and the empirical methods used, efforts to find observable predictors of teacher performance has been less successful. For example, much of the literature fails to find a relationship between a teacher's holding a master's degree and student achievement (Murnane (1975), Summers and Wolfe (1977), Ehrenberg and Brewer (1994), Aaronson, Barrow and Sander (2003)). Many studies also fail to find a clear relationship between teacher experience and student achievement. However, the studies that allow for non-linearities in the effect of experience tend to find

achievement data are needed on an annual basis, to be able to track gains for each student over a single school year. (Prior to the No Child Left Behind Act, many states tested at longer intervals, such as 4th and 8th grade.) Third, panel data on teachers are required as well, to be able to track performance of individual teachers over time. Teacher-level panel data are needed to account for school-level or classroom level shocks to student achievement that contribute to the measurement error in classroom-level measures. In earlier work (Kane and Staiger (2002)), we showed that conventional estimates of sampling error can not account for the lack of persistence in school-level value-added estimates. There appear to be other school-level and classroom-level sources of error.

[2] Aaronson, Barrow and Sander report the variance in teacher quality to be .02 to .06 grade-level equivalents (adjusted for sampling error). In Table 1, they report the standard deviation in grade-level equivalents of 8th grade students to be 1.55. ($\sqrt{.02}/1.55=.09, \sqrt{.06}/1.55=.16$)

returns to experience early in teachers' careers. (Rivkin, Hanushek and Kain (2004) and Rockoff (2004))

The literature on the predictive value of teacher certification is more mixed. For example, a recent paper by Darling-Hammond et. al. (2005) report that students of certified teachers in Houston, Texas outperformed the students assigned to uncertified teachers. Ballou and Podgursy (2000) summarize the literature and come to the opposite conclusion-- that teacher certification is not a reliable predictor of student achievement. A recent experimental evaluation of uncertified Teach for American Corps Members finds that they outperformed traditionally certified teachers—both novice and experienced teachers. (Decker, Mayer and Glazerman (2004))

Other research has found a relationship between teaching effectiveness and the selectivity of the college a teacher attended (for example, Summers and Wolfe (1977)) and tests of teachers' verbal ability (for example, Hanushek (1971)) or teacher's own ACT (American College Testing program) scores when applying to college (Ferguson and Ladd (1996)).

In 1987, a non-profit organization, the National Board for Professional Teaching Standards (NBPTS) was created 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, provide examples of written assignments and the feedback they provided to students as well as answer a number of essay questions in a testing center. A number of states and districts provide bonuses to teachers with NBPTS certification. A recent series of papers has suggested that student achievement is higher in classrooms taught by NBPTS certified teachers. (Goldhaber and

Anthony (2004), Cavalluzzo (2004) and Vandevoort, Amrein-Beardsley and Berliner (2004)).

There is little evidence that public school districts use the research on predictors of student achievement impacts in their recruitment and salary policies. In most public schools, teacher's wages are a simple function of years of experience and educational attainment. Moreover, Ballou (1996) finds little evidence that graduates from more selective colleges are any more successful in finding jobs at public schools. Admittedly, some districts, such as LAUSD, offer salary bonuses offered those with certification by the National Board on Professional Teaching Standards. However, most of those bonuses preceded the recent evidence that those certified by NBPTS are more effective in promoting student achievement.

However, it may not be surprising that school districts place little weight on observable characteristics in their hiring and pay decisions, since such traits explain relatively little of the estimated variation in student achievement associated with teacher quality.[3] The estimates of the heterogeneity in teacher impacts suggest that difference in achievement for those students assigned to the 10th and 90th percentile teacher would be .25 student-level standard deviations. Even if we were to accept the estimate of the student achievement effects of being a certified teacher in Darling-Hammond et. al. (2005), it was quite small-- roughly .025 student-level standard deviations. The estimate of the difference in student achievement associated with NBPTS certification in Goldhaber and Anthony (2004) was also small-- just .04 to .05 student-level standard

[3] Such traits also explains very little of the variation in principal ratings. Murnane (1975) reported that easily observable traits—such as master-degree attainment, gender, years of experience, being an undergraduate education major—explained only about 20 percent of the variance in principal ratings of teacher performance.

deviations. In Hanushek, Kain, O'Brien and Rivkin (2004), a whole year of teacher experience was associated with a .1 student-level standard deviation increase in performance (and less than .1 student-level standard deviation in Rockoff (2004)). So while there are a handful of traits that are related to teacher effectiveness, none has much predictive power.

One exception may be teacher scores on tests of verbal ability and college aptitude tests, which do seem to have more predictive power. For example, in Ferguson and Ladd (1996), one standard deviation in teachers' scores on the ACT exam was associated with a .10 standard deviation difference in students' reading scores (although no impact on math scores). Hanushek (1971) reported that one standard deviation in teachers' scores on a short test of teachers' verbal ability was associated with a .074 standard deviation impact on student achievement.[4]

Our goal in this paper is to compare the predictive power of various sources of information in identifying effective teachers. First, we evaluate the quality of information being used in teacher recruitment-- by studying differences in teacher effectiveness following fluctuations in district hiring. Our results suggest that traditional measures such as having a master's degree or having a traditional teaching certificate at the time of hiring have little impact on student achievement. Second, we study differences in achievement for those leaving the school district after hiring. Again, the difference is quite small, with "stayers" generating achievement gains roughly .05 student-level standard deviations higher than "leavers".

[4] From Hanushek (1971), p. 285, the coefficient on teacher test scores was .09, the teacher test had a standard deviation of 15.8 and the student testing outcome had a standard deviation of 19.1.

Improving teacher quality will require much better methods for discriminating between more and less effective teachers. Therefore, in the final section, we evaluate the prospects for using imperfect measures of teaching effectiveness on the job to screen out ineffective teachers in their first few years of teaching. Rather than a .05 student-level standard deviation difference, even a noisy measure of teacher effectiveness would allow one to identify a .3 to.4 student level standard deviation difference between leavers and stayers. We develop a model for determining when and where to draw the line to maximize student achievement.

III. Data

During the 2002-03 school year, the Los Angeles Unified School District (LAUSD) enrolled 746,831 students (kindergarten through grade 12) and employed 36,721 teachers in 689 schools scattered throughout Los Angeles County. There were 429 elementary schools alone. Student enrollment in LAUSD exceeds that of 29 states and the District of Columbia. We focus on students in grades 2 through 5, where a given student is assigned to a single teacher for the whole school day.

Between the spring of 1999 and the spring of 2002, the Los Angeles Unified School District administered the Stanford 9 achievement test to students in grades 1 through 5. Under state regulations, exemptions are not granted to students with disabilities or poor English skills. In May 2002, our comparison of enrollment data with the testing file suggests that test scores were available for 90 percent of students enrolled in grades 2 through 5. Since we are interested in using baseline test scores to capture students' prior educational inputs, we focus on the three academic years (1999-00 through 20001-02) for which we have both baseline and follow-up test scores. In the Spring of 2003, the district (and the state) switched from the Stanford 9 to the California Achievement Test. Both tests are reported in percentile and normal curve equivalent units. Given that tests cover slightly different material, however, such measures may not be directly comparable across tests. As a result, we use the 2003 test score data only for out-of-sample predictions and use the 1999 through 2002 testing data for all our estimatino.

Although there was considerable mobility of students within the school district (9 percent of students in grades 2 through 5 attended a different school than they did the previous year), the geographic size of LAUSD ensured that most students remained within the district when they moved. Conditional on having a baseline test score, we observed a follow-up test score for 90 percent of students.

We observed snapshots of classroom assignments at the end of the fall and spring semesters. Since we were interested in evaluating individual teacher impacts, we dropped those students who switched teachers (or schools) during the course of a school year (4 percent of students with test score outcomes). We also dropped those students in special education classes or with specifically identified disabilities (3 percent of students with valid scores). Finally, we dropped classrooms with extraordinarily large (more than 36) or extraordinarily small (less than 10) enrolled students (3 percent of students with valid scores).

We also obtained snapshots of all district employees from 1994 through 2003. Therefore, for teachers who were hired since 1993, we observed initial credential status (whether they were certified or not) and actual years of teaching experience since the time of hiring.   This was important since many of those who were initially hired as uncertified or emergency credentialed teachers subsequently achieved certification.

IV. Teacher Characteristics and Student Achievement

In this section, we investigate the relationship between student achievement and teacher characteristics, by regressing student-level math performance on teacher characteristics and various sets of student, classroom and school controls. The results are reported in Table 1. The first column in Table 1 reports the coefficients on the characteristics of the teacher assigned to classroom (Cjt) from a specification of the following form, including fixed effects for year and grade (τ_{tg}):

(Column 1) $S_{it} = C_j \beta + \tau_{tg} + \varepsilon_{it}$

The list of teacher characteristics include education level (an indicator for having a master's degree or a doctorate degree), certification status and years of experience. Those hired before 1994 are identified as veterans. For those hired since 1994, we included indicators for certification status at date of hire—uncertified teachers enrolling in intern programs, uncertified teachers not enrolled in an intern program, and a category for all others. The latter category includes Teach-for-America Corps Members, those who were substitute teachers before being hired, those who part of a training program for non-teaching personnel becoming teachers (the "Career Ladder" program) and those with experience teaching in other districts before being hired in LAUSD. The reference category is traditionally certified teachers.

With no controls for school characteristics or baseline student performance, the students assigned to interns and those with emergency credentials appear to perform 4 to

6 points worse that those assigned to credentialed teachers. (The test scores measures had a standard deviation of 20 at the student level). Moreover, those students assigned to more experienced teachers also performed 3 to 6 points higher than the students assigned to novice teachers.

Some portion of the difference in student achievement associated with teacher characteristics is due to between-school differences in student performance. Teachers with more seniority receive priority when positions become available elsewhere in the district. As a result, more experienced teachers often sort themselves into schools with higher baseline academic achievement. In the second column, we report similar results after adding fixed effects for each combination of school, grade, calendar track and year (δ_{sgct}).

(Column 2) $S_{it} = C_{jt}\beta + \delta_{sgct} + \varepsilon_{it}$

The differences in academic performance are not due primarily to differences in the types of schools to which uncertified teachers are assigned. Within a given school, grade, calendar track and academic year, the students assigned to less experienced and uncertified teachers appear to perform poorly relative to those assigned to traditionally certified teachers.

However, the poor performance of student assigned to teachers hired without credentials is largely due to the fact that, even within schools, such teachers are often assigned students with lower baseline scores. The third column adds controls for student baseline scores from the previous spring (math, reading and language arts), indicators for student demographics (gender, race, participation in gifted and talented programs, participation in the free/reduced price lunch program). To capture peer effects, we also include the mean demographic characteristics of students in each classroom. (All of the above regressors are interacted with grade level.)

(Column 3) $$S_{it} = C_{jt}\beta + \gamma_{1g} S_{it-1} + \gamma_{2g} X_{it} + \gamma_{3g} \overline{X}_{jt} + \delta_{sgct} + \varepsilon_{it}$$

After including controls for baseline performance, student demographics and classroom demographics, the estimated differences in performance between students assigned to certified and uncertified teachers are considerably smaller and no longer statistically distinguishable from zero. For example, controlling for years of experience, the difference in academic achievement between those with emergency credentials and those with traditional credentials is estimated to be just .3 points (with a standard error of just .2 points) Moreover, the performance for students assigned to teachers with master's and doctorate degrees is also indistinguishable from those assigned to teachers with no graduate degrees.

The addition of baseline test scores and demographic regressors also shrinks the estimated impact of teacher experience. Rather than performing 4 to 6 test score points higher, students assigned to teachers with 3 or more years of experience perform roughly 2 points higher than students assigned to novice teachers. That represents a tenth of a student-level standard deviation. Students assigned to those with five to nine years of experience perform just as well as those assigned to veteran teachers, hired before 1994.

Any measurement error in students' baseline performance may lead us to understate in absolute value the importance of baseline performance on subsequent performance. Therefore, in column four, we include the change in math scores for each individual student as the dependent variable (essentially forcing the coefficient on S_{it-1} to be equal to one).

$$\text{(Column 4)} \qquad S_{it} - S_{it-1} = C_{jt}\beta + \gamma_{2g} X_{it} + \gamma_{3g} \overline{X}_{jt} + \delta_{sgct} + \varepsilon_{it}$$

The above specification essentially assumes that baseline test performance is a sufficient statistic for all prior educational inputs, which may be correlated with teacher characteristics. This assumption may not hold and prior educational inputs may indeed be correlated with current teacher characteristics. As a result, in column (5), we remove the student baseline performance and add student-level fixed effects in test score levels. (We continue to include interactions between student demographic characteristics and grade level).

$$\text{(Column 5)} \qquad S_{it} = C_{jt}\beta + \gamma_{2g} X_{it} + \gamma_{3g} \overline{X}_{jt} + \delta_{i} + \varepsilon_{it}$$

In column (6), we include the change in student performance over their baseline performance as the dependent variable and include a student level fixed effect—to account for differences between students in their mean growth in performance in other years.

$$\text{(Column 6)} \qquad S_{it} - S_{it-1} = C_{jt}\beta + \gamma_{2g} X_{it} + \gamma_{3g} \overline{X}_{jt} + \delta_{i} + \varepsilon_{it}$$

The results in columns (4) through (6) are quite similar to the results in column (3). After accounting for differences in baseline performance or student fixed effects as well as the demographic characteristics of classroom peers, there are no differences in performance between traditionally certified teachers and uncertified teachers.

Each of the specifications above involve a different assumption regarding the nature of the education production function process. As reported in columns (3) through (6)—whether one uses individual baseline test performance as regressors, gain scores (forcing a coefficient of minus one on prior performance) or student fixed effects in levels or gains, the coefficients on teacher initial certification status and experience are quite similar.

The failure to find a relationship between teacher certification status and student achievement is somewhat disconcerting. In California, to qualify as a "highly qualified teacher" under the No Child Left Behind Act, a teacher must either be certified or participating in a training program to become certified. Moreover, in Los Angeles, teachers who already have their teaching certificate are paid considerably more. For first-year, novice teachers, the starting salary in 2002 was $41,177 for those with a traditional teaching certificate and $35,904 for those without one—a 15 percent difference in salary.

V. Teacher Cohort Size and Teacher Quality

Figure 1 reports the hire dates of elementary school teachers working for LAUSD in May of 2003. As is dramatically apparent, there was a large increase in the number of elementary school teachers hired between 1996 and 1997. In the years before 1997, the district hired 1200 to 1400 elementary school teachers per year.[5] However, beginning in the 1996-1997 academic year, the state of California provided cash incentives to school districts to keep class sizes in kindergarten through third grade to a maximum of 20 children. In order to take advantage of the state incentive, the district dramatically increased its hiring of new elementary teachers. In a single year, between 1996 and 1997, LAUSD nearly *tripled* the number of elementary school teachers it hired from 1,297 to 3,335. Before the California Classroom Size Reduction initiative, the district was already having difficulty meeting its hiring needs with traditionally certified teachers, hiring

[5] We coded someone as being hired in the 1997 academic year, if they were hired between July 1, 1996 and June 30, 1997. We defined the other academic years in the same way.

approximately 59 percent of its elementary teachers. But in response to the California Classroom Size Reduction Initiative (CSR), the district relied even more heavily on teachers without traditional teaching credentials. (The challenge was worsened by the fact that it was a statewide policy. All of the other surrounding districts would have been seeking to hire elementary teachers at the same time.) The proportion of new hires without teaching credentials rose from about 59 percent to 72 percent of all new hires. If the district were able to effectively discern teacher effectiveness in the hiring process, one might have expected such a large increase in hiring to have had a negative impact on the average quality of the teachers hired.   In this section, we test the district's ability to discriminate between more and less effective teachers in the recruitment process, by testing for any discernible difference in student achievement impact for the 1996 and 1997 cohorts.

Table 2 reports the mean characteristics of elementary teachers hired in 1996 and 1997 (including those who subsequently left the school district's employment).   By May 2000, the 1997 cohort was in its fourth year of teaching and the 1996 cohort was in its fifth year.   The earlier results suggested that the returns to experience flatten out by the third year of teaching, implying that the experience differentials should not play much of a role when comparing the two groups. However, to the extent that there are some small positive returns to experience by the fourth year of teaching, this would tend to bias the results against the 1997 cohort.   In May 2000, 70 percent of the 1996 cohort and 75 percent of the 1997 cohort were still employed by the district.   By May of 2003, 64 percent of the 1996 cohort remained, as compared with 66 percent of the 1997 cohort. Although these differences are small, they would also tend to bias the results against the

1997 cohort, since our results (reported below) suggest that those leaving the district had achievement impacts about 1 NCE point below those remaining with the district. (Hanushek, Kain and Rivkin (2005) also report that "leavers" have lower estimated impacts on student achievement than "stayers".)

The two cohorts of teachers differed in terms of certification status and racial composition. The 1997 cohort was more likely to have been uncertified at the time of hiring (72 percent versus 59 percent). The cohort of 1997 was also more likely to be African American (15 percent versus 9 percent) and less likely to be Latino (34 percent versus 40 percent).

The table also reports the mean characteristics of the students assigned to those teachers in the spring of 2000 through 2002. Although the large sample sizes result in some of the differences being statistically significant, the magnitude of any difference is generally quite small. The class sizes assigned to the two groups of teachers were quite similar—with an average of 19.7 students per class for the cohort of 1996 and 19.5 students for the cohort of 1997. The differences in racial composition were also quite small—less than 2 percentage point difference in the percent African American, Latino or white, non-Hispanic. There was no statistically significant difference in baseline math scores. There was a small difference (.4 to .5 points) in baseline reading and language arts scores. However, with a student-level standard deviation of 20, a .4 to .5 difference in baseline reading scores is unlikely to be of substantive importance. There was less than a 1.5 percentage point difference in the proportion of students in the Gifted and Talented Program or Free/Reduced Price Lunch Program assigned to the two cohorts.

Table 3 reports results comparing student performance in 2000-2002 for the students assigned to the teacher cohorts of 1996 and 1997. We used similar specifications to those used in Table 1 above—starting with no background controls, adding fixed effects for school/grade/calendar track/year, adding regressors for student and classroom characteristics, adding student fixed effects and using gain scores rather than test score levels as the dependent variable. Table 3 reports the results of specifications with both reading and math scores as the dependent variable. We find statistically significant differences in only one specification-- when using reading scores as the dependent variable and no controls for baseline test scores. When we add baseline test scores and classroom characteristics to that specification, the difference is no longer significant.

Despite the tripling of the size of the cohort of elementary teachers hired in 1997, there was no discernible evidence that the mean effectiveness of the larger cohort was any lower. There is little evidence that the selection process used for identifying new teachers is effectively screening for teaching effectiveness.

It is also apparent in Figure 1 that the number of teachers hired in LAUSD fluctuated in earlier years—although not nearly as dramatically. For example, there were declines in the number of teachers hired beginning in 1973, 1975, 1980 and 1991. Each of these earlier downturns reflected the fortunes of the California economy, which went into recession during those periods. For example, there were 30 to 75 elementary teachers still working for the district in 2003 who had been hired each year between 1975 and 1977 when the county and state would have been facing budget crises due to the business cycle. However, there were 289 elementary teachers who had been hired in the recovery year 1978 alone.

We re-estimated a specification similar to that in column (3) of Table 1, with math score in year t as the dependent variable and including test scores from year t, student and mean classroom demographic characteristics and fixed effects for each school/grade/calendar track/year combination.

$$S_{it} = Exper_{jt}\beta + \gamma_{1g}S_{it-1} + \gamma_{2g}X_{it} + \gamma_{3g}\overline{X}_{jt} + \phi_c + \delta_{sgct} + \varepsilon_{it}$$

where ϕ_c represent fixed effects for the year in which a teacher was hired. As before, $Exper_{jt}$ is a set of dummy variables for teacher years of experience, S_{it} measures math or reading test score performance for student i, S_{it-1} is a vector of math, reading and language arts test scores from the prior year, X_{it} is a set of demographic and program participation characteristics for student i, $\overline{X}_{ct}$ is a vector of the mean demographic and program participation characteristics for the students in the class, and δ_{sgct} represents a set of fixed effects for permutations of school, grade, calendar track and year. We used the estimated cohort effects ($\hat{\phi}_c$) to estimate the relationship between cohort size and mean teacher effectiveness.

Figure 2 reports the time series in estimated cohort effects superimposed over the time series of cohort sizes for those cohorts 1977 through 2002. (Small cohort sizes in earlier years led to quite imprecise estimates.) The vertical lines in the figure identify the years in which unemployment rates in California bottomed out. There is some evidence that, during the downturns in the early Eighties and early Nineties, average teacher effectiveness improved somewhat; and during the recoveries of the late Seventies and mid-Eighties, average teacher quality declined somewhat.

To test the relationship between a cohort's size and its average impact on student achievement, we estimated the following specification:

$$\hat{\phi}_c - \hat{\phi}_{c-1} = \beta_0 + \beta_1 Trend + \beta_2 (\ln CohortSize_c - \ln CohortSize_{c-1}) + \varepsilon_c$$

The coefficient on the first-difference in the log of cohort size (β_2) represents the difference in average effectiveness between the marginal and the average teacher hired. If the marginal teacher is less effective than the average teacher, we would expect the coefficient β_2 to be negative, implying that increases in hiring are associated with declines in performance. For the period 1977 through 1996 (before the dramatic increase in hiring in 1997) the results suggest that marginal teacher generated achievement gains .8 NCE points below the average teacher, with a standard error of .3. However, the results are quite sensitive to the time period used. If we were to include the data before 1977, or exclude the observations from the late Seventies, the coefficients remain negative but are no longer statistically significant.

Why might we find a relationship between cohort size and cohort quality in early years, but not between 1996 and 1997? One hypothesis is that the coefficient on cohort size from the earlier years reflects both demand and supply side selectivity. During recessions, the school district hired fewer people and may have been more selective in its decisions. However, during recessions, the district may have received a larger, more qualified pool of applications. The smaller cohorts may have outperformed the larger cohorts simply because there was a higher quality pool applying, not necessarily because the screening process identified the more effective teachers. The difference between 1996 and 1997 is more likely to provide a focused test of demand side selectivity.

Moreover, the estimated difference of .8 NCE points between the average and marginal teacher seems quite small relative to the student level standard deviation of 20 NCE points in math. The magnitude is roughly half as large as the estimated payoff to the first year of teaching experience reported in Table 1.

VI. Post-Employment Selection: Value-Added for Stayers vs. Leavers

In this section, we study the post-employment selection of teachers. It is not clear whether one would expect the "stayers" to be more or less effective teachers than "leavers", or whether the returns to experience are over-stated or under-stated by the cross-sectional evidence. To the extent that the skills required for effective teaching—personal organization, communication skills, etc.—have a market value outside of teaching, effective teachers may be drawn away from the teaching profession. This may be particularly likely, since the pay scale in LAUSD is a function of educational attainment and experience and has no "merit" component. Moreover, effective teachers who can demonstrate their skills may be drawn away to other teaching positions—particularly at private schools where there is more flexibility in setting wages. On the other hand, those who are not successful may find the personal rewards insufficient to continue in the profession, even if their pay is unaffected. Although our interviews with district staff suggested very few teachers were explicitly terminated, principals may find non-pecuniary means for discouraging ineffective teachers and encouraging them to leave.

Since we are able to track teachers as well as students over time, we investigate the returns to experience as well as the nature of the post-employment selection process within LAUSD. Like the earlier tables, the first column of Table 4 uses both the within- and between-teacher variation to identify the effect of experience on math achievement. (The specification includes baseline student scores, student demographic characteristics, mean classroom demographic characteristics, initial hiring status of teachers, and fixed effects for combinations of school/grade/calendar track/year.)

We estimate the impact on student achievement per year of experience using a linear spline. (This is different from the estimates in Table 1, which measured cumulative differences in achievement by year of experience relative to novice teachers.) In the first column, we include fixed effects for school, grade, calendar track and year, as well as the student and classroom controls used in column (3) of Table 1. The results suggest that between the first and second year of teaching, a teacher's students gain an additional 1.6 points in math achievement. That represents .08 student level standard deviations, quite similar to the .12 standard deviation impact reported in Hanushek, Kain and Rivkin (2005). Between the 2nd and 3rd year, students gain an additional .6 points in math achievement—for a total of approximately 2 points over the course of the first two years of teaching. After the third year of teaching, there is no estimated payoff for additional experience.

The coefficient on the indicator for "veteran" teachers—those hired before 1994—measures the difference between a veteran teacher and a novice, traditionally certified teacher. The estimate suggest that veteran teachers produce 2 additional points of student achievement in math than first-year teachers. Interestingly, one could not reject the hypothesis that the sum of the coefficients on the 1st through 5th year of experience—estimated for the cohorts hired since 1994—is equal to the difference between a veteran teacher and a novice teacher. (The p-value of the test of the hypothesis that the sum of the coefficients on years of experience was equal to the coefficient on being a veteran teacher was .349).

The second column of Table 4 adds fixed effects for each teacher. In that column, the payoffs to experience are identified by growth in student achievement impacts within-teacher (at least for those hired since 1994 for whom we can observe actual experience). The estimated payoff to the first year of experience is somewhat larger, but the payoff to the second year of experience is somewhat smaller. On net, the estimated impact of the first two years of experience for individual teachers over time, 1.81 (1.481+.332) is only slightly smaller than the estimated returns in the cross-section 2.05 (1.396+.654).

Earlier research has suggested that teachers use seniority preferences to move to higher achieving schools as they gain experience. As a result, part of the "within-teacher" payoffs may actually reflect movement across schools. In column (3) of Table 4, we include fixed effects for each teacher/school permutation, thereby identifying the effect of experience only with teachers remaining in the same schools. The results are largely unaffected, implying that little of the within-teacher returns to experience are derived from teachers switching schools.

In column (4), we include a dummy variable identifying those who were no longer employed with the district in May 2003. (The comparison category are those who were still employed with the district at that time.) Given that the returns to experience over the first few years of experience declined slightly when including fixed effects, we would expect that those who left to be less effective in raising student achievement than the teachers who leave the district. (The cross-sectional return was upward biased since the first year teachers contain a disproportionate share of ineffective teachers.) The estimates in column (4) indicate that those who left the district generate gains about .8 points below the average teacher in the years prior to leaving.

In column (5), we include an indicator for the "leavers" in the year before their final year. (In columns (4)-(6), we have limited the sample to students in years 2001-2002, for whom we have a record of every teachers' employment two years ahead.) This allows us to test if the "leavers" simply have a poor showing in their final year of teaching, or if they were performing poorly even before their final year. The coefficient is positive, but statistically insignificant, suggesting that the student achievement did not simply drop-off in the final year, but were performing poorly even before their final year.

The results above suggest that more effective teachers are more likely to remain in the district and that those leaving the district had lower impacts on student achievement for at least two years prior to their departure. In column (6), we limit the sample to veterans (those hired before 1994) to test if the same type of selection continues beyond the first few years of a teacher's career. Interestingly, even among veteran teachers, there is a 1 point deficit in student achievement for those teachers who subsequently left the district. Moreover, the coefficient on the indicator of leavers in the year before their final year is not statistically significant, implying that the leavers are not simply exhibiting poor performance in their final year.

If the same selection is going on throughout a teacher's career, why are we not observing a positive return to experience beyond the first few years of experience in the cross-section? The reason is that the exit rates of teachers diminish dramatically after the first few years teaching with the district, and the amount of bias in the cross-sectional results is likely to be quite small. Among those who were in their first year of teaching in 2002, 17 percent were not working with the district in May 2003. As a result, controlling for teacher fixed effects, we find slightly slightly smaller returns to experience in the first few years of teaching. In contrast, among those with 9 or more years of experience in 2002, only 4 percent were not with the district in May 2003. If the leavers truly were 1 point less effective than the stayers, one would only expect about a .04 point bias in the estimated payoff to teaching experience for the more experienced teachers. Such an impact would represent about 2 one-*thousandths* of a student-level standard deviation. Our estimates are simply not sufficiently precise to identify such an effect.

In Table 5, we report the results of interacting the incremental effect of a year of teaching experience with a teacher's initial hiring status—whether they were interns, emergency credentialed teachers not participating in internship programs or some other initial hiring status. (In the specification, we include teacher by school fixed effects, as well as all the student-level and classroom-level covariates included in column (3) of Table 4.) Since traditionally certified novice teachers are the left out category, the coefficients on years of experience in the first column are estimated for traditionally certified teachers and the interactions measure any difference for the other groups relative to traditionally certified teachers. These coefficients measuring the effect of each incremental year of experience are similar to those in Table 4, implying that the average teachers' impact on student achievement grows by 1 to 2 points in the first two years of teaching. The incremental impact of experience is indistinguishable from zero after the first two years. The remaining columns in Table 5 report the interactions between teaching experience and teachers' initial hiring status. The bottom row of the table reports the p-value for the hypothesis that all of the interactions in a particular column are equal to zero. The point estimate on the payoff to experience during the first year is positive for emergency credentialed teachers—but it is not statistically distinguishable from zero. Indeed, despite the large sample size, we are not able to reject the hypothesis that all five of the coefficients on experience are the same for traditionally certified teachers as for interns. Nor can we reject the hypothesis that the returns to experience are the same for traditionally certified teachers as for other emergency credentialed teachers. This is perhaps surprising, given that traditionally certified teachers receive some classroom exposure as part of their graduate training.

VII. Can We Use Early Performance to Screen for Effective Teachers?

The evidence presented thus far suggests that it is difficult to identify effective teachers based on characteristics that are observable at the time of hire. In this section, we investigate whether classroom performance can be used to reliably identify effective teachers. In particular, does the classroom performance of newly hired teachers forecast persistent differences in value added in future years? To answer this question, we begin by summarizing the statistical properties of teacher effects on value-added: Is there large variation in performance across teachers, do such differences persist over time, and can it be estimated reliably? We then use this information to construct simple forecasts of teacher performance based on classroom experience from 2000 and 2001, and evaluate the ability of these forecasts to predict teacher performance in 2002 and 2003.

Statistical Properties of Teacher Effects on Student Value-Added

We estimate each teacher's value-added in each year using a specification similar to that in column (3) of Table 1, with math score of the student as the dependent variable and including teacher fixed effects and student baseline characteristics as independent variables. Specifically, we estimate the following equation separately in each year:

$$S_{it} = \gamma_{1g} S_{it-1} + \gamma_{2g} X_{it} + \phi_{jt} + \varepsilon_{it}$$

The parameters ϕ_{jt} are teacher fixed effects. As before, S_{it} measures math test score performance for student i, S_{it-1} is a vector of math, reading and language arts test scores from the prior year, and X_{it} is a set of demographic and program participation characteristics for student i.

The estimated fixed effect ($\hat{\phi}_{jt}$) represents an estimate of the teacher's value added in that year. Some of the differences across teachers can be explained by observable characteristics of the teacher and classroom (Z). We removed this predictable component with a regression of the form:

$$\hat{\phi}_{jt} = Z_{jt} \hat{\beta} + \hat{\varepsilon}_{jt}$$

The unit of observation in this regression was a teacher-year. The regression included dummy variables for the first four years of experience, a full set of grade dummies for each year, and a vector of the mean demographic and program participation characteristics for the students in the class. In some specifications we also included a set of fixed effects for permutations of school, grade, calendar track and year – although this specification will remove systematic differences in teacher quality if teachers are sorted into school by their performance. The residual (ε_{jt}) represents each teacher's individual contribution to value-added – a teacher's performance residual in year t relative to other teachers with similar observable characteristics.

There is considerable variation in this teacher performance residual in any given year. The estimated root mean squared error of ε_{jt} is 6.9 from specifications that do not control for school-grade-track-year effects, and 6.3 in specifications that do control for school-grade-track-year effects. In other words, even within school and after controlling for observable difference in experience and grade, the standard deviation in teacher value-added in any given year is over 6 NCE points – or over one quarter of a student-level standard deviation. 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 year to year is to estimate correlations in the teacher performance residuals across years. Suppose we decompose ε_{jt} into two components: a persistent component (μ_{jt}) that represents teaching ability, and a non-persistent component (ξ_{jt}) that represents sampling variation, classroom dynamics, and other idiosyncratic shocks to classroom performance. The non-persistent component is independent from year to year. If the persistent component is fixed over time (e.g., unchanging teacher ability with $\mu_{jt} = \mu_j$), then it is straightforward to show that the correlation in ε between any two years should be constant, and equal to the variance of the persistent component as a proportion of the total variance (persistent and non-persistent). In addition, changes in ε (from t-1 to t) should be uncorrelated with levels of ε from prior years (t-2), while changes in ε from adjacent years (t-1 to t, and t-2 to t-1) should be correlated –0.5. In contrast, 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 $\mu_{jt} = \rho\mu_{jt-1} + \nu_{jt}$), then the correlation in ε between any two years should be strongest in adjacent years. In addition, changes in ε from t-1 to t should be negatively correlated with levels of ε from prior years (t-2), while changes in ε from adjacent years (t-1 to t, and t-2 to t-1) should be correlated between 0 and –0.5.

In Table 6, we report the correlations in the teacher performance residual (ε) between 2002 and 2001 (one lag) and between 2002 and 2000 (two lags), and the correlations between the change in ε from 2001 to 2002 and the level in 2000 and the change from 2000 to 2001. The first two columns report the results for teacher residuals that have not removed school-grade-track-year effects, while the last two columns remove school-grade-track-year effects. We calculate this correlation for the sample of all teachers that taught in grades 2-5 in all three years (columns 1 and 3), and for the sample that taught the same grade in all three years (columns 2 and 4). One might expect that teacher effects will be more persistent in the latter case.

The evidence in Table 6 suggests that there is a large persistent component in the teacher residuals. There is a strong correlation in teacher performance residuals across years, with the correlation appearing to be slightly stronger at a one-year lag than at a two-year lag. For example, in column 1 the correlation is 0.57 at one lag, and remains 0.52 at two lags. In the remaining columns the pattern is similar, although the correlations are somewhat stronger for teachers teaching in the same grade in each year, and somewhat weaker after removing school-grade-track-year effects. The small decline in correlation at the second lag implies that the persistent component changes over time, leading to a stronger correlation of performance in adjacent years.

The correlation between changes in the teacher performance residual in 2002 and the level in 2000 is slightly negative, ranging from –0.03 to –0.06. Thus, there is some evidence of mean reversion in the persistent component: teachers with a high residual in 2000 tended to decline slightly from 2001 to 2002. The correlation of changes with lagged changes is also slightly above what it should be if the persistent component was fixed (-0.44 to –0.46 rather than –0.50).

Taken together, this evidence suggests that the persistent component of teacher performance changes slightly over time and is not purely a fixed effect. One reason for this may be that teachers accumulate knowledge about teaching over time, and their classroom performance reflects changes in this accumulated knowledge. If this were the case, we might expect the persistent component to change most for newer teachers who are still learning, but change little for more experienced teachers. The bottom two panels of Table 6 split the sample into teachers with 1-5 years of experience and teachers with 6+ years of experience in 2002. In fact, the persistent component for more experienced teachers looks more like it is fixed: there is less of a decline in the correlation of the teacher residual between one and two lags, and the correlations of changes with lagged changes is closer to –0.5. In contrast, among less experienced teachers there is a more notable decline in the correlation at the second lag and a less negative correlation in adjacent changes in teacher effectiveness.

Taken together, this evidence suggests that there is some change over time in the persistent component, particularly early in a teacher's career. Nevertheless, assuming that the persistent component is fixed may be a useful approximation – especially at short time horizons of 2-3 years where the evidence is not particularly at odds with this assumption.

Forecasts of Teacher Value-Added

Because the teacher effect is both large and persistent, estimates of the teacher effect from one or two years may be quite useful in forecasting future performance. In particular, the optimal forecast of each teacher's performance is simply the posterior mean of the teacher's persistent effect, conditional on prior year's performance. This takes a simple form if the persistent effect is fixed over time:

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

In other words, the optimal forecast is a rescaled version of the average teacher performance residual. The scaling factor depends only on the number of years the teacher has been observed, and the ratio of noise (non-persistent) variance to signal (persistent) variance. The estimate from Table 6 that 57% of the total variance was persistent implies that the ratio of noise to signal is .75 – the variance of the non-persistent component is about 75% as large as the persistent component. Thus, the optimal forecast is very easy to compute: with one year of prior data it is (1/1.75)*ε_1, with two years of prior data it is (2/2.75)*(ε_1+ ε_2)/2, etc.

Using this simple formula for the posterior mean, we constructed forecasts for each teacher based on their residuals in the first two years of our data (2000 and 2001), and then ranked all teachers into quartiles – with the first quartile having the worst performance and the fourth quartile the best. For this calculation we used residuals that did not remove school*grade*track*year effects (e.g. column 1-2 of Table 6). We then conducted an out-of-sample forecasting exercise to investigate if prior performance can accurately predict large differences in teacher performance in future years. More specifically, we added the quartile dummies (Q) to a regression of the form:

$$\hat{\phi}_{jt} = Q_j\beta_1 + Z_{jt}\beta_2 + \varepsilon_{jt}$$

where this regression was estimated with the teacher effects from 2002 or 2003. The unit of observation in this regression was a teacher. The regression controlled for dummy variables for the first four years of experience, a full set of grade dummies, a vector of the mean demographic and program participation characteristics for the students in the class, and a set of fixed effects for permutations of school, grade, calendar track and year. Thus, in this regression we estimate whether teachers who are highly ranked based on their 2000 and 2001 performance perform significantly better than other teachers within the same school and grade in 2002 and 2003.

The results of this forecasting exercise are reported in Table 7. The omitted category is the fourth quartile – teachers with the highest value-added. The first column provides the expected difference across the quartiles based on the posterior mean estimates from 2000/2001. Based on the variation in performance from prior years and our estimate of the persistence, we expect that students in classrooms taught by the bottom quartile of teachers will be more than 10 NCE points below students taught by the top quartile of teachers – nearly a half a student-level standard deviation. This is very close to what we observe in 2002, while the effects in 2003 (which are based on a new test) are only somewhat smaller. These differences are an order of magnitude larger than the differences associated with observable characteristics such as experience and credentials. This evidence highlights the potential usefulness of using imperfect information on early career performance to identify effective teachers.

Although the regressions in Table 7 controlled for average student characteristics in each teacher's classroom, it is possible that some of this difference may be the result of student sorting, if students who can make larger test score gains are systematically sorted into the classrooms of better teachers. Table 8 provides some evidence that this may be occurring. In this table, the dependent variable was various average classroom characteristics in 2002. The independent variables included the quartile dummies, experience dummies, and a full set of school-grade-track dummies. The coefficients on the quartile dummies show some evidence that bottom quartile teachers also taught less academically able students: On average, their students had lower baseline test scores, were less likely to be in the gifted and talented program, and were slightly more likely to be African American. However, the differences between a bottom and top quartile teacher were similar to differences observed between a second year teacher and a veteran teacher (note that 1st year teachers are not included in this regression because they do not have prior data on which to be ranked into quartiles). Thus, it seems unlikely that this type of selection would lead to such large differences between top and bottom quartile teachers, yet small differences based on experience. However, the pattern of student selection is interesting in itself, and suggests that principals or parents may take teacher quality into account in student assignment.

VIII. Using Imperfect Information at the Tenure Decision: A Simple Model

In this section, we develop a simple model for using imperfect estimates of teacher effectiveness to screen out ineffective teachers and maximize student achievement. Given our estimates of the measurement error in teachers' impacts on student achievement, we ask where the district would draw the line in screening out ineffective teachers at the tenure decision, if its goal were to use such measures to maximize student achievement. Moreover, we ask when that decision is optimally made—at the end of one year or many years of teaching. The model captures many of the trade-offs implicit in the school district's decision: First, as noted above, the model recognizes the fact that teacher-level estimates of value-added are measured with error, and that accumulating information over several years increases the ratio of signal to noise in such measures. Second, keeping an ineffective teacher for a second year also has a cost in terms of lowered student achievement. Third, given the returns to experience estimated above, laying off a second or third year teacher requires the district to forfeit the benefits of any on-the-job learning teachers do during their first years of teaching. The average novice teacher is less productive than the average second or third year teacher. A district would use information on teacher impacts only if it were sufficiently reliable to ensure that the payoffs from screening out the low-performing teachers exceed the costs in terms of loss of experience.

We impose a number of simplifying assumptions to make the solution more tractable. First, we assume that teacher effectiveness is simply a fixed effect plus the return to experience. As we saw in the previous section, this simple formulation seems to match the empirical evidence on teacher impacts fairly closely. Second, we assume that naturally occurring teacher turnover is exogenous (or, at least, unrelated to student achievement). We presented evidence earlier in the paper that "leavers" were somewhat less effective than "stayers"—although the differences were small (about .05 student level standard deviations). To simplify the model, we assume that those differences are zero. Third, we will assume that the variance of the measurement error is the same for all teachers—that is, that all teachers have the same size class and that they are all subject to the same i.i.d. shocks to classroom level performance. Fourth, we will assume that there is no effect of classroom size on achievement and that there is no benefit to pairing teachers within the same class. We rule out the possibility that the district could raise a teacher's effectiveness by lowering class size. This relieves us of the need to estimate the effect of class size on student achievement. All teachers have the same class size. Fifth, we assume that the district must draw a line only once in teachers' careers-- much like a tenure decision—and apply the same tenure clock and same standard to all teachers. Sixth, we assume that the supply of teachers is exogenous. In particular, we assume that the supply would not be negatively impacted by the uncertain prospect of being able to remain with the district beyond the tenure decision. Finally, we assume zero hiring costs—although we have simulated the effect of loosening that assumption in the discussion of the results below.

The model we describe is a simplified version of the model in Jovanovic (1979). In that model, Jovanovic was considering the case of infinitely lived workers, taking draws from a job distribution—while we are considering the case of a job taking draws from the worker distribution. We simplify by allowing the tenure decision to be made at the same point (T) for all workers. This eliminates the option value of waiting to fire a worker until a future period, which complicates the solution.

Model Setup

Suppose that Y_t represents our estimate of a teacher's impact on student achievement after t periods of teaching experience. Moreover, suppose that the unconditional mean impact on student achievement impact is given by $\exists_t$, the return to experience, which is known (that is $E(Y_t) = \beta_t$). Let : measure a teacher's fixed effect, and suppose that it is normally distributed with mean zero and variance σ_μ^2 (that is, $\mu \sim N(0, \sigma_\mu^2)$). And suppose that Y_t is an imperfect measure of a teacher's impact, subject to the measurement error, „$_t$:

$$Y_t = \mu + B_t + \varepsilon_t$$

We will also assume that the measurement error, ε_t, is normally distributed with mean 0 and variance σ_ε^2, so that the distribution of Y conditional on : is $N(\beta_t + \mu, \sigma_\varepsilon^2)$. Let V_t represent the experience-adjusted estimate of a teacher's impact on student achievement ($V_t = Y_t - \beta_t$). Under the assumptions above, the expected value of a given teacher's fixed effect after T periods of imperfectly measured impacts on student achievement can be expressed as below:

$$E(\mu \mid Y_1, \ldots, Y_T) = \overline{V}_T \frac{T}{T + \sigma_\varepsilon^2 / \sigma_\mu^2}$$

where $\overline{V}_T$ is just the mean of the experience-adjusted estimates of a teacher's impact after T periods ($\overline{V}_T = \frac{1}{T}\sum_{t=1}^{T} V_T$). The term multiplying $\overline{V}_T$ is a shrinkage factor, capturing the proportion of the total variance in $\overline{V}_T$ that is signal variance. As T grows large $\overline{V}_T$

becomes a more reliable measure of μ and the shrinkage factor goes to one. The unconditional distribution of $\overline{V}_T$ will be normal, with a mean of zero and a variance of $\sigma_T^2 = \sigma_\mu^2 + \sigma_\varepsilon^2 / T$. In other words, with each additional period of estimates, then variance of the distribution of $\overline{V}_T$ converges toward the signal variance, σ_μ^2.

Moreover, given our assumption that a teacher's impact on student performance is just equal to a teacher fixed effect plus the return to experience, then our expected value of a teacher's estimated performance in any future period will bee:

$$E\left(Y_{T+k} \mid Y_1,...,Y_T\right) = \beta_{T+k} + \overline{V}_T \frac{T}{T + \sigma_\varepsilon^2 / \sigma_\mu^2} \text{ for all k>0.}$$

This setup implies that if we were able to identify some cut-off, c, above which teachers were offered tenure after T periods (and below which teachers' contracts were not renewed), then the expected impact on student achievement of tenured teachers would be equal to $E\left(Y_{T+k} \mid \overline{V}_T > c\right) = \beta_{T+k} + \frac{T}{T + \sigma_\varepsilon^2 / \sigma_\mu^2} E\left(\overline{V}_t \mid \overline{V}_t > c\right)$, where

$E\left(\overline{V}_t \mid \overline{V}_t > c\right) = \sigma_T \frac{\phi\left(-c/\sigma_T\right)}{\Phi\left(-c/\sigma_T\right)}$, and the probability that someone would make tenure could be described as $\Pr(\overline{V}_T > c) = \Phi\left(-c/\sigma_T\right)$. (In the above expressions, φ() and Φ() represent the normal probability density function and cumulative distribution function respectively.)

The District's Optimization Problem

The district's problem is to choose a cut-off for estimated teacher performance, c, and a tenure clock, T, to maximize the average productivity of its entire workforce, $\overline{Y}$. The workforce consists of two groups of workers: those who are pre-tenure, whose expected performance is just $E(Y_t)$ and workers who survived the tenure cut-off, whose expected performance would be $E(Y_t \mid \overline{V}_t > c)$. Therefore, the productivity of the workforce will be equal to:

$$(1)\ \overline{Y} = \sum_{t=1}^{T} \pi_t E(Y_t) + \sum_{t=T+1}^{N} \pi_t E\left(Y_t \mid \overline{V}_t > c\right)$$

where π_t is the proportion of teachers in the workforce with experience t.

If * is the proportion of teachers exogenously choosing to remain with the district after each period, then we can derive the B_t for the periods before and after the tenure cut in period T as follows:

$$\pi_t = \frac{\delta^{t-1}}{D} \; for\, t \le T,\, and$$
$$\pi_t = \frac{\delta^{t-1} pr(\overline{V}_T > c)}{D} \; for\, t > T,\, and$$
$$D = \sum_{t=1}^{T} \delta^{t-1} + \sum_{t=T+1}^{N} \delta^{t-1} pr(\overline{V}_T > c)$$

Differentiating the expression (1) above with respect to c and setting it equal to zero, the first order condition which determines a cut-off, c, for a given T, would imply the following:

$$\overline{Y} = c\left(\frac{T}{T + \sigma_{\varepsilon}^{2} / \sigma_{\mu}^{2}} \right) + \left(\frac{\sum_{t=T+1}^{N} \delta^{t-1} \beta_t}{\sum_{t=T+1}^{N} \delta^{t-1}} \right)$$

The above expression has a fairly straightforward interpretation. The expression on the left is the average productivity of the workforce. The expression on the right is the productivity of the *marginal* teacher, with estimated performance ($\overline{V}_T$) at the cut-off, c. The latter consists of two parts. The first part is the expected fixed effect of the marginal teacher; while the second part is their expected return to experience over the remainder of their career, if they were not fired. So, in other words, the district should set the cut-off, c, where the productivity of the *marginal* teacher is equal to the *average* teacher.

Imagine if this were not true. That is, suppose the marginal teacher were less productive than the average teacher. The district could raise performance, by raising its standard by a small amount. Likewise, if the marginal teacher were more productive than the average teacher, then the district could raise average performance more by lowering the cut-off and adding one more above-average teacher. This result is analogous to the usual result that average costs are minimized at the point where marginal cost equals average cost.

The above first order condition has a number of implications for the determinants of the cut-off level of performance required for tenure. First, as the signal variance increases, the district should set a higher threshold. Second, the threshold level of performance for tenure declines with the return to experience. The greater the loss in productivity for replacing an experienced teacher with a novice teacher, the higher the proportion of experienced teachers the district would want to retain. Third, the cut-off is decreasing in the exogenous turnover rate. The higher the proportion of teachers who exogenously decide to stay in the district each year, the greater the cost to keeping a low-performing teacher and the higher the threshold one would want to set.

Simulations

We use our estimates of the relevant parameters, to simulate average productivity, with varying cut-offs and varying tenure clocks. From Table 6, we set the estimated proportion of the variance in estimated teacher effects that is signal to be .57. This implies a signal variance (σ_{μ}^{2}) of 27 (.57 times a variance in estimated teacher effects of 47.3). We assume an exogenous turnover rate of 5 percent, which is approximately the proportion of experienced teachers who leave the district each year. We use the returns to experience estimated in column (3) of Table 4. We also assume a maximum teaching career of 30 years.

Figure 3 reports the average productivity of the teaching force on the vertical axis by the proportion of teachers not retained at the tenure decision on the horizontal axis. The different curves in Figure 3 correspond to different tenure clock lengths, ranging from a tenure decision after one year (the top curve) to a tenure decision after five years (the bottom curve). Figure 3 has several implications worth noting. First, there are fairly large impacts on average teacher productivity to using the information on teaching performance in the first few years of teaching as part of the tenure decision. For example, if the district were to retain only the top fifth of teachers at the end of their first year of teaching, the district could raise average achievement by nearly 4 points—one fifth of a student level standard deviation. Second, the simulation implies that the district should offer tenure to a fairly low proportion of all teachers. For example, if the tenure decision is being made at the end of the first year, only about one-fifth of the highest scoring teachers should be retained. This low tenure rate reflects the low expected turnover rate post-tenure and the modestly high signal-to-noise ratio even in the first year of teaching. Third, rather than wait until years two through five, the district is better off making the tenure decision at the end of only one year. This latter result is perhaps a bit surprising. But the costs of retaining low-performing teachers is apparently sufficiently high to offset the additional information the district would obtain by waiting to see a second year of performance data.

The returns to experience generate some interesting equity issues. Those students who are assigned novice teachers can expect lower achievement gains than before. If the tenure decision is used aggressively by districts and only the highest performing teachers are retained, this gap in performance between more and less experienced teachers will be accentuated, with a difference of about a third of a standard deviation in student performance between tenured teachers and novice teachers. So there would be very high stakes attached to being assigned a novice teacher. However, the overall variance in teacher effectiveness would be about the same in the new equilibrium, since only the highest performing teachers would be tenured which reduces variation among the tenured teachers.

Figure 4 reports the proportion of teachers who would be novice teachers with different tenure rates, with a tenure clock of one year. At the optimal tenure rate of approximately 20 percent of teachers retained, about one quarter of the teaching force in the new steady state would be novice teachers. Currently, only about 9 percent of the teaching force in LAUSD is made up of novice teachers. This implies that the district would have to roughly triple its hiring of new teachers.

Discussion

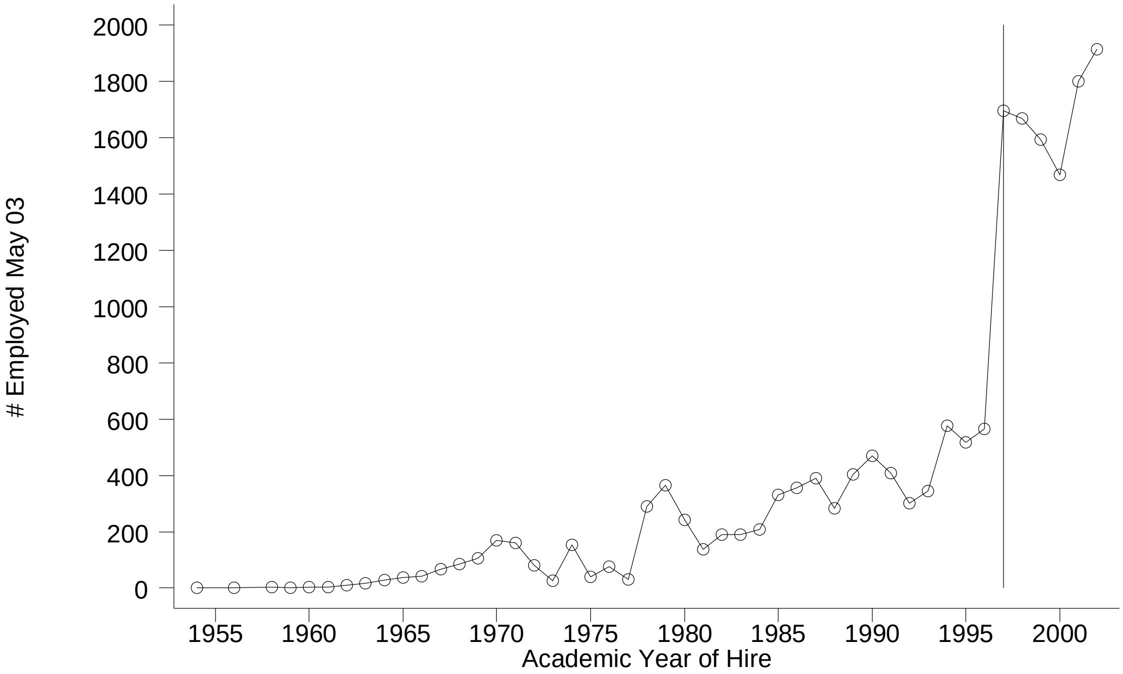
One of the assumptions in the model above was that there were zero costs to hiring (or firing) teachers before the tenure decision. That is likely to be unrealistic. However, it would be relatively straightforward to incorporate hiring costs into the model, by simply raising the payoff to experience. The difficulty comes in translating the dollar costs of hiring to test score units. We simulated the implication of using a hiring cost of 10 points. Such a cost would be quite large in dollar terms. In Kane and Staiger (2002), we argued that one standard deviation in test performance is worth nearly $100,000 in lifetime earnings. So 10 points would be worth roughly $50,000 per student – or roughly $1 million per teacher! Nevertheless, even with a hiring cost of 10 points, the optimal tenure rate would remain at only 45 percent (with the district not renewing the contracts of 55 percent of first-year teachers). In such a regime, about 20 percent of teachers would be novice teachers.

IX. Conclusion

Our results suggest a very different approach to raising the quality of the teaching force. Most school districts attempt to screen out ineffective teachers at the point of hiring-- and then do little to screen out ineffective teachers afterwards. This strategy is consistent with an education process that requires teachers to make large specific investments prior to becoming teachers. Our evidence suggests that there may be little point to screening teachers at hiring, since there is little difference in the effectiveness of the marginal and average teacher hired. Moreover, although there is evidence that those who subsequently left LAUSD were less effective than those who remained, the difference in achievement impacts between those leaving and those staying was small. Rather, our evidence suggests that one could identify much larger differences between the marginal and average teacher by observing a single year of teaching performance and retaining only the highest-scoring teachers. Despite the fact that our estimates of teacher performance are fairly noisy, they can still be used aggressively to identify effective teachers and increase the overall quality of teaching. This approach is consistent with an initial process of hiring that is not selective – and in particular does not require teachers to make specific education investments prior to being hired. Given the small chance of being retained, any specific investment being made by teachers must wait until after they have been identified for retention.

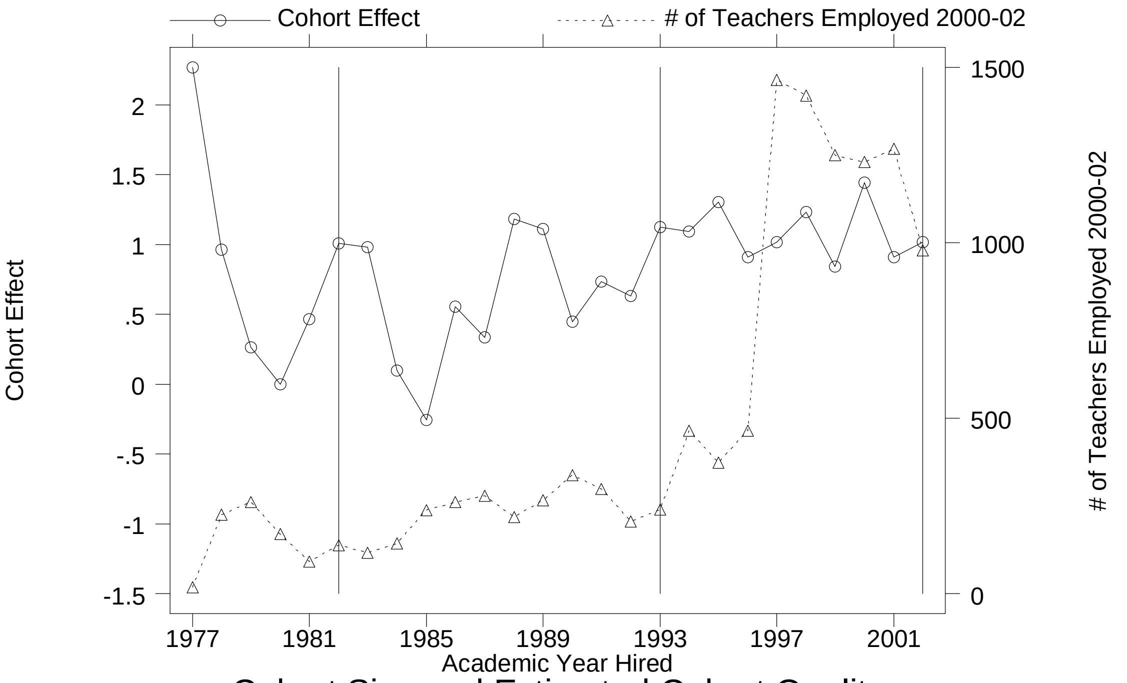
Of course, some districts will be uncomfortable using the purely statistical approach we outline to determine retention of new teachers. Other measures of teacher performance may be available, such as parent or principal evaluations, in classroom observation, and the like. Our approach can be readily modified to incorporate other performance measures to form an overall index as discussed in Kane and Staiger (2001) and McClellan and Staiger (1999). While the details may be somewhat more complicated, it would be unlikely that the general message would change – that such a measure should be used to aggressively identify and retain only the best teachers early in their career.

Figure 1.



Hiring Dates of Elementary Teachers Employed May 2003

Figure 2.



Cohort Size and Estimated Cohort Quality

Figure 3.

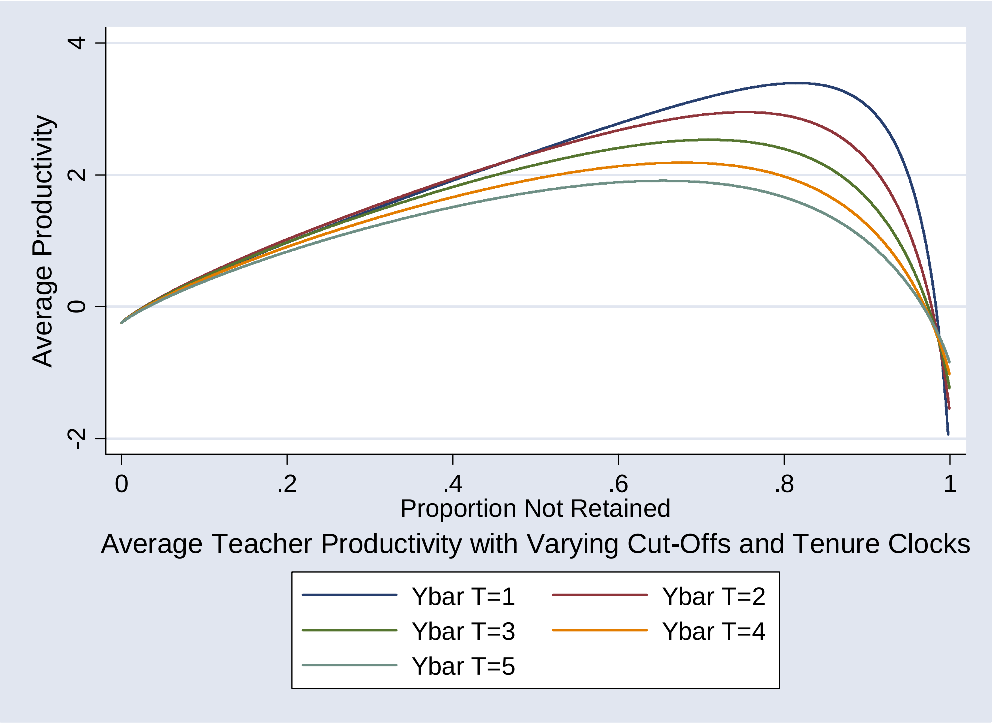
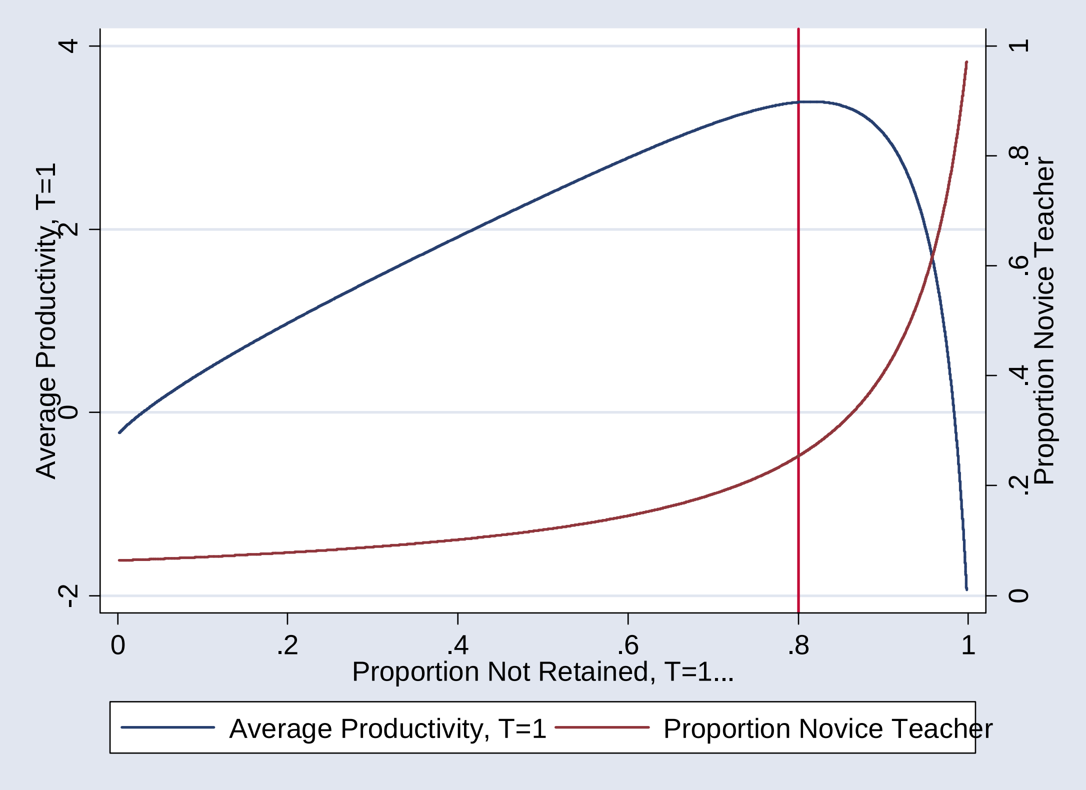

Average Productivity
-2
0
2
4
0
.2
.4
.6
.8
1
Proportion Not Retained
Average Teacher Productivity with Varying Cut-Offs and Tenure Clocks
Ybar T=1
Ybar T=2
Ybar T=3
Ybar T=4
Ybar T=5

Figure 4.


Average Productivity, T=1
4
2
0
-2
Proportion Novice Teacher
1
.8
.6
.4
.2
0
0
.2
.4
.6
.8
1
Proportion Not Retained, T=1...
Average Productivity, T=1
Proportion Novice Teacher

References:

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

Ballou, Dale. "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics* (1996) Vol. 111, No. 1, pp. 97-133.

Ballou, Dale and Michael Podgursky. "Reforming teacher preparation and licensing: What is the evidence?" *Teachers College Record* (2000) Vol. 102 , No. 1, pp. 1-27.

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

Darling-Hammond, Linda, Deborah J. Holtzman, Su Jin Gatlin and Julian Vasquez Heilig, "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, "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 "Do School and Teacher Characteristics Matter?: Evidence from High School and Beyond" *Economics of Education Review* (1994), Vol. 13, No. 1, pp. 1-17.

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

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

Jovanovic, Boyan "Job Matching and the Theory of Turnover" *Journal of Political Economy* (1979), Vol. 87, No. 5, pp. 972-990.

Rockoff, Jonah. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data" Harvard University Working Paper, March 2004.

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

Hanushek, Eric, John F. Kain , Daniel O'Brien and Steven Rivkin, "Are Better Teachers More Likely to Exit Large Urban Districts?" Stanford Univerisity Working Paper, April 2004.

Hanushek, Eric, John Kain, Daniel O'Brien and Steven Rivkin, "The Market for Teacher Quality" Unpublished Paper, Stanford University, January 2005.

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

Podgursky, Michael, Ryan Monroe, Donald Watson, "The Academic Quality of Public School Teachers: An Analysis of Entry and Exit Behavior" *Economics of Education Review* (2004) Vol. 23, pp. 507-518.

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

Sanders, William, Arnold Saxton, and Sandra Horn "The Tennessee Value-Added Assessment System: A Quantitative, Outcomes-Based Approach to Educational Assessment" in *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (Thousand Oaks, CA: Corwin Press, 1997).

Sanders, William and Sandra Horn "The Tennessee Value-Added Assessment System: Mixed-Model Methodology in Educational Assessment" *Journal of Personnel Evaluation in Education* (1994), pp. 299-311.

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

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

Table 1. Student Achievement and Teacher Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
Outcome	Math NCE	Math NCE	Math NCE	Gain Math	Math NCE	Gain Math
Educational Degree: (Relative to Bachelor's Degee Holders)						
Masters	0.662	0.508	-0.015	-0.047	-0.011	-0.281
	(0.205)	(0.190)	(0.116)	(0.122)	(0.139)	(0.227)
Doctorate	-0.853	-0.581	-0.134	-0.156	-0.572	-1.076
	(0.730)	(0.759)	(0.474)	(0.501)	(0.524)	(0.895)
Initial Status for Those Hired Since 1994: (Relative to Traditionally Certified Novices)						
Intern	-4.552	-0.433	-0.091	-0.079	-0.135	-0.207
	(0.379)	(0.328)	(0.208)	(0.221)	(0.242)	(0.403)
Emerg Cred	-6.932	-1.540	0.300	0.245	0.328	0.472
	(0.338)	(0.289)	(0.186)	(0.196)	(0.219)	(0.366)
Other	-4.492	-0.880	0.420	0.292	0.492	0.798
	(0.381)	(0.333)	(0.210)	(0.222)	(0.247)	(0.418)
Experience for Those Hired Since 1994:						
2nd Year	2.778	2.359	1.396	1.372	1.192	1.978
	(0.303)	(0.304)	(0.198)	(0.209)	(0.226)	(0.366)
3rd Year	4.435	3.859	2.047	1.843	1.660	2.103
	(0.306)	(0.319)	(0.208)	(0.218)	(0.241)	(0.389)
4th Year	5.697	4.666	2.110	1.909	1.976	2.653
	(0.317)	(0.319)	(0.207)	(0.218)	(0.240)	(0.396)
5th to 9th Yr	6.344	4.878	2.253	1.974	1.917	2.778
	(0.286)	(0.298)	(0.190)	(0.199)	(0.220)	(0.356)
Veteran (Hired	3.442	5.091	2.090	1.789	2.006	2.715
Before 1994)	(0.383)	(0.345)	(0.221)	(0.232)	(0.253)	(0.423)
Baseline Scores			X			
Student & Peer Characteristics			X	X	X	X
Fixed Effects?	Year*Grade	School*Gr* Track*Year	School*Gr* Track*Year	School*Gr* Track*Year	Stud*School and Year*Grade	Stud*School and Year*Grade
Sample Size	481411	481411	481411	481411	481411	481411
R^2	0.05	0.28	0.70	0.18	0.92	0.53

Note: Table was estimated using student-level data for those in grades 2-5 in spring 2000-2002. Baseline test scores included math, reading and language arts scores (and each interacted with grade level). Student characteristics included indicators for gender, six racial/ethnic categories, free/reduced price lunch status, english language develoment level (five levels), an indicator for those repeating the current grade and an indicator for those in the gifted and talented program (and each of the above interacted with grade level). The peer characteristics included the classroom level means of the student level characteristics (each interacted with grade level). Standard errors were calculated allowing for clustering at the school/grade/track level.

Table 2. Differences Between 1996 and 1997 Cohorts

	Hired in:			
Teacher Characteristics:	1996	1997	Difference:	P-value:
Still Employed May 2000	0.698	0.753	0.055	0.002
Still Employed May 2001	0.669	0.697	0.028	0.071
Still Employed May 2002	0.637	0.663	0.026	0.092
Graduate Degree	0.212	0.202	-0.011	0.422
Trad Certified (No Prior Exp)	0.198	0.129	-0.069	0.000
Intern or Emerg Credential	0.589	0.722	0.133	0.000
Teacher White	0.414	0.407	-0.006	0.691
Teacher African American	0.094	0.148	0.054	0.000
Teacher Latino	0.399	0.345	-0.054	0.001
Teacher Asian	0.082	0.091	0.008	0.351
N: (Teachers)	1277	3297		
Classroom Characteristics in 2000-2002:				
Number of students in classroom	19.691	19.516	-0.176	0.000
Percent African American	0.095	0.111	0.016	0.000
Percent White, Non-Hispanic	0.116	0.109	-0.007	0.014
Percent Latino	0.732	0.713	-0.020	0.000
Engl Lang Dev 1-2	0.121	0.094	-0.027	0.000
Grade Level	3.389	3.414	0.024	0.016
Baseline Math NCE	47.563	47.675	0.112	0.554
Baseline Reading NCE	42.456	42.830	0.373	0.040
Baseline Language NCE	44.741	45.239	0.498	0.009
Percent in Gifted and Talented	0.079	0.071	-0.008	0.001
Percent in Free/Reduced Price Lur	0.830	0.844	0.014	0.000
N: (Students)	15615	51614		

Table 3. Comparing Student Achievement for the Teacher Cohorts Hired in 1996 and 1997

	(1)	(2)	(3)	(4)	(5)	(6)
Dep Var:	Math NCE	Math NCE	Math NCE	Gain Math	Math NCE	Gain Math
Relative to 1996 Cohort:						
Cohort 1997	0.439	1.109	0.223	0.032	-0.513	-0.639
	(0.524)	(0.841)	(0.537)	(0.549)	(1.402)	(2.294)
Sample Size	67933	67933	67933	67933	67933	67933
R^2	0.03	0.38	0.74	0.29	0.98	0.90

	(7)	(8)	(9)	(10)	(11)	(12)
Dep Var:	Read NCE	Read NCE	Read NCE	Gain Read	Read NCE	Gain Read
Relative to 1996 Cohort:						
Cohort 1997	0.634	1.526	0.372	0.386	0.760	0.684
	(0.508)	(0.700)	(0.365)	(0.399)	(0.952)	(1.445)
Sample Size	67229	67229	67229	67229	67229	67229
R^2	0.03	0.42	0.78	0.23	0.99	0.91
Baseline Scores			X			
Student & Peer Characteristics			X	X	X	X
Fixed Effects'	Year*Grade	School*Gr* Track*Year	School*Gr* Track*Year	School*Gr* Track*Year	Stud*School and Year*Grade	Stud*School and Year*Grade

Note: Table was estimated with student-level data for those in grades 2-5 with scores in spring 2000-2002 who were assigned to a teacher hired in 1996 or 1997. Baseline test scores included math, reading and language arts scores (and each interacted with grade level). Student characteristics included indicators for gender, six racial/ethnic categories, free/reduced price lunch status, english language develoment level (five levels), an indicator for those repeating the current grade and an indicator for those in the gifted and talented program (and each of the above interacted with grade level). The peer characteristics included the classroom level means of the student level characteristics (each interacted with grade level). Standard errors were calculated allowing for clustering at the school/grade/track level.

Table 4. Selection and the Payoff to Experience

	(1)	(2)	(3)	(4)	(5)	(6)
						Veterans
Sample:	All	All	All	2001-02	2001-02	2001-02
Incremental Effect Per Year of Experience (Using Linear Spline):						
1st Year	1.396	1.481	1.511	1.630	1.624	
	(0.198)	(0.144)	(0.144)	(0.240)	(0.240)	
2nd Year	0.654	0.332	0.348	0.190	0.199	
	(0.197)	(0.139)	(0.140)	(0.239)	(0.239)	
3rd Year	0.060	0.184	0.174	0.191	0.185	
	(0.202)	(0.137)	(0.139)	(0.240)	(0.240)	
4th Year	0.332	0.329	0.324	0.324	0.326	
	(0.213)	(0.135)	(0.136)	(0.277)	(0.277)	
5th to 9th Year	-0.183	-0.031	0.010	-0.251	-0.252	
	(0.095)	(0.109)	(0.110)	(0.148)	(0.148)	
Veteran	2.075	---	---	2.197	2.197	
	(0.221)			(0.282)	(0.282)	
Leaving LAUSD				-0.792	-0.958	-1.065
(Not employed 5/03)				(0.153)	(0.188)	(0.428)
Leaving t+2					0.420	0.189
					(0.262)	(0.579)
Baseline Scores	X	X	X	X	X	X
Student & Peer Characteristics	X	X	X	X	X	X
Fixed Effects?	School*Gr* Track*Year	Teacher	Teacher* School	School*Gr* Track*Year	School*Gr* Track*Year	School*Gr* Track*Year
Sample Size	481411	481411	481411	308873	308873	126731
R^2	0.70	0.73	0.73	0.70	0.70	0.73

Note: Table was estimated using student-level data for those in grades 2-5 in spring 2000-2002. Columns (1) and (2-4) also included dummies for teacher degree and initial hiring status (interns, emgergency credentials and other). Baseline test scores included math, reading and language arts scores (and each interacted with grade level). Student characteristics included indicators for gender, six racial/ethnic categories, free/reduced price lunch status, english language develoment level (five levels), an indicator for those repeating the current grade and an indicator for those in the gifted and talented program (and each of the above interacted with grade level). The peer characteristics included the classroom level means of the student level characteristics (each interacted with grade level). Standard errors were calculated allowing for clustering at the school/grade/track level.

Table 5. Interactions between Payoff to Experience and Initial Hiring Status

		Interacted with:		
		Emergency		
Sample:		Credential	Intern	Other
Incremental Effect Per Year of Experience (Using Linear Spline):				
1st Year	1.329	0.635	-0.244	-0.246
	(0.333)	(0.384)	(0.431)	(0.440)
2nd Year	-0.575	0.855	1.212	1.242
	(0.397)	(0.440)	(0.470)	(0.498)
3rd Year	0.455	-0.278	0.033	-0.851
	(0.385)	(0.425)	(0.462)	(0.478)
4th Year	0.163	0.225	0.063	0.185
	(0.365)	(0.406)	(0.445)	(0.471)
5th to 9th Year	0.127	-0.151	0.071	-0.303
	(0.240)	(0.279)	(0.333)	(0.318)
p-value of H_o				
Interactions=0		0.1538	0.2155	0.1213
		N=481411, R^2=.73		

Note: Table was estimated using student-level data for those in grades 2-5 in spring 2000-2002. The specification also included teacher by school fixed effects, grade and year dummies, student baseline test scores and demographics and mean baseline scores and demographics for the classroom. Baseline test scores included math, reading and language arts scores (and each interacted with grade level). Student characteristics included indicators for gender, six racial/ethnic categories, free/reduced price lunch status, english language develoment level (five levels), an indicator for those repeating the current grade and an indicator for those in the gifted and talented program (and each of the above interacted with grade level). The peer characteristics included the classroom level means of the student level characteristics (each interacted with grade level). Standard errors were calculated allowing for clustering at the school/grade/track level.

Table 6. Correlation of Teacher Effects Over Time

	Not Removing Effect of School*Gr*Track*Year		Removing Effect of School*Gr*Track*Year	
Grades taught:	Any Grade	Same Grade	Any Grade	Same Grade
	A. All Teachers			
Correlation of Teacher FE from 2002 with:				
Teacher FE from 2001	0.57	0.61	0.48	0.54
Teacher FE from 2000	0.52	0.57	0.42	0.47
Correlation of Δ Teacher FE from 2002 with:				
Teacher FE from 2000	-0.03	-0.03	-0.05	-0.06
Δ Teacher FE from 2001	-0.45	-0.46	-0.44	-0.44
# teachers	5370	3630	5370	3630
	B. Teachers with <6 years experience in 2002			
Correlation of Teacher FE from 2002 with:				
Teacher FE from 2001	0.57	0.63	0.48	0.56
Teacher FE from 2000	0.46	0.52	0.38	0.44
Correlation of Δ Teacher FE from 2002 with:				
Teacher FE from 2000	-0.01	-0.04	-0.03	-0.05
Δ Teacher FE from 2001	-0.44	-0.42	-0.44	-0.42
# teachers	1728	993	1728	993
	C. Teachers with >5 years experience in 2002			
Correlation of Teacher FE from 2002 with:				
Teacher FE from 2001	0.57	0.60	0.48	0.53
Teacher FE from 2000	0.55	0.58	0.43	0.49
Correlation of Δ Teacher FE from 2002 with:				
Teacher FE from 2000	-0.04	-0.03	-0.06	-0.06
Δ Teacher FE from 2001	-0.46	-0.47	-0.44	-0.45
# teachers	3642	2637	3642	2637

Note: Correlations in the first two columns are based on residuals after teacher value-added fixed effects were regressed on dummy variables for the first four years of experience, a full set of grade dummies for each year, and a vector of the mean demographic and program participation characteristics for the students in the class. The second two columns also removed a set of fixed effects for permutations of school, grade, calendar track and year.

Table 7. Using Posterior Mean Estimated From 2000 & 2001

	Expected	2002		2003	
	Difference	all	same grade	all	same grade
Quartile of posterior mean (relative to best quartile)					
1st quartile (lowest)	*-10.924*	-9.678	-10.540	-8.165	-8.643
		(0.343)	(0.517)	(0.345)	(0.541)
2nd quartile	*-7.318*	-6.749	-7.296	-5.312	-6.006
		(0.319)	(0.496)	(0.291)	(0.442)
3rd quartile	*-4.721*	-4.038	-4.677	-3.200	-3.669
		(0.334)	(0.462)	(0.304)	(0.478)
Observations		7483	5105	8126	5090

Note: Quartiles were based on posterior mean calculated from each teacher's fixed-effects estimated from the 2000 and 2001 school years. The first column reports the average posterior mean in each quartile, which is the predicted teacher effect. The remaining columns report coefficients on quartile dummy variables from a regression in which the dependent variable was the teacher's fixed effect in 2002 or 2003. These regressions also included dummy variables for the first four years of experience, a full set of grade dummies for each year, a vector of the mean demographic and program participation characteristics for the students in the class, and a set of fixed effects for permutations of school, grade, calendar track and year. Standard errors were clustered at the school level.

Table 8. Predicting Average Classroom Characteristics in 2002 Using Posterior Mean Estimated From 2000 & 2001

dependent variable:	Average Math Score	Average Reading Score	Average Language Score	Proportion % Gifted & Talented	Proportion Male	Proportion Black	Proportion Latino	Proportion Eligible for Meal Program
Quartile of posterior mean (relative to best quartile)								
1st quartile (lowest)	-2.247	-1.837	-1.980	-0.036	0.002	0.020	-0.002	0.010
	(0.546)	(0.554)	(0.569)	(0.008)	(0.006)	(0.007)	(0.010)	(0.005)
2nd quartile	-1.299	-0.695	-0.936	-0.023	-0.002	0.015	-0.004	0.007
	(0.521)	(0.532)	(0.538)	(0.008)	(0.005)	(0.006)	(0.010)	(0.005)
3rd quartile	-0.875	-0.499	-0.589	-0.010	-0.002	0.011	0.000	0.006
	(0.532)	(0.542)	(0.535)	(0.009)	(0.005)	(0.006)	(0.010)	(0.005)
Experience								
2nd year	-2.106	-1.790	-2.067	-0.034	0.001	0.004	0.003	0.005
	(0.616)	(0.605)	(0.628)	(0.011)	(0.006)	(0.009)	(0.012)	(0.006)
3rd year	-1.765	-1.849	-1.940	-0.046	0.010	0.002	0.015	0.010
	(0.611)	(0.636)	(0.645)	(0.011)	(0.007)	(0.006)	(0.010)	(0.007)
4th year	-1.296	-1.273	-1.117	-0.022	-0.001	0.004	0.018	0.015
	(0.704)	(0.682)	(0.710)	(0.011)	(0.007)	(0.009)	(0.012)	(0.007)
Observations	7483	7483	7483	7483	7483	7483	7483	7483

Note: Quartiles were based on posterior mean calculated from each teacher's fixed-effects estimated from the 2000 and 2001 school years. The dependent variable, listed at the top of each column, was an average student characteristic in the teacher's class in 2002. Each regression also included a full set of grade dummies for each year and a set of fixed effects for permutations of school, grade, calendar track and year. Standard errors were clustered at the school level.