THE HAMILTON PROJECT

Advancing Opportunity, Prosperity and Growth

DISCUSSION PAPER 2006-01 APRIL 2006

Robert Gordon
Thomas J. Kane
Douglas O. Staiger

Identifying Effective Teachers Using Performance on the Job

THE BROOKINGS INSTITUTION

The Hamilton Project seeks to advance America's promise of opportunity, prosperity, and growth. The Project's economic strategy reflects a judgment that long-term prosperity is best achieved by making economic growth broad-based, by enhancing individual economic security, and by embracing a role for effective government in making needed public investments. Our strategy—strikingly different from the theories driving current economic policy—calls for fiscal discipline and for increased public investment in key growth-enhancing areas. The Project will put forward innovative policy ideas from leading economic thinkers throughout the United States—ideas based on experience and evidence, not ideology and doctrine—to introduce new, sometimes controversial, policy options into the national debate with the goal of improving our country's economic policy.

The Project is named after Alexander Hamilton, the nation's first treasury secretary, who laid the foundation for the modern American economy. Consistent with the guiding principles of the Project, Hamilton stood for sound fiscal policy, believed that broad-based opportunity for advancement would drive American economic growth, and recognized that "prudent aids and encouragements on the part of government" are necessary to enhance and guide market forces.

THE HAMILTON PROJECT

THE HAMILTON PROJECT

Identifying Effective Teachers Using Performance on the Job

Robert Gordon
Center for American Progress

Thomas J. Kane
Harvard Graduate School of Education

Douglas O. Staiger
Dartmouth College

The Brookings Institution
APRIL 2006

Abstract

Traditionally, policymakers have attempted to improve the quality of the teaching force by raising minimum credentials for entering teachers. Recent research, however, suggests that such paper qualifications have little predictive power in identifying effective teachers. We propose federal support to help states measure the effectiveness of individual teachers—based on their impact on student achievement, subjective evaluations by principals and peers, and parental evaluations. States would be given considerable discretion to develop their own measures, as long as student achievement impacts (using so-called "value-added" measures) are a key component. The federal government would pay for bonuses to highly rated teachers willing to teach in high-poverty schools. In return for federal support, schools would not be able to offer tenure to new teachers who receive poor evaluations during their first two years on the job without obtaining district approval and informing parents in the schools. States would open further the door to teaching for those who lack traditional certification but can demonstrate success on the job. This approach would facilitate entry into teaching by those pursuing other careers. The new measures of teacher performance would also provide key data for teachers and schools to use in their efforts to improve their performance.

Contents

I. Introduction 5

II. Recent Evidence on Teacher Quality 7

III. Recommendations 10

IV. Implementation and Costs of Our Five Recommendations 24

V. Questions and Concerns 26

VI. Conclusion 30

Technical Appendix 31

References 33

I. Introduction

Over the last two decades, policymakers have fretted over the quality of elementary and secondary education in the United States. Worried that the public education system has become a constraint on future productivity growth and a root cause of income inequality, leaders have championed a succession of reforms—from test-based accountability to smaller class sizes. But, ultimately, the success of U.S. public education depends upon the skills of the 3.1 million teachers managing classrooms in elementary and secondary schools around the country. Everything else—educational standards, testing, class size, greater accountability—is background, intended to support the crucial interactions between teachers and their students. Without the right people standing in front of the classroom, school reform is a futile exercise.

Traditionally, policymakers have attempted to raise the quality of the teaching force by raising the hurdles for those seeking to enter the profession. For instance, the federal No Child Left Behind Act (NCLBA) requires all teachers of the core academic subjects to be "highly qualified"—with a minimum of a bachelor's degree, full state licensure and certification (generally requiring that teachers graduate from a teacher education program), and demonstrated subject-area competence (through completing academic coursework or passing a standardized test).

Once teachers are hired, however, school districts typically do very little additional screening. Tenure is awarded as a matter of course after two or three years of teaching. Very few teachers are involuntarily discharged from a school or school district. And the very best teachers receive no financial incentives to go where they are needed most.

The current credential-centered regime is built upon two questionable premises. The first premise is that the paper qualifications required for certification (passage of a standardized test and completion of a specified set of courses) are strongly related to a teacher's effectiveness. The second premise is that school districts learn nothing more about teachers' effectiveness after the initial hire.

A growing body of research, however, suggests that neither of these premises is valid. According to recent evidence, certification of teachers bears little relationship to teacher effectiveness (measured by impacts on student achievement). There are effective certified teachers and there are ineffective certified teachers; similarly, there are effective uncertified teachers and ineffective uncertified teachers. The differences between the stronger teachers and the weaker teachers only become clear once teachers have been in the classroom for a couple of years.

In response to this evidence, our proposal aims to improve average teacher effectiveness by increasing the inflow of new teachers and requiring minimum demonstrated competency on the job (rather than relying solely on screens at the point of hiring). It also aims to alter the *distribution* of high-performing teachers by encouraging more of the most effective teachers to work in high-poverty schools. Moreover, by removing barriers to entering the teaching profession, our proposal would enable many people interested in pursuing teaching as a second career (or as one of several careers) to become teachers. This is particularly important at a time when our nation faces a looming teacher shortage because a large share of our nation's teachers are nearing retirement.

These policies require consistent and reliable measurement of teacher performance. States and districts will need funding and technical support to build the requisite data infrastructure if these policies are to succeed. This infrastructure will not only make decisions about tenure and pay easier, but will also help identify which teachers need help, which teachers are succeeding and should serve as mentors to others, and which teaching approaches are proving most effective.

We make five specific recommendations:

Recommendation 1: Reduce the barriers to entry into teaching for those without traditional teacher certification. The evidence suggests that there is no reason to limit initial entrance into teaching to those who have completed traditional certification programs or are willing to take such courses in their first years on the job. Many districts already face growing shortages of certified teachers, and removing unwise entry requirements into teaching would also help to address this problem.

Recommendation 2: Make it harder to promote the least effective teachers to tenured positions. In most school districts, tenure is typically granted as a matter of course to those who remain employed for a specified term—usually three years. The tenure process should be changed, since school districts have much better information about a teacher's effectiveness after two years on the job than at the point of recruitment. If schools simply set a minimum tenure standard and denied tenure to teachers below that standard, student achievement would rise substantially. Of course, such a system should be flexible enough to allow for special cases, and should provide sufficient professional development opportunities for teachers in their early years of teaching.

Recommendation 3: Provide bonuses to highly effective teachers willing to teach in schools with a high proportion of low-income students. Today, the lowest achieving teachers are clustered in the poorest schools where students are most in need of effective teaching. Yet even the best teachers at these poor schools are typically paid no more, and sometimes less, than teachers at wealthier schools. Together with other policies, paying more to high-achieving teachers in these schools could draw more effective teachers into these schools and into the teaching profession.

Recommendation 4: Evaluate individual teachers using various measures of teacher performance on the job. There is no consensus yet on the one best way to evaluate teacher performance, so many measures of teacher performance might be used, such as principal evaluations, parent evaluations, classroom observations, and the number of times a teacher is absent. However, measures of outputs and performance rather than credentials would need to be used. Moreover, some measure of "value-added," or the average gain in performance for students assigned to each teacher, would need to be a significant component of that scale. That requirement leads to our last recommendation.

Recommendation 5: Provide federal grants to help states that link student performance with the effectiveness of individual teachers over time. Only a few states currently have the ability to measure the effect of individual teachers on the performance of their students (by comparing performance of classrooms of students with similar incoming performance). This capacity must be built both to facilitate the evaluation of teachers and to supply schools and teachers with better data about what works and what does not.

Our proposals for tenure and pay represent significant departures from current practices. The federal government should initially fund implementation of these more controversial measures in up to ten states. Those efforts should be carefully evaluated and adjusted based on their record. If the concepts prove sound, then with adjustments based on experience, these proposals should be implemented nationally.

II. Recent Evidence on Teacher Quality

Recent evidence demonstrates that teacher certification is a poor predictor of teacher effectiveness. Figure 1 plots the distribution of teacher impacts on average student math performance in grades three through five in Los Angeles Unified School District. The figure is based on the performance of roughly 150,000 students in 9,400 classrooms each year from 2000 through 2003. Figure 1 shows the distribution of teacher impacts for three different groups of teachers—those who were certified when hired, those who were uncertified when hired but participating in an alternative certification program, and those who were uncertified and not participating in an alternative certification program.[1] Controlling for baseline characteristics of students and comparing classrooms within schools, there is no statistically significant difference in achievement for students assigned to certified and uncertified teachers (Kane and Staiger 2005).[2]

While the differences *between* the three groups are small, the differences *within* the three groups are quite dramatic. In other words, there is not much difference between certified and uncertified teachers overall. But effectiveness varies substantially among certified teachers and also among uncertified teachers.

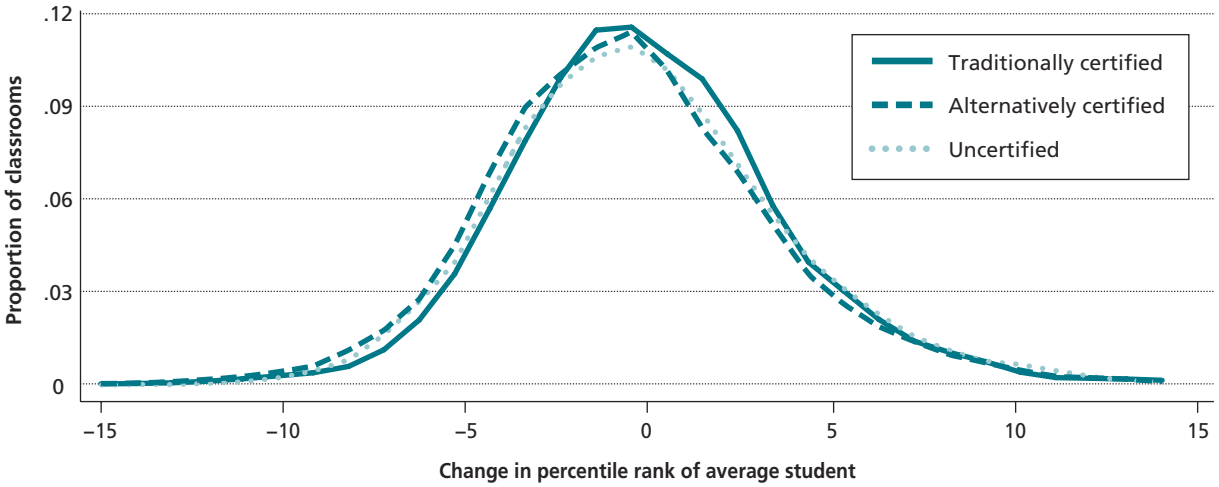
The difference between the 75th percentile teacher and the 50th percentile teacher for all three groups of teachers was roughly five times as large as the difference between the average certified teacher and the average uncertified teacher. The difference between the 25th percentile teacher and the 50th percentile teacher is also about five times as large. And those larger differences are evident even after adjusting for the obvious socioeconomic and educational factors that affect student performance. A similar analysis for distributions of reading scores yielded similar results: that is, certification does not seem to affect classroom performance much, but there is wide variation across teacher effectiveness even after adjusting for many other factors that affect student performance.

To put it simply, teachers vary considerably in the extent to which they promote student learning, but whether a teacher is certified or not is largely irrelevant to predicting his or her effectiveness. But could school district leaders learn anything useful about a teacher's likely future impacts by measuring that teacher's impact on student test scores in the past? How long would it take to make reliable distinctions between more and less effective teachers? To test how well a district could predict future effectiveness using performance during the first couple of years on the job, we focused on a sample of teachers whom we observed in their first, second, and third year of teaching. We measured their students' performance during each of those three years, controlling for students' previous test scores and demographics. We then ranked teachers based on their estimated impact on their students during their first two years of teaching, sorting them into quartiles. Figure 2 reports the distribution of estimated impacts of teachers during their third year, using four separate curves, with each one representing the quarter of the distribution of effectiveness in which the teacher was categorized during the first two years of teaching.

While certification status was not very helpful in predicting teacher impacts on student performance, teachers' rankings during their first two years of teaching does

1. The figure reports the differences in performance that emerge when similar students—with similar baseline scores and similar demographics—are assigned to different teachers. The impacts in figure 1 represent the estimated impact of teachers on the mean percentile score of students in a classroom. Each student's score is being measured on a percentile basis, with each score representing the percentage of students with scores at the student's level or lower in the national norm sample. On the horizontal axis, a value of 5 implies that the average student in the class moved ahead 5 percentile points relative to students with similar baseline scores and demographics. A value of -5 implies that the average student fell behind an additional 5 percent of students with similar baseline scores and demographics. The height of the curves represents the proportion of teachers with a given impact. As reported in figure 1, about 90 percent of the teachers' estimated impacts were between -5 and 5 percentile points.
2. In the analysis reported here, we controlled for demographic factors. We later re-ran our analysis without including such factors and found only modest differences from the results in this paper. We conclude that when the policies advocated here are implemented by the state, rather than simply proposed by researchers, controls for demographic factors should not be used. We discuss this point further below.

Figure 1. Teacher Impacts on Math Performance by Initial Certification



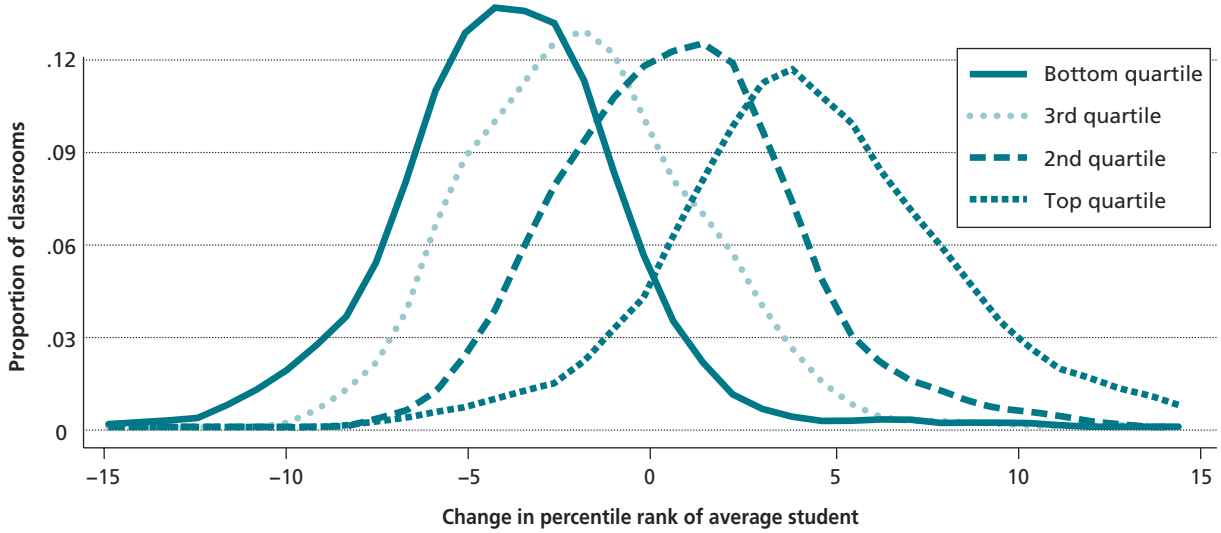
Note: Classroom-level impacts on average student performance, controlling for baseline scores, student demographics, and program participation. LAUSD elementary teachers, grade three through five. For details of how an ordinary least squares regression was used to adjust for student background, baseline performance, and other factors, see the appendix.

provide a lot of information about their likely impact during their third year. The average student assigned to a teacher who was in the bottom quartile during his or her first two years lost on average 5 percentile points relative to students with similar baseline scores and demographics. In contrast, the average student assigned to a top-quartile teacher gained 5 percentile points relative to students with similar baseline scores and demographics. Therefore, the average difference between being assigned a top-quartile or a bottom-quartile teacher is 10 percentile points.

Moving up (or down) 10 percentile points in one year is a massive impact. For some perspective, the black-white achievement gap nationally is roughly 34 percentile points. Therefore, if the effects were to accumulate, having a top-quartile teacher rather than a bottom-quartile teacher four years in a row would be enough to close the black-white test score gap. A random assignment evaluation of a classroom size reduction in Tennessee found that schools could improve achievement by half as much—5 percentile points—by shrinking class size in early grades (Krueger 1999). But class size reduction of the magnitude considered in that experiment is expensive: shrinking average class size from twenty-two to sixteen students per class would require a 38 percent increase in the number of teachers and the amount of classroom space in those early grades.

Although these data come from only one school district, they illustrate three conclusions widely accepted among education researchers and consistent with results from many other places. First, there is wide variation in the effectiveness of teachers, even after adjusting for student characteristics such as baseline test performance, race/ethnicity, family income, gender, and so on. Rockoff (2004) found similar results using data from two school districts in New Jersey. Using data from Texas and Chicago respectively, Rivkin, Hanushek, and Kain (2005) and Aaronson, Barrow, and Sander (2003) report very similar estimates of the variation in teacher impacts on student achievement. Using data from New York City, Kane, Rockoff, and Staiger (2005) find somewhat smaller differences between elementary teachers ranked in the top and bottom quartile. While all of the above were based on nonexperimental methods (that is, they use statistical techniques to control for student characteristics and baseline performance), Nye, Konstantopoulos, and Hedges (2004) analyzed teacher impacts from a random assignment experiment in Tennessee. They found similar variation in teacher impacts on student achievement to those found in the nonexperimental studies.

Figure 2. Teacher Impacts on Math Performance in Third Year By Ranking after First Two Years



Note: Classroom-level impacts on average student performance, controlling for baseline scores, student demographics, and program participation. LAUSD elementary teachers, < 4 years' experience.

Second, with only one or two years of student outcome data, a district learns a lot about which teachers are likely to generate large student learning gains and which are not (as shown in figure 2). And, third, these differences in teacher effectiveness are largely unrelated to whether a teacher is certified. The above results—that those with traditional certification do not outperform those without such certification in promoting student achievement—are mirrored in several recent papers (Jepsen and Rivkin 2002; Hanushek et al. 2005; Ballou and Podgursky 2000; Raymond, Fletcher, and Luque 2001). But even when researchers have found differences in mean performance between certified and uncertified teachers, those differences are usually quite small. For example, a recent study by Darling-Hammond et al. (2005, table 5) found that students assigned to uncertified teachers performed 0.5 percentile points worse on an achievement test than those assigned to traditionally certified teachers and those assigned to alternatively certified teachers underperformed by 2.5 points.[3] Indeed, even in our own work in New York City, we have found that the average traditionally certified teacher raised reading scores about 1 percentile point more than the average alternatively certified teacher (Kane, Rockoff, and Staiger 2005). But a statistically significant difference is not necessarily an important difference: a 1 percentile point difference between groups is dwarfed by the differences within groups. Moreover, a recent random assignment evaluation found that Teach for America corps members considerably outperformed traditionally certified teachers (Decker, Mayer, and Glazerman 2004).

In related research, Hanushek and Rivkin (2004) summarize the research on the predictive power of master's degree completion and find little consistent evidence that graduate degree attainment can identify effective teachers. Similar results are reported in Murnane (1975), Summers and Wolfe (1977), Ehrenberg and Brewer (1994), and Aaronson, Barrow, and Sander (2003).

The evidence described above sets the stage for the five recommendations in our policy proposal for improving the quality of the teacher workforce. The next five sections of this paper lay out each of these five recommendations in more detail. We then pose and answer a number of questions, including how much this proposal would cost, how practical it is, and other issues.

3. These are Stanford 9 math and reading NCE points.

III. Recommendations

Recommendation 1: Reduce the Barriers to Entry into Teaching for Those Without Traditional Teacher Certification

The central provision of the No Child Left Behind Act related to teacher quality is the requirement that teachers of core academic subjects be "highly qualified" by the close of the 2005-06 school year. "Highly qualified" means having a bachelor's degree and obtaining (or being on the way to obtaining) full state certification. It then means different things for different teachers depending on when they were hired and whom they teach. For new elementary school teachers, "highly qualified" also requires passage of a "rigorous" subject-matter test; for new middle- and high-school teachers, passage of such a test or an academic major in the relevant subject; and, for veteran teachers, compliance with these standards or with an alternative "high objective uniform state standard of evaluation" (HOUSSE) established by the state. Although the Department of Education's data show a sharp increase in the number of teachers deemed "highly qualified," it is unclear how much this increase corresponds to any increase in actual teaching effectiveness, as opposed to teachers and administrators becoming more skilled at checking statutory boxes.

We would broaden the definition of a "highly qualified" teacher. Under our proposal, a new teacher would continue to be required to have a four-year undergraduate bachelor's degree and to demonstrate content knowledge. There is fairly consistent evidence that teacher test scores and subject-matter expertise are modestly related to their classroom performance (Goldhaber and Anthony 2004; Cavalluzzo 2004; Vandevoort, Amrein-Beardsley, and Berliner 2004). Such evidence is somewhat more robust for students in later grades. Therefore, we would allow teachers who met these basic requirements to be deemed "highly qualified" if they also demonstrate effectiveness in the classroom, regardless of whether they had met a state's other certification requirements. Specifically, any new teacher scoring above the 50th percentile on the scale of teacher effectiveness at the end of two years would be deemed "highly qualified"—regardless of his or her ability to meet existing certification requirements. Moreover, all current experienced teachers who are rated above the median would be deemed "highly qualified" regardless of their certification status or compliance with other state systems.

Why a Performance-Based Option Is Preferable for Teachers

Under the regime we propose, novice teachers will have two routes into teaching. One point of entry would follow the current model, in which they follow the existing rules leading to certification. However, another route would be provided to novice teachers who have the undergraduate degree and subject knowledge to look for a teaching job and get hired.

Schools will of course remain free to screen for the qualities they deem most important in the classroom; certification simply will not be an iron-clad requirement. And school systems likely will provide training short of that required for full certification. Most principals judge novice teachers who complete Teach for America's intensive six-week summer training program, for example, as at least as well trained as other novice teachers.

Once hired, teachers will have a trial period of a couple of years, and then they can receive tenure based on performance. We expect that this additional option will encourage many of those who suspect that they might have the makings of a good teacher, but are unwilling to commit several years to education school, to enter the teaching profession. Given the large variation in teacher effectiveness, we expect that the full range of good, average, and ineffective teachers will enter the teaching profession in this way. But as a group, those who enter the teaching profession in this way will not be noticeably less effective than those who have pursued traditional certification.

For experienced teachers, who nonetheless need to be certified as "high quality" under the No Child Left Behind Act, most states have established HOUSSE ("high objective uniform state standard of evaluation") standards that are easily met based on assorted past activities, which provide little evidence of genuine subject-matter expertise (Walsh and Snyder 2004; Education Trust 2003). At the same time, the HOUSSE standards have managed to be genuinely burdensome to many good teachers who are forced to rummage through transcripts of classes they took years or decades earlier to demonstrate knowledge that they deploy every day (National Education Association 2005).

Allowing experienced teachers with above-average results to be deemed highly qualified—whether or not they satisfy the other HOUSSE provisions—would simplify the lives of many high-quality experienced teachers by requiring less paperwork and hassle. In addition, meeting a performance-based tenure standard is a better guarantee of a quality teacher than HOUSSE because it reflects actual success in raising student performance. This could be a "win-win" for many teachers and schools.

The Coming Teacher Shortage

Encouraging more recent college graduates and midcareer professionals to enter a teaching career, without requiring them to take (or commit to taking) years of education school classes, should substantially expand the pool of eligible candidates. Recent experience has shown that there is a reserve army of Americans who are interested in teaching. When the Los Angeles Unified School District needed to triple its hiring of elementary teachers following the state's class-size reduction initiative in 1997, the district was able to do so without experiencing a reduction in mean teacher effectiveness, even though a disproportionate share of the new recruits were not certified (Kane and Staiger 2005). New York City's Teaching Fellows program, geared to young and midcareer professionals and still requiring alternative certification, had 16,700 applicants for 1,850 spots. Similarly, Teach for America had 17,000 applicants last year for only 2,000 openings.

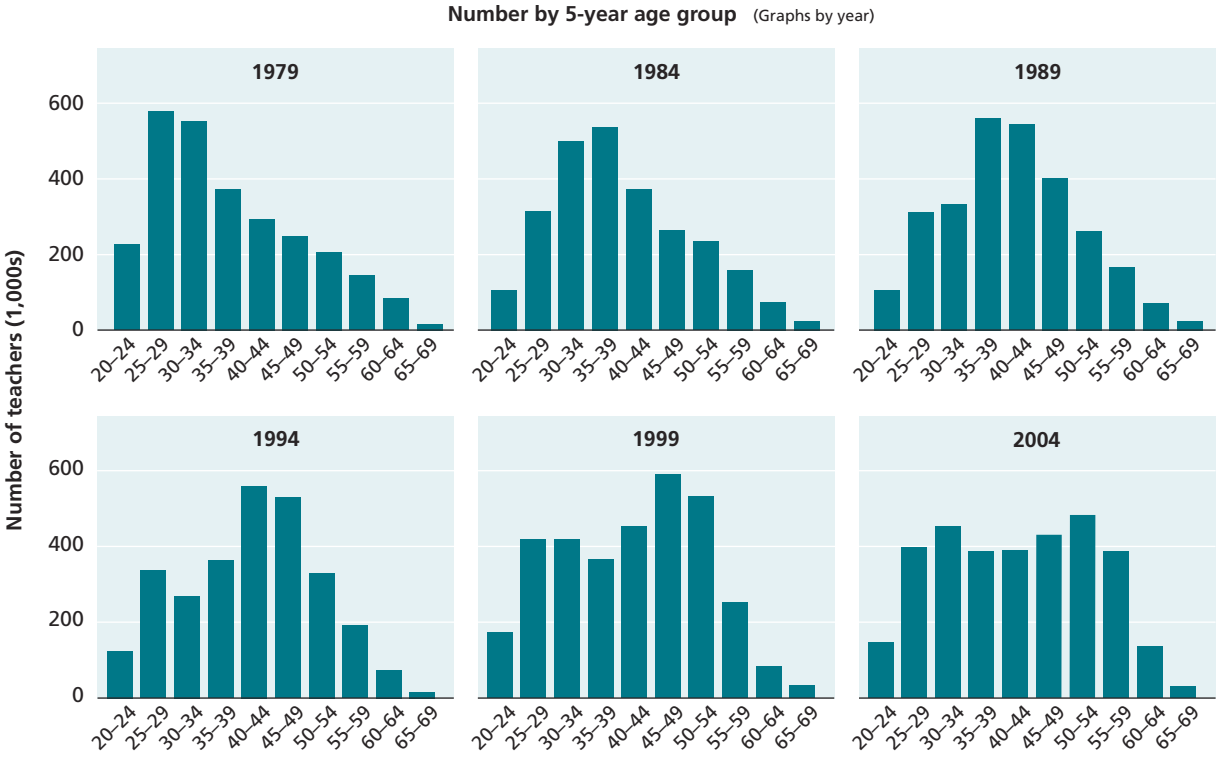
Expanding the pool of teacher recruits is especially important now because America's schools will soon face a growing teacher shortage. The age of primary and secondary school teachers has increased substantially over the last twenty-five years. The median age of a public school teacher (that is, the threshold at which half the teachers are older and half are younger) rose from thirty-three in 1976 to forty-six in 2001 (Snyder, Tan, and Hoffman 2004). There are two underlying reasons for this demographic bubble. First, there was a persistent decline in the proportion of younger women choosing teaching as a career, which occurred in the late 1960s and early 1970s. As career opportunities for women expanded (Blau and Ferber 1992), the proportion of female college freshmen interested in teaching fell precipitously in the early 1970s. Despite a small rebound in interest since that time, the proportion remains below the high levels of the early 1960s (Higher Education Research Institute 2002). Second, elementary and secondary school enrollment started declining in 1970, and districts were hiring fewer teachers (Murnane et. al. 1991). Indeed, the decline in job opportunities in teaching may have accelerated the declining interest of college students in teaching.

Thus, the college freshman of the late sixties were the last cohorts to enter teaching in large numbers. That group is now nearing sixty. Therefore, it is not surprising that 40 percent of public school teachers plan to exit the profession within five years (National Center for Education Information 2005). Similar trends have occurred in other professions traditionally dominated by women, such as nursing (Buerhaus, Staiger, and Auerbach 2000; Staiger, Auerbach, and Buerhaus 2000).

Figure 3 shows the evolution of the age of teachers in recent decades. We plot the number of teachers by five-year age groups for selected years from 1979 through 2004. Two large cohorts can be seen working through the age distribution, e.g. 25-29/30-34 in 1979, 30-34/35-39 five years later in 1984, and eventually 50-54/55-59 in 2004. These are the same cohorts who expressed the highest interest in teaching as college freshman in the late 1960s. These large cohorts are also now heading in to retirement.

Figure 3. The Age Distribution of Elementary and Secondary School Teachers, Selected Years 1979–2004, by 5-Year Age Groups

Teachers are measured in "full-time equivalents," which means that a full-time teacher is counted as 1, a part-time teacher is counted as one-half.



Source: The Current Population Survey Outgoing Rotation Group files, a household-based survey administered monthly by the Bureau of the Census that covers a nationally representative sample of more than 100,000 individuals including more than 5,000 teachers each year.

Over the next twenty years, the U.S. Census Bureau projects that the school-age population age five through seventeen will grow by 10 percent. To maintain pupil-teacher ratios at their current levels, the number of teachers must also grow by 10 percent, from their current level of 3.1 million to 3.4 million. Based on the data in figure 3, we extrapolated the future supply of teachers by aging the current cohorts and assuming that new cohorts will enter teaching at about the same rate as people have for the last two decades. Under this scenario, the supply of teachers will decline over the next decade and then remain at about 3 million through 2025, or nearly half a million teachers below what would be required to maintain current student-teacher ratios.

The bottom line is rather stark: *Simply to maintain pupil-teacher ratios, we must increase the number of people entering teaching by roughly 35 percent—back to levels not seen since the cohorts that came out of high school in the 1960s.* Rather than dig further down in the pool of those willing to consider teacher certification programs or raise class sizes, we need to expand the pool of those eligible to teach. It is time to encourage young people to begin a teaching career without needing to invest in two years of education school first, and to encourage older people to try teaching as a second career.

Recommendation 2: Make It Harder to Promote the Least Effective Teachers to Tenured Positions

Because paper qualifications are not very useful in identifying effective teachers, school districts will inevitably

make some mistakes in choosing whom to hire in the first place. But our results above suggest that if states or school districts were to assemble the evidence available to them, by linking student performance and teacher effectiveness data over time, and estimating differences in performance when similar students are assigned to different teachers, they could learn a lot about which teachers are likely to be effective in the future. Thus, we believe states should establish a presumption, but not a requirement, that teachers in the bottom quartile of effectiveness after two years do not qualify for tenure and are not allowed to continue teaching.

Current Teacher Tenure Laws and Procedures

State statutes typically provide considerable protections to teachers who are already granted tenure (often formally known as "permanent," "continuing contract," or "postprobationary" teachers). For example, tenured teachers may be removed only after an adversarial hearing before a neutral arbiter and often only on specific grounds, which may be exacting (such as requiring proof of "incompetence"). However, these same statutes frequently provide fewer constraints during the first two or three years of a teacher's career, before a teacher is granted tenure. In most states, nontenured teachers may be removed for any reason except grounds prohibited by generally applicable federal or state laws or by the Constitution—for example, racial discrimination, sex discrimination, or politically motivated discharge. Furthermore, nontenured teachers generally are entitled to no hearing on their discharge. Most states award tenure after three years of teaching; smaller numbers of states require two or four years; a few states do not have tenure at all.

Even though school districts have the opportunity to discharge nontenured teachers, they seldom do so. It is rare for public or private school teachers to report being laid off or transferred involuntarily. Table 1 reports tabulations of a survey of public and private school teachers who left teaching or moved between schools following the 1999-2000 school year. Very few of either public or private school teachers report that being laid off or involuntarily transferred was a "very important" or "extremely important" reason for their decision. Less than 1 percent of public or private school teachers moved and cited being laid off or transferred as the reason. For those in their first three years of teaching, less than 2 percent report that they moved schools because of a layoff or involuntary transfer.

It may be that teachers are hesitant to admit that they were laid off or involuntarily transferred or principals may find ways to persuade ineffective teachers to move "voluntarily." However, even if the proportion being laid off or transferred involuntarily is understated by a factor of five, less than 10 percent of new teachers are terminated involuntarily. Given the evidence on the wide variability in teacher performance, it seems clear that schools regularly award tenure to teachers who are quite ineffective in the classroom compared with other teachers who have similarly situated students.

The Impact on Student Achievement of a More Selective Teacher Tenure Policy

Schools could substantially increase student achievement by denying tenure to the least effective teachers. Suppose that teachers were ranked at the end of their first two years of classroom performance as measured by test-score gains among their students (with the scores adjusted so that they do not reflect family income, race/ethnicity, gender, or baseline scores at the beginning of the year). What might a school system expect to gain if the bottom quarter were not renewed for the following year? The outcome would depend on two effects. First, establishing a minimum threshold of effectiveness will raise the average quality of the remaining retained teachers. But second, each teacher not given tenure would be replaced with a novice teacher, who would have less experience. Depending upon the magnitude of learning teachers do on the job, this latter effect could be quite costly. How is the net balance of these two effects likely to work out?

Considering the risks of false positives and negatives, the potential effects on recruitment of teachers, and the overriding goal of increasing student achievement, we focused on quartiles of teachers. Based on the Los Angeles data, a policy that dropped the bottom quartile of teachers after their first year of teach-

Table 1. Percentage of Teachers in 1999–2000 Leaving the Profession or Moving between Schools in the Following Year

	Moved Schools or Left Teaching	Moved to a a New School	Cited Layoff or InvoluntaryTransfer	Left Teaching	Cited Layoff or Involuntary Transfer
Public Schools					
Total	15.1	7.7	0.8	7.4	0.2
By Teacher Experience:					
1–3 years	22.7	13.8	1.9	8.9	0.8
4–9 years	17.7	10.8	0.9	6.9	0.3
10–19 years	12.8	6.7	0.6	6.1	0.1
20+ years	11.5	3.5	0.4	8.0	0.1
Private Schools					
Total	20.9	8.4	1.0	12.5	1.4
By Teacher Experience:					
1–3 years	34.1	10.8	0.7	23.3	1.5
4–9 years	24.6	11.6	1.9	13.0	0.3
10–19 years	12.5	5.9	0.4	6.6	0.7
20+ years	13.6	5.4	0.8	8.2	0.9

Note: Movers are teachers who were still teaching in 2000–01, but who had moved to a different school. Leavers are teachers who left the teaching profession between 1999–2000 and 2000–01. Respondents were asked to rate the importance of various reasons in their decision to move or to leave teaching. Columns 3 and 5 includes those who stated that a layoff or involuntary transfer was "very important" or "extremely important" in their decision. Estimates in columns 1 through 5 are drawn from Luekens, Lyter, and Fox (2004), tables 6 and 7. Estimates in column 5 are from unpublished estimates from the Schools and Staffing Survey provided by Deanna Lyter.

ing would increase the average impact of retained teachers by about 1.5 percentile points.[4] This policy would also require an increase in the hiring of novice teachers in order to maintain class size. The evidence suggests that the average "value-added" of novices is about 4 percentile points lower than for teachers with two years of experience. For example, Los Angeles would have to increase its number of novice teachers from the current level of 9 percent of teachers to 12 percent of teachers, since one-quarter (3 percent) of these would not be retained under the new policy. We could expect a 1.5 percentile point increase in higher student performance among 88 percent of teachers who were not novices. This would be offset by a 4 percentile point decline among the 3 percent of additional novice teachers, for a net increase in student test score gains of around 1.2 percentile points per year.[5]

The cumulative impact of such a policy could be substantial. If the effects of a good teacher in early grades were to persist through high school (a hypothesis that has not been tested in the education research), an annual increase in test scores of 1.2 percentile points at each grade over the course of twelve years in a school system would raise student test scores by roughly 14 percentile points by the time students graduated.

The economic value of such an increase could be enormous. To estimate the dollar value of an increase in academic achievement, we needed some means of converting test scores into dollars. To do so, we used alternative estimates of the relationship between test scores and earnings among young adults from Murnane, Willett and Levy (1995) and Neal and Johnson (1996). Both sets of authors provide estimates of the relationship between earnings and academic achievement in a given year. Using these, we calculated the value of a test-score increase over a student's career.[6] We estimate that the increase in career earnings from a 14 percentile point increase in

4. The gain would be (3.16+5.46+10.08)/3-(0+3.16+5.46+10.08)/4=1.49 percentile points.
5. The gain would be .88*1.49-.03*4=1.2 percentile points.

achievement test scores would be worth about $72,000 to $169,000 per high school graduate.[7] When multiplied by 3 million public high school graduates per year, such an increase would be worth $216 billion to $507 billion per year, if the policy were applied nationwide.

These rough estimates may overstate the gains somewhat. For instance, the estimates assume that the impact of having a particularly effective teacher in an early grade persists over a student's career. But it is not uncommon for learning gains produced in one year to fade somewhat over time. In addition, we may be too optimistic about the quality of the novice teachers that would be attracted, particularly at a time when larger numbers of novices would need to be hired. Moreover, the greater uncertainty about tenure prospects for new teachers might also make it harder to recruit teachers (though our proposal to increase pay substantially for effective teachers at high-poverty schools could help counter this effect). Finally, there are difficulties with implementing the system among younger and older students (because of the lack of availability of baseline test scores or standardized tests that all students take). Nevertheless, even if only a quarter of the gains suggested above were realized, such an improvement in student performance would represent substantial economic value.

Changing the Default for Ineffective Teachers

After a phase-in period, states receiving federal teacher quality funding would no longer be able to grant tenure so easily to teachers performing in the bottom quartile during their first two years. As a general matter, these teachers should not be able to continue teaching in the jurisdiction. However, we would not want to require districts to fire these teachers, because there may be circumstances in which teachers should be retained or granted tenure notwithstanding being ranked in the bottom quarter. For example, a principal may be able to identify certain cases where teachers were inaccurately identified as ineffective or where there were factors beyond the teacher's control affecting classroom's performance.

However, when a principal wishes to allow a low-performing teacher to continue teaching, we would require the principal to meet two requirements. First, the principal must receive a waiver from local district authorities. Second, the principal would have to provide public notice of the waiver, through both letters to parents and some other form of public notification (perhaps on a school Web site, for example). Such a rule would create costs to keeping lower-achieving teachers in the classroom or granting them tenure, but would permit overrides when principals can make a case for them.

Within such a regime, teachers should receive the support needed to maximize their chances for success. New teachers should have access to mentoring and support during their first year of teaching. Such support is particularly important for those without traditional certification, who often will not have had prior experience in the classroom (Johnson, Birkeland, and Peske 2005). Schools should give teachers notice of how they are performing as frequently as possible and, at the latest, after their first year in the classroom. Indeed, just as the presence of high-stakes testing for students may encourage schools to target necessary resources to students in danger of failing, we hope that raising the stakes for new teachers will increase the pressure on districts to ensure that new teachers receive the support they need. Teachers should have notice of their achievement and a reasonable opportunity to improve before the tenure decision at the end of the second year.

Although denying tenure to many low-achieving teachers would mark a sharp break from what actually happens in schools today, it is consistent with the views of many of the key players in education. In a Public Agenda survey, 78 percent of teachers recognized that at least some other teachers in their own buildings "fail to do a good job" (2003). Principals report that they believe many teachers remain in the classroom who do not belong there. (Gordon 2005; Bradley 1999). According to one new study, principals regularly deal with low achievers by "passing them around from school to school" rather than

6. For details on the net present value calculation, see Kane and Staiger (2002).
7. In Kane and Staiger (2002), we estimate that a 1 standard deviation increase in test scores is associated with a $110,000 to $256,000 increase in the present value of lifetime earnings for an eighteen-year-old. A 14 percentile point increase would represent 0.66 standard deviations in normal curve equivalents.

terminating them (Levin, Mulhern, and Schunck 2005). The problem may be, in the words of Michael Ward, North Carolina's superintendent of public instruction, the limited "willingness of school leaders to confront unpleasant tasks associated with dealing with performance problems" (Bradley 1999). Changing the default rule will make it much easier to confront those tasks. The result may be simply to deny tenure to teachers whom peer teachers and principals already recognize are not effective in the classroom.

Currently, in most school districts, the presumption is that new teachers will be offered tenure at the end of two or three years. Such a system rewards longevity, not results. It can be costly in terms of time or personal relationships for a principal to terminate an ineffective teacher. Our proposal would shift the default for bottom-quartile teachers: rather than make it costly to terminate such teachers, we would make it costly to keep them.

Maintaining Commitments to Teachers Who Already Have Tenure

We do not suggest that the policy suggested here be applied to already tenured teachers. These teachers have legal rights and legitimate expectations under both state statutes and local collective bargaining agreements. Moreover, given the coming wave of teacher retirements, the new teachers hired will quickly become the majority of the teaching force anyway.

Recommendation 3: Provide Bonuses to Highly Effective Teachers Willing to Teach in Schools with a High Proportion of Low-Income Students

If current tenure practices screen out too few of the weakest teachers, current pay practices encourage too few of the strongest teachers to work in the schools where they are needed most. Teacher pay scales typically increase salaries based on only two criteria—years of experience and educational qualifications—neither of which is strongly related to teacher effectiveness beyond the first few years of teaching.

Today, only a few school districts offer rewards for high-performing teachers, and these are often modest. Denver is one of the few to do so, yet even there the performance bonus amounts to only 5 percent of base pay (Jupp 2005). According to recent surveys, only eight states provide bonuses of at least $5,000 for teachers with certification from the National Board for Professional Teaching Standards, which has been shown to be correlated with improved performance (Goldhaber and Anthony 2004; Cavalluzzo 2004; Vandevoort, Amrein-Beardsley, and Berliner 2004). At present, a distinct minority of districts offer differential pay to teachers in schools with a high proportion of low-income students, and among those that do so, many fail to screen for teacher quality (Rotherham 2005).

Salary increases for high-performing teachers are particularly critical in schools where a large share of the children come from low-income families. These schools tend to have the weakest teachers. They have the fewest teachers with relevant subject-matter expertise (Education Trust 2003). They also have the fewest teachers certified by the National Board for Professional Teaching Standards (Humphrey, Koppich, and Hough 2005). Using our own data, we find that in Los Angeles, students in the poorest schools (where more than 90 percent of the students come from families that qualify for free or reduced-price school lunch) were more than 2.5 times as likely to have teachers in the bottom quarter of all teachers than were students in the wealthiest schools (where fewer than 10 percent of students came from families that qualified for free or reduced-price lunch).

The inequitable distribution of effective teachers within school districts has many causes. Uniform salary schedules, under which teachers with the same experience and educational attainment are paid the same regardless of their skills or where they work, are an important contributing factor. Uniform pay may sound fair, but it leads to an inequitable distribution of teachers. It may seem counterintuitive that uniform pay could be inequitable, but the reason is that teachers' compensation is determined by their wages and their working conditions. And working conditions are partially determined by the prior

preparation of the students that are assigned to them. For many teachers, high-achieving students with parents who are supportive of education are simply easier to teach. Schools with those students often also have better facilities and safer environments. If teacher salaries are based solely on educational attainment and experience of the teacher, and any teacher would earn the same salary in a high- and low-achieving school, there is no way for low-achieving, low-income schools to compensate teachers for the additional challenges of working in those schools. If they are paying the same wages, principals in high-income schools can effectively offer higher total compensation, since working conditions are generally more desirable. Understandably, once teachers accumulate sufficient seniority, they frequently exercise contractual rights and transfer into wealthier schools (Lankford, Loeb, and Wycoff 2002; Levin and Quinn 2003; Prince 2002).

School finance rules facilitate the inequitable distribution of teachers. Because dollars typically follow teachers within districts, more experienced and better-paid teachers who transfer into schools with less taxing teaching environments effectively bring their higher salaries with them. Schools with students from low-income families not only are left with less costly, less experienced teachers, but also receive no additional funding to raise salaries, hire additional staff, or provide additional services (Roza and Hill 2004).

Salary increases targeted to high-performing teachers in poor schools could help counter all these effects. They could also attract more high-performing individuals to become teachers rather than go into other professions. There is some evidence that the inverse of that effect has already occurred. Hoxby and Leigh (2005) find that as the teacher pay scale became compressed and the premium available to women teachers educated at elite schools declined, the number of elite-educated women going into teaching also dropped. This finding is consistent with a broader literature concluding that the aptitude of individuals entering public sector fields like teaching has declined as compensation in those fields relative to other professions has dropped (Bok 1993; Miller 2003).

To encourage better teaching and to attract more high-quality teachers, we recommend bonus pay for teachers who are ranked in the top quarter by effectiveness and who teach in schools where at least 75 percent of the students come from families with incomes low enough to be eligible for free or reduced-price school lunches. Some states now offer bonuses to teachers willing to work in high-poverty schools, but we do not see the point in offering bonuses to *any* teacher willing to do so—there will be a lot of low-performing as well as high-performing teachers willing to take that offer. Our proposal would provide large bonuses only to teachers with a proven track record who are willing to teach in high-poverty schools.

How Large Should Bonuses Be and How Should They Be Distributed?

There is no settled answer to the question of how large incentives must be to attract and retain high-quality teachers in low-performing schools. Kate Walsh (2005) of the National Center for Teacher Quality suggests that bonuses would need to be 10 to 20 percent of base pay. Others have suggested that even 15 percent is inadequate (Miller 2003), that bonuses would need to be at least $20,000 to have an impact (Rothstein 2004), or that bonuses would need to range between 20 and 50 percent of base salary to attract teachers to the highest-poverty schools (Hanushek, Kain, and Rivkin 2001).

We propose that top-quartile teachers willing to teach in high-poverty schools be provided at least $15,000 in bonus money above and beyond their current salaries. In a profession where salaries currently start at about $30,000 and average about $45,000, this is a substantial increase. As noted above, we would define a high-poverty school as one where more than 75 percent of the students qualify for the federal free or reduced-price lunch program. Such schools represent about 21 percent of public school enrollment.

Alternative approaches to raising pay are possible. One could offer bonuses of differing amounts, graduated according to the poverty rate in the school, with some bonuses for teachers at all schools. At schools with 51 to 75

percent of students receiving subsidized lunches, for example, the federal government could subsidize bonuses of up to $7,500. An additional 19 percent of students attend such schools. In schools with fewer than half of students receiving subsidized lunches, support for very modest bonuses might be available, up to $2,000. Still another approach, with more flexibility, would be to send money for salary bonuses to the district and school based on the percentage of students in poverty, and then require those districts and schools to allocate the bonuses to the highest-achieving teachers.

Instead of providing a fixed sum like $15,000 for all teachers in the top quartile, policymakers could provide a bonus as a percentage of the teacher's salary instead. This approach would have the advantage of keying to base teacher pay, which will bear some relationship to the cost of living in the area. But this approach would also provide larger bonuses to teachers who are earning more because of their seniority. Given our evidence that teachers do not substantially improve their performance after their third year in the classroom, that skew in performance-based bonuses does not seem wise. Our proposal would in any event provide bonuses only after the second year, when teachers have already typically achieved their largest improvement. To address regional variation, however, the $15,000 might be reformulated as a percentage of base pay for starting teachers.

Teachers who wish to be eligible for additional compensation would need to be reassessed periodically. As a matter of fairness, we would give new teachers two years to get their feet under them, provide notice of their performance after the first year, and make decisions after their second year. We also would require reassessments every five years. Such reassessments would recognize when teachers burn out or when they sharply improve over time. But the reassessments would not be so frequent that they would become a constant presence in a teacher's life.

We do not suggest that increasing pay alone is a complete strategy for attracting more high-quality teachers into poor districts. The quality of school facilities, school supports, and school safety all play important roles in teachers' choices of where to go. Our proposal is not a cure-all for the maldistribution of teachers, but it will help significantly.

Recommendation 4: Evaluate Individual Teachers Using Various Measures of Teacher Performance on the Job

Each of the first three steps relies on a working definition of classroom effectiveness. States and districts will need to implement a practical definition of classroom effectiveness. In establishing such systems, several challenges arise, such as 1) balancing objective and subjective factors; 2) using appropriate control factors; 3) applying the system to teachers in grade levels and subjects where there is currently no testing; 4) measuring performance relative to other teachers or relative to an absolute standard; 5) addressing concerns about fairness; 6) addressing the role of principals; and 7) choosing the appropriate level at which the measures should operate—state, district, or school. We consider these issues in turn.

Objective and Subjective Factors

Impacts on measured student achievement should be a substantial factor in teacher evaluations. Such changes are the most tangible evidence of a teacher's accomplishment. Simply providing such estimates to principals may prove particularly valuable in teacher promotion decisions. A measure of students' growth in performance, benchmarked against the performance of similar classrooms of students elsewhere, may be the first piece of "objective" evidence principals have been given to make difficult decisions regarding tenure.

However, no single measure of performance is a perfect measure of what students should be learning, and statistical evidence from student scores should not be the only measure by which teachers are evaluated (Walsh 2005; Feldman 2004). There is growing evidence that the tests and assessments now in use are not adequately aligned with state standards and not sufficiently sophisticated to measure high-level student skills (Toch 2005). And as states have implemented systems to raise accountability for student test scores, researchers have documented

troubling evidence of teachers and principals cheating (Jacob and Levitt 2003), narrowing of the curriculum to tested subjects such as reading and math (Koretz 2002), and increasing instruction geared to particular tests. If the stakes on student tests are too high, the looming presence of such tests can distort the classroom learning experience.

A wide range of other methods of evaluating teachers are possible. Principals, teachers, and other educators, from inside or outside the school, can evaluate teacher performance based on both classroom observation and reviews of student work. The use of multiple evaluators from inside and outside of a particular school can reduce the risk that any individual evaluator lets personal biases color his or her judgment. Parent evaluations can also be taken into account. The National Board for Professional Teaching Standards has its own multifaceted method for certifying effective teachers, including videotapes of classroom instruction, examples of student assignments, and teacher feedback to students.

Sound objective and subjective measures of teacher quality are likely to converge, at least for those teachers at the top and bottom of the distribution of teacher quality. For example, Jacob and Lefgren (2005) recently asked principals to subjectively rate teachers' ability to raise the math and reading achievement of their students. Nearly 70 percent of those who received top ratings from their principals in their ability to raise math achievement were in the top of the distribution of value-added on test scores. In reading, more than 50 percent of those who were in the top of the subjective ratings were in the top of the value-added metric using test scores. In general, although there was more disagreement in the middle of the distribution, principals' subjective impressions lined up with the quantitative evidence for the most and least effective teachers. Murnane (1975) and Armor et al. (1976) also found that subjective ratings by principals were correlated with value-added measures.

We propose that states be offered funding to establish systems for evaluating teacher performance. As there is no consensus on the single best way to evaluate teachers, states should be allowed to develop different methods of evaluation that weight different items in different ways. We would impose only three substantive constraints. First, although states would be permitted to incorporate any outcome-based measures of teacher performance, like those just mentioned, they would not be permitted to use measures such as licensure status, degrees awarded, or courses or tests taken. (We would allow certification by the National Board to be used, since that particular certification does include some performance assessments and since Goldhaber and Anthony [2004] and others have shown that such certification is related to teacher effectiveness.) Second, a substantial portion of the evaluation, but not the entirety—perhaps one-third to two-thirds of a total score—should be tied to student test scores in one form or another. Third, states would be required to ensure that data collected over a period of time, not just a single school year or a few months within a year, represents a substantial aspect of the evaluation.

Many school districts already provide evaluations of individual teachers. Unfortunately, in many districts, virtually every teacher gets a satisfactory evaluation because principals have little incentive to make distinctions among teachers. Under our proposal, if all teachers were evaluated as "satisfactory," such evaluations would play little role in determining who was in each quartile. The measures along which teachers varied—such as student achievement impact—would account for much more of the variation in teacher rankings. However, it is hard to imagine that a system that was driven solely by the test-based measures of value-added would ever be viewed as fully legitimate. To earn legitimacy, school systems will have to develop alternative ways to discern among their teachers beyond simply test scores.

Use of Control Factors

A performance-ranking system must control for baseline test scores, so that teachers are held accountable for their ability to raise achievement, not for students' preexisting knowledge and skills. Thornier questions arise about whether to control for other characteristics such as income, gender, and race.

Controlling for these characteristics, as we have done in this paper, ensures that each teacher is in effect compared only against other teachers with demographically similar classrooms. School-lunch status and race, for example, provide some information about students' income and socioeconomic status.

In theory, if background characteristics are not controlled, expectations for teachers with disadvantaged students could be higher than the historical performance of those students could justify. Teachers might then be effectively punished for having poorer students. A teacher of disadvantaged students who is performing well relative to his or her peers teaching similar students might not quality for tenure, for example, only because the achievement of poor students in general is predictably lower on average. This could bring about the perverse effect of discouraging teachers from going into these students' classrooms.

On the other hand, by using control factors, the government would effectively be instituting different standards for students based on race, gender, or income. For example, where poorer students have shown lower gains in the past—perhaps in part due to lower expectations—their teachers would face a lower threshold of expected gains. Particularly given the abundant evidence that academic expectations can be self-fulfilling, such controls could send a destructive signal to teachers and students.

We considered a practical question: To what extent does this trade-off actually arise? How much does controlling for racial composition and other student background characteristics actually matter for the teacher evaluations?

To gain some insight into this question, we first estimated teacher impacts on math performance, controlling only for student baseline test scores in reading, math, and language arts and an indicator for whether the student is currently repeating a grade (as well as interactions with all these with academic year and grade level). We did not include any direct socioeconomic background measures. Second, we added indicators for student race/ethnicity, gender, participation in federal lunch-subsidy programs, and English Language Learner status. The correlation between the two measures was 0.98. Ninety percent of those who were in the top (and bottom) quartile on one measure were in the top (and bottom) quartile on the other measure. So, as long as the estimates are controlling for student baseline test scores, it made only a modest difference whether or not there were additional controls for demographic characteristics and family background.

Given the evidence that expectations can be self-fulfilling, and given the absence of evidence that correcting for socioeconomic characteristics significantly affects which teachers are rewarded, we recommend that the state not control for income, gender, and race.

Evaluating Teachers in Early Grades and High Schools

Nearly every state now tests students annually in reading and math in grades three through eight. Therefore, it should be possible to construct a system to evaluate the performance of those teaching math and reading in grades four through eight. Such an analysis can adjust for baseline academic performance relying on data on the performance of students from the prior spring. But in most states, a number of K-12 teachers will not be covered well by the current tests, including teachers in kindergarten through second grade, middle school teachers teaching subjects other than math and reading, and many teachers working at the high school level.

For those teaching in elementary schools, a state could require probationary teachers to start teaching in grades four or five, where their performance could be monitored using the student test-score data. However, to the extent that there are specific talents and skills appropriate for teaching in kindergarten through third grade, this option may not be attractive. In middle schools, the typical student receives instruction from several different teachers over the course of a day. To the extent that the quality of instruction in one subject (like science), spills over and affects a student's performance in another subject (like math), it may be difficult to separate out the contributions of individual

teachers. In high schools, there is the additional problem that students generally self-select into courses and take courses at different difficulty levels. The problem of controlling for all the relevant baseline differences between students, which is a distinct challenge in elementary grades, would be even more of a challenge in high school.

For those teachers working in grades and subject areas that do not lend themselves to value-added assessments, states and districts will have to rely on other measures to evaluate their performance. For these teachers, evaluations by principals, peers, or parents will necessarily play a larger role.

One option is simply to focus the new evaluation systems on teachers in tested grades and subjects. This would create unhelpful incentives for low-achieving teachers to leave the tested fields and high-achieving teachers to enter them. It is important to avoid such distortions, and, more important, to develop sound methods for evaluating teaching performance in all fields. After all, even though we do not currently have national mandates for testing of first-graders or eleventh-grade Social Studies students, there is no reason to believe that the distribution of quality among teachers in these fields is less broad than the distribution for teachers in the tested subjects and grades.

For these reasons, we would encourage states to develop alternative evaluation systems for teachers in nontested grades and subjects where value-added measures may not be practical. One potential model is Connecticut's Beginning Educator Support and Training (BEST) program in which new teachers submit portfolios of their work, including lesson logs, videotaped segments of teaching, examples of student work, and reflective commentaries on the goals during the lesson. Portfolios are scored by multiple external assessors with experience in the same content area as the beginning teacher. The assessors go through approximately fifty hours of training to able to score portfolios. Measures along these lines provide a promising model for evaluations on grounds other than test scores.

Absolute and Relative Standards

Should teachers be evaluated on an absolute scale, where in theory all could succeed or all could fail? Or should they be graded on a curve and evaluated on a relative scale, where inevitably some will be at the top, the middle, and the bottom? Each approach has advantages and disadvantages. With an absolute standard, evaluators may be pushed by political and personal considerations to dilute the standards so that few teachers face negative consequences. (This is a real concern: states have already responded to No Child Left Behind's demand for rising student "proficiency" by defining the definition of proficiency downward.)

But relative standards have other pitfalls. If performance is measured relative to other teachers in the same school or district, teachers will be competing for a finite number of tenure positions or performance awards. In such a system, teachers may be discouraged from collaborating. The ultimate goal of performance reviews is not to pit teachers against one another, but to encourage excellence among all teachers.

An alternative approach would be to use a combination of relative and absolute standards. A threshold could be established using a relative comparison in the first year of a program, but then could be held constant over time. For example, a state might set an absolute cutoff at the level of achievement growth achieved by the 25th percentile teacher in the first year. If average teacher effectiveness improves, more than 75 percent of teachers might exceed that threshold in future years. But such systems also have problems: to the extent that subjective measures like peer evaluations are included, future evaluations could be artificially inflated. In addition, as performance measures are added or improved, it will be difficult to continue using the original benchmark.

Although we recognize that no solution is without problems, we believe it is essential to use a measure that resists manipulation. For that reason, we would require evaluation of teachers relative to each other and would impose consequences based on relative rankings at the

state or district (but not the school) level. Potentially unfair consequences would be mitigated by permitting principals to make exceptions when they were willing to justify their actions to district officials and to parents in their schools.

Ensuring Quality, Fairness, and Teacher Participation

A rigorous performance-based system will succeed over the long-run only if it is perceived as fair by teachers who must live with it. As a result, it will be critical that performance measures be developed through an open process in which teachers fully participate. Indeed, the full array of stakeholders—including parents, teachers, and principals—should be involved in the design of such measures. The plan using performance measures recently approved by Denver voters, for example, was developed with extensive involvement by teachers themselves. In addition, not only should the process be open, but the measures themselves should also be transparent. "Merit pay" has often become a synonym for principals handing out rewards to favorite teachers based on grounds only the principals themselves know. The grounds for performance measurements should be subject to public review.

The Role of Principals

Beyond their important general role in schools, principals also play a critical part in the success of evaluation systems as we propose. They are likely to be invested with significant authority for evaluating teachers and deciding whether exceptions to quartile rankings should be offered. Principals must have incentives to make judgments based on teacher performance rather than personal preference. Although beyond the scope of this paper, principals should ultimately be subject to a parallel incentive system regarding any tenure they may enjoy, as well as pay. Principals could be evaluated based on the performance of the teachers they allowed to earn tenure on their watch. So, even if a principal were to move between schools, their evaluation could depend upon the learning gains generated by all the teachers they ever recommended for tenure.

State, District, or School Evaluations

Another key design question is whether teachers are measured against teachers in the same school, the same district, or the same state. To ensure that students compared are as similar as possible, one may be tempted to compare teachers within the same school, on the assumption that students attending the same school may be similar in ways that justify such a comparison. This might also help ensure that the teachers are operating in similar facilities and with similar administrative supports. However, making comparisons within the same school has disadvantages as well. As noted above, it weakens incentives for teachers to collaborate, since teachers know they are being ranked relative to each other. Moreover, making comparisons within schools disadvantages those teaching in schools where the average teacher is high-performing—and gives undeserved credit to those teaching in schools where the average teacher is low-performing.

Our preference would be to measure teacher quality at the district or state level. To address concerns about comparing teachers with very different student populations, a state or district could rank teachers within peer groups of comparable schools (e.g., based on size or location). A big advantage of a state system is that it would be able to track students moving across school district lines. However, doing so requires the states to have a statewide student identification system. At least initially, districts may be in a better position to launch such a system. As a result, districts would be eligible to apply for federal funds to develop their data systems if the states are not in a position to do so. This necessity leads to our final proposal.

Recommendation 5: Develop Data Systems to Link Student Performance with the Effectiveness of Individual Teachers over Time

If a system for evaluating teacher effectiveness is to work well, data systems are needed that can track the performance of individual students from year to year and link these results with their teachers. Technical as-

sistance must also be provided on how to use these data systems.

Although the No Child Left Behind Act requires states to test in grades three through eight in reading and math, only a subset of districts and states have linked student outcomes to teacher identifiers and followed students over time. Tennessee began doing so in 1992, developing measures of teacher "value-added" similar in spirit to those described above. Ohio, Florida, and Colorado have created or are creating such tracking systems. Some cities, such as Dallas, have created these systems as well (Toch 2005; Carey 2004). However, most states currently do not have the needed longitudinal data systems (National Center for Educational Accountability 2005).

The Institute for Education Sciences at the U.S. Department of Education has a modest grant program designed to support states' development of data systems that provide raw material for measuring the value that teachers add in the classroom. The emphasis on development at the state level makes sense: individual school districts, particularly those in urban areas, may have a difficult time developing such systems because students may move frequently from one district to another.

The federal government should expand its support of state efforts to assemble data and provide sufficient funding for all states to develop and implement longitudinal data systems linking teachers and students. In the context of the U.S. K-12 education system, the costs of supporting the development of improved data for tracking student performance and linking it to teachers nationwide would not be great. Dallas estimates that its value-added system, serving 160,000 students, cost about $210,000 to start up and now costs about $100,000 per year to operate. Hoxby (2002) has suggested that the costs of starting up and administering a wide range of accountability systems has been similarly small.

IV. Implementation and Costs of Our Five Recommendations

We propose a two-phase implementation of our proposals. The first phase would last three years. During this phase, the federal government would support all states in adopting our fifth recommendation: developing the data infrastructure required to track students on a longitudinal basis. Initially, participation would be voluntary. Some states would be in a better position to implement such a system than others. However, by 2009, we would require states receiving funding under Title II (the teacher quality title) of the Elementary and Secondary Education Act (ESEA) to have in place operational longitudinal data systems for linking student performance and teacher effectiveness. Assuming a cost of approximately $4 per youth for start-up and $2 per youth for operating the system, the cost of the implementation of these data systems should be $200 million in one-time start-up costs and $100 million in annual operating costs.

During phase one, the federal government would also fully fund implementation of our four other recommendations in up to ten states. These states would be selected by the Department of Education based on a competitive application process. To encourage states to compete based on the quality of their proposals rather than their financial resources, the federal government would pay for the main programming costs in this phase. This means the federal government would pay for the implementation of teacher ranking systems (recommendation 4), the pay bonuses for teachers (recommendation 3), and the new tenure policies (recommendation 2). Without cost, we would also modify NCLB so these states could (and would be required to) provide a performance-based path to "highly qualified" status (recommendation 1).

At the end of the first implementation phase, we would evaluate the success of the initiative, including comparisons between states inside and outside the new initiatives. We could see, for example, if states that denied tenure to bottom-quartile teachers saw higher gains in student achievement than states that did not. As part of the first phase, some states outside the pilot might also be funded to establish evaluation systems for their teachers but not yet act on the results of those evaluations. These states could keep track of the teachers ranked in the bottom quartile, who would not continue teaching under the proposed policy, and could see how their students perform in subsequent years.

The cost of our proposals depends on many variables. We assume the following: participation by ten states having typical proportions of the U.S. and low-income populations; $15,000 bonuses for top-quartile teachers in the 21 percent of schools meeting our high-poverty definition, with proportional representation of top-quartile teachers in these schools; and additional funding, equal to 25 percent of the amounts spent on bonuses, to implement the evaluation system and enhance professional development. Based on these assumptions, costs in phase one would be about $600 million per year. Together with the costs of the data systems nationally, the costs in the first five years would total perhaps $800 million per year. Of course, if bonuses were offered to teachers at schools with lower levels of poverty, if poorer schools came to have disproportionately large rather than disproportionately small numbers of high-performing teachers, or if more extensive technical assistance proved necessary, the costs of the program could be significantly higher.

Based on the results of phase one, we would expect modifications to be made. If the initiative broadly succeeded, we would propose to take it national. In the national phase, we would require states receiving any Title II funding to have in place new systems for evaluation, tenure, and pay in poor districts. We would also allow teachers in all these states to be deemed "highly qualified" based on performance.

Our proposal assumes that the federal government would bear the full cost of this program, just as it now pays the entire cost of existing teacher quality programs under

ESEA. When fully implemented, the salary bonuses and operation of the data systems would cost slightly more than $3 billion per year. Even if costs ultimately proved higher for various reasons, they would still be relatively small in context. The nation currently spends more than $500 billion per year on K-12 education, of which the federal government pays nearly $38 billion per year. For a small fraction of those sums, our proposal could begin to change the way American schools induct, tenure, and pay American teachers.

V. Questions and Concerns

Won't these proposals undermine the status of teaching as a profession?

To attract and retain highly skilled individuals, teaching must be an honored profession. Many high-status professions, like law and medicine, have high barriers to entry. One concern is that our proposal, by lowering barriers to entry into teaching, will diminish the social status of teaching as a profession.

Barriers to entry may be one element of high-status professions, but they are neither necessary nor sufficient. There are plenty of modestly regarded professions with distinctive certification requirements, from forensic scientists to real estate agents. Today, obtaining an education degree and certification, however time-consuming, is not perceived as a large challenge for talented individuals. Adoption of our proposal would signal that long-term standing in the teaching profession depends on a more challenging achievement—some success in the classroom. Our proposal would also enable teachers who demonstrate excellence in the most challenging classrooms to earn higher pay. That higher pay could also be coupled with other steps to elevate such high-performing teachers, such as use of career ladders and master-teacher status. These measures together could improve the standing of teaching as a profession built not on paper qualification, but on excellence.

Aren't involuntary layoffs rare in the private sector? Why are teachers different?

According to the U.S. Bureau of Labor Statistics (2005), 1 to 1.4 percent of employed persons are laid off or discharged every month.[8] In table 1, we reported that a similar percentage of teachers report moving schools or leaving teaching involuntarily in a whole year.

However, the production process in education is very different from other sectors. An employee hired in the mail room in a modern corporation can remain in the mail room or be promoted. The same is true for employees hired to be stock analysts, accountants, or salespeople. It is typically assumed that as they gain skills and experience, employees will move on to more responsible tasks. When they meet expectations, they are promoted; when they fall below expectations, they remain at the entry level. Firing may be rare, but it is not at all rare for employees to be passed over for promotion.

For teachers, there is no equivalent to the mail room. A low-performing teacher has as much responsibility for a class of students as a high-performing teacher. (If a high-performing teacher has leverage to influence classroom assignments, the low-performing teacher may actually get larger class sizes or the students with the poorest prior performance.) When a low-performing teacher is retained, his or her students pay the price. All else equal, particularly given the difficulty in identifying effective teachers based on paper qualifications, one might even expect to see higher discharge rates in schools than in other industries. At present, they seem to be considerably lower.

How reliable are quantitative measures of teacher effectiveness?

One concern is that quantitative measures of teacher effectiveness will be unreliable because of statistical noise. Even if a teacher's skills and effort remain largely the same from one year to the next (after the teacher has a few years of experience, at least), the average performance of students in the classroom will differ from year to year. In a typical fifth-grade classroom, with only about twenty-five students taking the test each year, a few particularly bright or particularly rowdy pupils can substantially affect the average performance of the class. A teacher may look good either because the students in that year did unusually poorly on the baseline test or unusually well on the follow-up test.

8. This is based on the layoff and discharge rate from the Job Opening and Labor Turnover Survey (JOLTS), 2000–2005.

Some years a construction crew may be working loudly across the street on the day students take their evaluation exam, driving down the test scores for the class, or perhaps two of the low-scoring students in a class will come down with flu on the day of the test, bringing up the class average.

These extraneous sources of variation mean that evidence on teacher effectiveness mixes together true differences between teachers and other, potentially random, factors. As described in the technical appendix, we have attempted to adjust downward the variation in teacher effectiveness reported in figures 1 and 2, using our best estimates of the proportion of the variation that is due to nonpersistent or random factors. While we may still be making mistakes due to random errors in categorizing individual teachers, the total variation depicted in these figures does not reflect these other factors (that is, the total variation has been "shrunk" to match our estimate of the persistent variation).

We also find that the teachers who seem to have a positive impact on math achievement also seem to be able to raise reading achievement. On a scale where a correlation of 0 means that those teachers who have a positive impact on math are completely random in how they perform in teaching reading, and a correlation of 1 means that all the best teachers in math are also all the best teachers in reading, the correlation between teacher effectiveness math and reading achievement was roughly 0.6.

While there is a degree of randomness in evaluating teachers, the evaluation does capture actual aspects of performance. Our proposal does not attempt to use measures of teacher performance to make fine gradations, but instead focuses on who will look either quite effective or quite ineffective largely regardless of the evaluation system that is used.

Why not focus on improving teacher quality by investing in training for existing teachers?

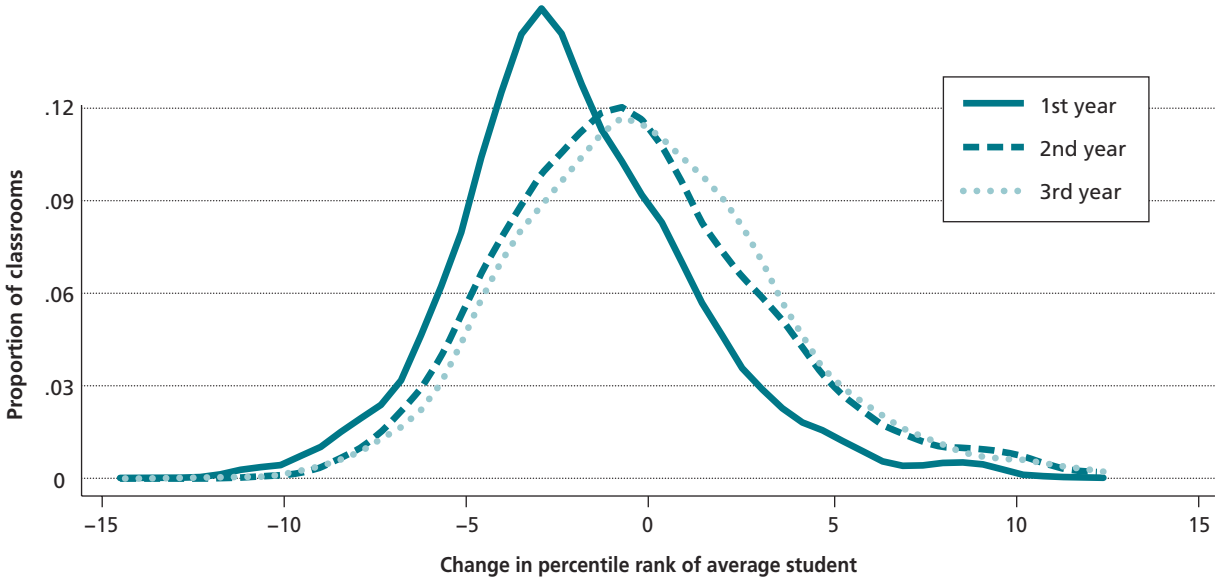
Many school districts currently invest heavily in professional development for existing teachers. However, we believe that efforts to selectively retain the most effective teachers are more likely to generate large increases in average teacher effectiveness than additional training of the existing teaching force.

As evidence for this admittedly provocative proposition, we present data from the Los Angeles Unified School District about how teacher effectiveness changes with experience. In figure 4, we report the distribution of the estimated impacts on performance by year of experience for the sample of teachers whom we observed during all three of their first years of teaching. (By focusing on those whom we saw all three years, we have to worry less about the changing composition of teachers who exit after their first or second year [Hanushek et al. 2004; Podgursky, Monroe, and Watson 2004].) As is done throughout this paper, these average scores for a classroom can be taken to represent the statistically average student—that is, they have been adjusted for race/ethnicity, gender, family income, and scores on an earlier baseline test.

Figure 4 illustrates three interesting facts. First, there are large gains in teacher effectiveness between the first and second year of teaching, but much smaller gains between the second and third year. The difference in mean math impacts is approximately 3 percentile points between the first and second year of teaching and roughly 1 percentile point between the second and third year of teaching.

Second, the distribution of teacher effectiveness does not seem to become more narrow by the third year: the curve for teachers in their third year is just about as wide as the curve for teachers in their first year. (In fact, it is slightly wider.) In other words, as teachers gain experience on the job, their effectiveness does not seem to converge. This has potentially important implications. For example, suppose that some teachers started out effective and remained so and other teachers started out ineffective, but got better. We would expect the distribution of teacher impacts to become narrower with each year of experience. This does not happen. In other work, we have shown that the reverse is true: those who start out effective in their first years of teaching tend to get better faster than those who start out ineffective (Kane and

Figure 4. Teacher Impacts on Math Performance by Year of Experience



Note: Classroom-level impacts on average student performance, controlling for baseline scores, student demographics, and program participation. LAUSD elementary teachers, < 4 years' experience.

Staiger 2005; Kane, Rockoff and Staiger, 2005). In other words, the teachers to start out more effective seem to improve at a slightly faster rate than those who start out less effective.

Third, the magnitude of the payoff to experience—about 4 percentile points over the first three years of teaching—is small relative to the difference in effectiveness between those identified in the top and bottom quartile. Remember from a previous section that the difference in teacher effectiveness, as measured by impact on the math score between a teacher identified as having been in the top and bottom quartiles in their first two years, is 10 percentile points. That is, the return to moving from one to three years of experience is less than half as large as the difference between teachers identified to have been in the top and bottom quartile in their first two years.

Districts invest considerable resources in the professional development of their teachers (much of it through salary points for teachers completing graduate coursework). Without attempting here to assess the vast evidence on how well these programs work, or how they might work if they were better designed, it is hard to imagine that such retraining efforts could generate the same learning that each teacher goes through on the first year on the job. Anyone who has ever taught knows how steep the learning curve is during the first year or two in the classroom. Thus, the return to experience during the first few years of teaching is surely an upper bound on the potential effectiveness of later investments in professional development. However, as noted above, the return to the first few years of experience is less than half as large as the difference between the highest- and lowest-performing quartiles of teachers in their first two years.

All this said, changes to tenure policies should be complements to, not substitutes for, teacher training efforts. One possible use of the evaluation systems described in this paper would be to identify the highest-achieving teachers and single them out to provide mentoring to teachers who are struggling. Our point is that, rather than simply invest in professional development in the hope of solving the problem of ineffective teaching, districts should place greater emphasis on selectively retaining effective teachers and then invest in professional development for them.

What other potential uses do we see for new teacher evaluation systems?

The system for evaluating teacher effectiveness that we have outlined in this paper could be put to many uses. Given that there is a degree of randomness in any method of evaluation, we would not support using the results of this evaluation system to, say, fine-tune teacher salaries on an annual basis. However, we would tend to welcome uses of this system that rely mainly on very large distinctions, such as those between teachers who are consistently in the lowest quarter or the highest quarter of effectiveness over time.

For example, the data could also be used to address the consequences of current transfer policies. Under the current regime, principals will frequently transfer or excess tenured low-performing teachers, rather than going through the complex process to discharge them. (Levin, Mulhern, and Schunck 2005). Many have argued that low-income schools bear the brunt of these involuntary transfers. (Saunders 2005). It may be that the best and simplest way to deal with such involuntary transfers is simply to stop schools from being forced to accept teachers they do not want. Short of that solution, a district might adopt a policy that a school cannot be required to accept a teacher in the bottom quartile of teacher effectiveness. Such a policy would require principals to address the consequences of subpar teaching in their own schools, rather than shifting those teachers elsewhere. This would add to principals' incentive to take the tenure decision seriously, since they could not count on ridding their schools of ineffective teachers earning tenure.

We expect that the availability of better data on teacher effectiveness will set off a cascade of other activities at the district and school level: to study the characteristics of effective teachers, to measure teacher effectiveness more carefully, to target effective teachers as mentors, to identify teachers who need additional assistance during the school year, and so on. One problem with NCLB today is that principals and teachers rarely receive timely, useful information about how they are doing. The systems proposed here would furnish school professionals with detailed and critical new data. In fact, these systems will permit more sophisticated evaluation of school performance than the "adequate yearly progress" measure now used under the No Child Left Behind Act.

Are there potential legal barriers to implementing this proposal?

One important advantage of our proposal is that it is consistent with many existing tenure laws and collective bargaining agreements. We are not contemplating a wholesale shift to performance-based pay. Likewise, we are not proposing to revoke tenure for existing teachers or dismantle the system for future teachers. Indeed, the system may well help provide legitimacy to teacher tenure in the future, by ensuring that teachers clear a real hurdle before being granted tenure.

Most collective bargaining agreements already allow for careful scrutiny during the initial probationary period; our proposals would simply engage in the scrutiny that these agreements allow. The proposals also do not alter the fact that teachers will be paid according to years of experience and paper qualifications—except for the bonuses proposed here. Moreover, the proposal would help schools meet the federal requirement that all teachers be "highly qualified" by offering an alternative avenue for other professionals to get into teaching.

VI. Conclusion

Although it can be difficult to know with much certainty who is likely to be an effective teacher during a job interview, we have shown that school districts can learn a lot about teachers' future effectiveness simply by scrutinizing their record during their first few years on the job. Currently, such information is not being used. Indeed, it is usually not even assembled—since most districts now cannot link individual student test scores to teachers.

Over many years, American schools have experimented with various reform strategies, from increasing accountability to reducing class sizes. Given that history, we are unlikely to get dramatic new results from pushing a little harder on these familiar levers for reform. For instance, in school systems that already have good accountability systems, further ratcheting up the pressure is not likely to produce sudden improvements. Moreover, raising the hurdles for entry into the teaching profession a little higher is not likely to generate a watershed improvement in teacher quality. But partially because most districts have never assembled the data required to calculate the "value-added" by individual teachers, the payoff to beginning to do so could be enormous.

Traditionally, policymakers have tried to raise teacher quality by raising the hurdles for those entering teaching. But our results suggest that those hurdles are often not related to teacher effectiveness. Rather than continuing to focus on teacher *credentials*, our proposal would build the infrastructure to measure teacher *effectiveness* on the job and to encourage states and districts to use that information.

Technical Appendix

To estimate each teacher's impact on student achievement in each school year ($\hat{\delta}_{teacher,year}$), we used student-level data to estimate the following equation by ordinary least squares:

$$S_{it} = \beta_{1gr,yr} Math_{it-1} + \beta_{2gr,yr} \text{Re}ad_{it-1} + \beta_{3gr,yr} LangArt_{it-1} + \lambda_{1gr,yr} Race/Eth_i + \lambda_{2gr,yr} ELD_{it} + \lambda_{3gr,yr} FreeLnch + \lambda_{4gr,yr} Male_i + \lambda_{5gr,yr} GATE_{it} + \lambda_{6gr,yr} \text{Re}peat_{it} + \delta_{teacher,year} + \varepsilon_{it}$$

The dependent variable is the math score for person *i* in year *t*, Race/Eth is a vector of six racial/ethnic categories, ELD is a vector of five categories for English language development level, and Repeat is a dummy indicating whether the person is currently repeating a grade. In the specification, we also included the math, reading, and language arts score for the student from the previous spring. For 2000-02, we used scores from the Stanford 9 test. For 2003, we used the scores on the California Achievement Test. A separate specification was estimated for each year, and the coefficients on all the covariates were allowed to vary by grade level.

The dependent variable is measured in "normal curve equivalents" (NCE). A normal curve equivalent is a linear function of test performance, which approximates the percentile for a normal distribution. If Z is a test score with mean zero and standard deviation of one, then the normal curve equivalent is calculated as NCE=50+21.06*Z. If Z is distributed normally, then NCE=1 at the first percentile of Z, NCE=99 at the ninety-ninth percentile and NCE=50 at the fiftieth percentile. One NCE point is used to approximate one percentile point.

In estimating teacher impacts, we did not control for school fixed effects. Many analysts do include school fixed effects to control for unmeasured differences between the students attending different schools. (For more on value-added specifications and their use in teacher evaluation, see McCaffrey et al. 2004; Sanders and Horn 1994; and Sanders, Saxton, and Horn 1997.) However, doing so implicitly assumes that the mean teacher quality is the same in each school. With school fixed effects, comparisons of teacher effectiveness are all made within schools. There is some evidence of systematic differences in teacher impacts ($\hat{\delta}_{teacher,year}$) between schools. However, only 5 percent of the total variation in $\hat{\delta}_{teacher,year}$ is between schools; the vast majority of the variation in estimated teacher impacts (95 percent) is observed within schools.

Similarly, many others also control for the mean characteristics of the students in each class. However, if worse teachers are assigned classes with more disadvantaged students, doing so may give too much credit to poor teachers. Controlling for school fixed effects and classroom characteristics tends to lower the estimated returns to experience and makes uncertified and alternatively certified teachers look better relative to certified teachers, since the latter tend to be assigned to poorer performing schools.

The variation in the estimated effects on student performance by teacher and year ($\hat{\delta}_{teacher,year}$) includes estimation error and other sources of nonpersistent variation in test performance, in addition to persistent differences in performance between teachers. We assume that teachers' impacts on student performance are made up of an unknown fixed effect ($\delta_{teacher}$), a return to experience ($Exper_{teacher,year} * \gamma$), and a random component, which includes estimation error and other sources of nonpersistent variation in student performance ($\varepsilon_{teacher,year}$).

To simplify, we focus on those in their first three years of teaching experience, thereby limiting the contribution of years of experience. Among those with similar years of experience, the expected value of a teacher's fixed effect, conditional on the estimated fixed effect, can be expressed as follows:

$$E(\delta_{teacher} \mid \hat{\delta}_{teacher,year}) = \mu_{\delta} + \frac{\sigma_{\delta}^2}{\sigma_{\delta}^2 + \delta_{\varepsilon}^2}(\hat{\delta}_{teacher,year} - \mu_{\delta}) = \mu_{\delta} + .57(\hat{\delta}_{teacher,year} - \mu_{\delta})$$

This is the empirical Bayesian estimator of the teacher effect, where μ_{δ} is the population mean of the teacher impacts and $\frac{\sigma_{\delta}^2}{\sigma_{\delta}^2 + \delta_{\varepsilon}^2}$ is the proportion of the total variation in that is attributable to persistent differences between teachers.

This is sometimes known as the "shrinkage" estimator, because it essentially shrinks our estimates back to the population mean. The shrinkage factor is closer to one when a larger share of the variation in the estimated teacher impacts is attributable to persistent differences between teachers; it is closer to zero when most of the variation in estimated teacher impacts is nonpersistent. We estimate that about 57 percent of the variance in teacher effects in a given year is due to persistent differences between teachers.

In figures 1, 2, and 4, we did not want to exaggerate differences between teachers by including the variation in teacher impacts that were not persistent. However, at the same time, we did not want to understate any differences between groups by "shrinking" those differences too. As a result, for each distribution portrayed—e.g., uncertified, certified, and alternatively certified teachers—we shrank back to the mean estimated impact for the relevant subgroup of teachers, using the mean of each subgroup as our estimate of $\mu^{j}{}_{\delta}$.

References

Aaronson, Daniel, Lisa Barrow, and William Sander. 2003. Teachers and student achievement in the Chicago public high schools. Working Paper 2002–28. Chicago: Federal Reserve Bank of Chicago.

Armor, David, Patricia Conry-Oseguera, Millicent Fox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the school preferred reading program in selected Los Angeles minority schools.* Santa Monica, Calif.: Rand Corporation.

Ballou, Dale, and Michael Podgursky. 2000. Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record* 102 (1): 5–27.

Blau, F. D., and M. A. Ferber. 1992. *The economics of women, men and work*. 2nd ed. Englewood Cliffs, N.J.: Prentice Hall.

Bok, Derek. 1993. *The cost of talent.* New York, NY: Free Press.

Bradley, Ann. 1999. Confronting a tough issue: Teacher tenure. *Education Week*, January 11.

Buerhaus, Peter I., Douglas O. Staiger, and David I. Auerbach. 2000. Implications of an Aging Registered Nurse Workforce. *Journal of the American Medical Association* 283 (22): 2948–54.

Bureau of Labor Statistics. 2005. Job Openings and Labor Turnover Survey. Washington D.C.: Bureau of Labor Statistics, Department of Labor. http://www.bls.gov/jlt.

Carey, Kevin. 2004. The real value of teachers. Thinking K–16 8 (1): 3–42.

Cavalluzzo, Linda C. 2004. Is national board certification an effective signal of teacher quality? Working Paper. Alexandria, Va.: CNA Corporation.

Darling-Hammond, Linda, Deborah J. Holtzman, Su Jin Gatlin, and Julian Vasquez Heilig. 2005. Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. Working Paper. Stanford, Calif: Stanford University.

Decker, Paul T., Daniel P. Mayer, and Steven Glazerman. 2004. *The effects of Teach for America on students: Findings from a national evaluation*. Princeton, N.J.: Mathematica Policy Research.

Education Trust. 2003. *Telling the whole truth (or not) about highly qualified teachers.* Washington, D.C.: Education Trust.

Ehrenberg, Ronald G., and Dominic J. Brewer. 1994. Do school and teacher characteristics matter?: Evidence from high school and beyond. *Economics of Education Review* 13 (1): 1–17.

Feldman, Sandra. 2004. Rethinking teacher compensation. *American Teacher*, March.

Goldhaber, Dan, and Emily Anthony. 2004. Can teacher quality be effectively assessed? Working Paper. Washington, D.C.: Urban Institute.

Gordon, Robert. 2005. Class struggle. *The New Republic*, June 6.

Hanushek, Eric A. , John F. Kain, Daniel M. O'Brien, and Steven G. Rivkin. 2004. Are better teachers more likely to exit large urban districts? Working Paper. Stanford, Calif: Stanford University.

————. 2005. The market for teacher quality. Working Paper 11154. Cambridge, Mass.: National Bureau of Economic Research.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2001. Why public schools lose teachers. Working Paper 8599. Cambridge, Mass.: National Bureau of Economic Research.

Hanushek, Eric A., and Steven G. Rivkin. 2004. How to improve the supply of high quality teachers. In *Brookings Papers on Education Policy: 2004*, ed. Diane Ravitch. Washington, D.C.: Brookings Institution.

Higher Education Research Institute. 2002. *The American freshman: 35 year trends.* Los Angeles: Higher Education Research Institute.

Hoxby, Caroline M. 2002. The cost of accountability. In *School accountability*, ed. Williamson M. Evers and Herbert J. Walberg. Stanford, Calif: Hoover Institution Press.

Hoxby, Carolyn M., and Andrew Leigh. 2005. Wage distortion. *Education Next* 5 (2): 50–6.

Humphrey, Daniel C., Julia E. Koppich, and Heather J. Hough. 2005. Sharing the wealth: National Board certified teachers and the students who need them most. *Education Policy Analysis Archives* 13 (18).

Jacob, Brian A., and Lars Lefgren. 2005. Principals as agents: Subjective performance measurement in education. Working Paper 11463. Cambridge, Mass.: National Bureau of Economic Research.

Jacob, Brian A., and Steven D. Levitt. 2003. Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics* 118 (3): 843–78.

Jepsen, Christopher, and Steven G. Rivkin. 2002. What is the tradeoff between smaller classes and teacher quality? Working Paper 9205. Cambridge, Mass.: National Bureau of Economic Research.

Johnson, Susan Moore, Sarah E. Birkeland, and Heather G. Peske. 2005. *A difficult balance: Incentives and quality control in alternative certification programs.* Cambridge, Mass.: Project on the Next Generation of Teachers, Graduate School of Education, Havard University.

Jupp, Brad. 2005. The uniform salary schedule. *Education Next* 5 (1): 10–2.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2005. Identifying effective teachers in New York City. Paper presented at NBER Summer Institute.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16 (4): 91–114.

————. 2005. Using imperfect information to identify effective teachers. Unpublished Paper. Los Angeles: School of Public Affairs, University of California–Los Angeles.

Koretz, Daniel M. 2002. Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources* 37 (4): 752–77.

Krueger, Alan B. 1999. Experimental estimates of education production functions. *Quarterly Journal of Economics* 114 (2): 497–532.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. Teacher sorting and the plight of urban schools: A descriptive analysis. *Education Evaluation and Policy Analysis* 24 (1): 37–62.

Levin, Jessica, Jennifer Mulhern, and Joan Schunck. 2005. *Unintended consequences: The case for reforming the staffing rules in urban teachers union contracts*. New York: The New Teacher Project.

Levin, Jessica, and Meredith Quinn. 2003. *Missed opportunities: How we keep high-quality teachers out of urban classrooms*. New York: The New Teacher Project.

Luekens, Michael T., Deanna M. Lyter, and Erin E. Fox. 2004. *Teacher attrition and mobility: Results from the teacher follow-up survey, 2000–01*. Washington, D.C.: National Center for Education Statistics, Department of Education.

McCaffrey, Daniel F., Daniel M. Koretz, J. R. Lockwood, and Laura S. Hamilton. 2003. *Evaluating value-added models for teacher accountability*. Santa Monica, Calif.: Rand Corporation.

Miller, Matt. 2003. *The two percent solution.* New York: Public Affairs.

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

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

Murnane, Richard J., John B. Willett, and Frank Levy. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77 (2): 251–66.

National Center for Education Information. 2005. *Profile of teachers in the U.S. 2005*. Washington, D.C.: National Center for Education Information.

National Center for Educational Accountability. 2005. Results of 2005 NCEA survey of state data collection issues related to longitudinal analysis. Washington, D.C.: National Center for Educational Accountability. http://www.dataqualitycampaign.org/activities/survey_result_2005.cfm.

National Education Association. 2005. Time is almost up! Do you meet the No Child Left Behind "highly qualified" teacher rules? Washington, D.C.: National Education Association. http://www.nea.org/esea/qualification/teacher/index.html.

Neal, Derek A., and William R. Johnson. 1996. The role of premarket factors in black-white wage differences. *Journal of Political Economy* 104 (5): 869–95.

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges. 2004. How large are teacher effects? *Educational Evaluation and Policy Analysis* 26 (3): 237–57.

Podgursky, Michael, Ryan Monroe, and Donald Watson. 2004. The academic quality of public school teachers: An analysis of entry and exit behavior. *Economics of Education Review* 23 (5): 507–18.

Prince, Cynthia D. 2002. Higher pay in hard-to-staff schools: The case for financial incentives. Arlington, Va.: American Association of School Administrators.

Public Agenda. 2003. *Stand by Me*. New York: Public Agenda.

Raymond, Margaret, Stephen H. Fletcher, and Javier Luque. 2001. *Teach for America: An evaluation of teacher differences and student outcomes in Houston, Texas.* Stanford, Calif.: Center for Research and Education Outcomes, Hoover Institution, Stanford University.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools and academic achievement. *Econometrica* 73 (2): 417–58.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94 (2): 247–252.

Rotherham, Andrew J. 2005. Credit where it's due: Putting nationally certified teachers into the classrooms that need them most. *Education Week*, March 30.

Rothstein, Richard. 2004. *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap.* Washington, D.C.: Economic Policy Institute.

Roza, Margeurite and Paul T. Hill. 2004. How within-district spending inequities help some schools to fail. In *Brookings Papers on Education Policy*: 2004, ed. Diane Ravitch. Washington, D.C.: Brookings Institution.

Sanders, William L., and Sandra P. Horn. 1994. The Tennessee Value-Added Assessment System (TVAAS): Mixed-model methodology in educational assessment. *Journal of Personnel Evaluation in Education* 8 (3): 299–311.

Sanders, William L., Arnold M. Saxton, and Sandra P. Horn. 1997. The Tennessee Value-Added Assessment System (TVAAS): A quantitative, outcomes-based approach to educational assessment. In *Grading teachers, grading schools: Is student achievement a valid evaluation measure?*, ed. Jason Millman. Thousand Oaks, Calif.: Corwin Press.

Saunders, Debra J. 2005. A vote to end the teacher shuffle. *San Francisco Chronicle*, September 18.

Snyder, Thomas D., Alexandra G. Tan, and Charlene M. Hoffman. 2004. *Digest of education statistics 2003*. Washington, D.C.: National Center for Education Statistics, Department of Education.

Staiger, Douglas O., David I. Auerbach, and Peter I. Buerhaus. 2000. Expanding career opportunities for women and the declining interest in nursing as a career. *Nursing Economics* 18 (5): 230–6.

Summers, Anita A., and Barbara L. Wolfe. 1977. Do schools make a difference? *American Economic Review* 67 (4): 639–52.

Toch, Thomas. 2005. Measure for measure. *The Washington Monthly*, October/November.

Vandevoort, Leslie G., Audrey Amrein-Beardsley, and David C. Berliner. 2004. National board-certified teachers and their students' achievement. *Education Policy Analysis Archives* 12 (46).

Walsh, Kate. 2005. Merit pay: Not so fast, governors! *The Education Gadfly* 5 (27).

Walsh, Kate, and Emma Snyder. 2004. *Searching the attic: How states are responding to the nation's goal of placing a highly qualified teacher in every classroom.* Washington, D.C.: National Council on Teacher Quality.

Acknowledgments

The authors thank Jason Bordoff, Cindy Brown, Michael Cohen, Michael Deich, Kati Haycock, Richard Kahlenberg, Goodwin Liu, Richard Murnane, Ann O'Leary, Peter Orszag, Mike Petrilli, Michelle Rhee, Andrew Rotherham, Richard Rothstein, Ross Wiener, and Amy Wilkins for comments. Timothy Taylor suggested a number of helpful revisions to an earlier draft. Amanda Major and Jerilyn Libby provided excellent research assistance.

Authors

ROBERT GORDON

Senior Vice President for Economic Policy, Center for American Progress

Robert Gordon is a Senior Vice President at the Center for American Progress. He was the domestic policy director for the Kerry-Edwards campaign and the policy director for John Edwards during Edwards' presidential primary campaign and in the Senate. Gordon was previously a law clerk for Justice Ruth Bader Ginsburg, a Skadden fellow at the Juvenile Rights Division of the Legal Aid Society, and an aide at the National Economic Council during the Clinton administration. Gordon is a graduate of Yale Law School and Harvard College.

THOMAS J. KANE

Professor, Harvard Graduate School of Education

Thomas J. Kane is Professor of Education and Economics at the Harvard Graduate School of Education and faculty director of a new center for school research at Harvard. In his work, he has focused on both higher education and elementary/secondary education. In higher education, he has studied the labor market payoff to a community college education, the impact of tuition and financial aid policy on college enrollment rates, and the impact of affirmative action in college admissions. His book on state and federal financial aid policies, *The Price of Admission: Rethinking How Americans Pay for College*, was published by the Brookings Institution in 1999. Recently, he has been studying the design of accountability systems in elementary and secondary education and teacher impacts on student achievement in Los Angeles and New York. Before moving to Harvard, Kane was a professor of public policy at UCLA. In addition, he has been a visiting fellow at the Brookings Institution and at the Hoover Institution at Stanford University. In 1995–96, he served as the senior staff economist for labor, education, and welfare policy issues within President Clinton's Council of Economic Advisers.

DOUGLAS O. STAIGER

Professor, Dartmouth College

Douglas O. Staiger is a Professor of Economics at Dartmouth College and a Research Associate of the National Bureau of Economic Research. Staiger's work focuses on school accountability, the quality of medical care, and the labor market for nurses.

THE HAMILTON PROJECT

ADVISORY COUNCIL

THE HAMILTON PROJECT

The Brookings Institution
1775 Massachusetts Ave., NW, Washington, DC 20036
(202) 797-6279 ▪ www.hamiltonproject.org

The Brookings Institution