STAND™ for children
LEADERSHIP CENTER

EDUCATION POLICIES AND PRACTICES AND THE QUALITY OF THE TEACHER WORKFORCE:

An Update

by Dan Goldhaber

A White Paper from
Stand for Children Leadership Center

June 2013

Introduction

The importance of teacher quality for student achievement is now a well-documented and widely cited fact. Decades-old research[1] has been augmented by a tremendous amount of new empirical evidence[2] showing that teacher effectiveness, as measured by impact on student learning, is the most important school-based factor when it comes to improving student achievement. When students have effective teachers, the results are dramatic—children with more effective teachers can gain up to a year's worth of learning compared to peers with weaker teachers.[3] While much of the literature on the value of teacher effectiveness measures teachers based on their impact on student test scores, new evidence[4] suggests that teacher effectiveness also influences students' later life outcomes, such as college-going behavior and earnings.

It is no surprise that the quality and distribution of the teacher workforce have moved to the forefront of education policy.[5] This paper provides a brief overview of what is known about the teacher pipeline and about the type of college graduates who have historically entered the teaching profession. It also addresses the research on various ways to affect teacher quality: changing the makeup of the teacher workforce through recruitment, selection, and deselection policies;[6] and improving the effectiveness of current teachers.

Effective Teaching

Understanding of Effective Teaching Is Increasing—But Many Questions Remain

The past decade has yielded a tremendous amount of research on teacher quality, much of it enabled by the year-over-year testing associated with state-accountability systems enacted as part of the No Child Left Behind Act. Year-over-year testing allows statisticians to use statistical techniques to try to distinguish the contribution that teachers make to student-achievement gains on standardized tests from other student, family, or schooling factors that might also influence student achievement. These direct measures of teacher effectiveness have shown just how important it is to have a high-quality teacher, and how much teacher quality varies from school to school and classroom to classroom. However, because the data to construct value added measures of teachers are relatively new, they cannot always answer questions about longer-term trends in the teacher workforce and its distribution. Available evidence related to these questions is necessarily based on readily quantifiable teacher attributes, such as teaching experience or where teachers received their degrees (often referred to as "input-based" measures).

The Teacher Pipeline, and Long-Term Trends on Who Becomes a Teacher

Judgments about the long-term trends in the quality of the teacher workforce are based on academic measures, such as teachers' college entrance-exam scores and the selectivity of the colleges from which they graduate.[7] The relatively few studies on long-term changes in teacher quality generally find (based on different measures) a modest decline since the early 1960s, relative to other occupations.[8]

The average changes in the academic quality of the teacher workforce obscure a more dramatic drop-off in the probability that individuals who perform well academically end up teaching. The likelihood that a female teacher was from the top decile of her high-school graduating class—judged by standardized-test scores—dropped by more than half from 1964 to 2000, from about a 20 percent probability to about a 10 percent probability. [9] From 1963 to 2000, the share of all teachers from the lowest-aptitude group rose from 16 to 36 percent; the share from the highest-aptitude group fell from 5 to 1 percent. [10]

Several factors that are not necessarily mutually exclusive may explain the declining academic caliber of the teacher workforce. First, the labor-market opportunities for women have expanded significantly since 1960—meaning other industries are competing for talent with what was, and remains, a female-dominated profession.[11] Second, teaching may be a less financially attractive occupation than it once was. The percent of college graduates earning less than the average teacher fell for women from about 55 percent in 1950 to 45 percent in 2000.[12] It is also possible that the structure of compensation in education—which, relative to other occupations, often compresses wages toward the mean—tends to discourage the most academically proficient individuals from a career in teaching, because they can command a higher salary in other sectors.[13] Finally, selection amongst applicants for teaching positions may not strongly favor those with better academic credentials.[14]

The long-term trends in teacher quality are reflected in recent studies that follow the teacher pipeline. High school seniors intending to major in education score lower on college entrance exams than their peers,[15] and entrance exam scores are lower for those who prepared to teach, were teaching, or were considering teaching than for other undergraduate students.[16] Additionally, teachers with stronger academic records are more likely to leave the profession than their peers.[17] A now two-decades-old quote from Murnane and Singer—"College graduates with high test scores are less likely to take jobs, employed teachers with high test scores are less likely to stay, and former teachers with high test scores are less likely to return"—is an apt summary of the recent academic findings on the teacher pipeline. By contrast, in countries whose students outperform the United States in international assessments, the teaching force tends to be comprised of individuals drawn disproportionately from the upper end of the academic-performance distribution.[18]

These findings are a concern, given that at least moderate evidence supports the notion that

"smarter" teachers tend to be more effective.[19] Some direct empirical evidence suggests that large increases in salaries have an impact on the academic caliber of the teacher workforce.[20] Unfortunately, there is relatively little direct evidence about how much salaries would need to increase to draw more academically talented people into the teaching profession. Market research by McKinsey shows that significantly increasing the percentage of top-third college graduates who enter teaching would be costly; for instance, the findings indicate that an increase of 15 percent in the top third would require starting salaries of about $65,000, which is much higher than today's average of about $40,000.[21] The implication is that reversing the trends described above would almost certainly require either significant new K–12 educational investments or a major restructuring of the way teachers are paid.

Licensure and Alternative Pathways Into the Teaching Profession

States try to guarantee a basic level of teacher competence by regulating who is able to teach, through licensure and certification. Research on teacher licensure as a means of quality control goes back more than a decade, with many studies predating the growth of well-defined alternative pathways into the profession.[22] Few of the studies from this early literature focused on student outcomes, and those that did found little systematic evidence that students taught by fully licensed teachers outperformed those taught by teachers holding provisional or emergency credentials.[23] This is perhaps not terribly surprising, since state licensure policies (both traditional and alternative) vary.

Individual teachers choose to participate in either traditional or alternative preparation routes, and that choice may be shaped by personal factors that also influence effectiveness. Consequently, it is not easy to deduce the extent to which differences in teacher effectiveness across routes into the classroom are a result of program selection versus differences in pedagogical training. A 2010 report from the National Research Council concludes that we know relatively little about how specific approaches to teacher preparation are related to effectiveness in the field.[24]

More recent research on licensure tends to use data from large administrative state databases to compare fully licensed teachers with those who hold a particular alternative credential. Alternative licensure programs allow individuals interested in teaching to move into the classroom more quickly and easily, and they are based on the idea that allowing people to bypass or postpone at least some of the coursework requirements associated with traditional licensure programs may increase the pool of high-quality applicants.[25] As is the case with traditional licensure, there is no single alternative licensure policy.

Teach for America, arguably the best-known alternative pathway to the classroom, has received a great deal of research attention. The research on TFA teachers' impacts on student achievement shows pretty consistent evidence that TFA teachers compare favorably with other teachers[26]—they are, on average, as good as or more effective than other teachers in the same schools.[27]

But Teach for America does not represent most alternative pathways, because its teachers represent a very selective group in terms of academic preparation.[28] Thus, findings for TFA do not necessarily generalize to other alternative programs. One recent study found little difference between the test achievement of students whose teachers received traditional training and the test achievement of students whose teachers entered the profession through alternative routes.[29] This study also finds little relationship between the amount, or content, of teacher training coursework and student achievement.

In general, differences among individual teachers entering the profession through a specific route swamp the average differences that exist between cohorts of teachers entering through different routes. In comparing teachers who followed a variety of pathways into New York City schools—including participants in traditional preparation and licensure, Teach for America, the New Teacher Project Teaching Fellows program, and teachers who were not licensed—researchers have found that variations in effectiveness among teachers who followed the same pathway far exceed differences in average effectiveness of teachers from different pathways.[30]

The clear implication of the research on licensure is that the current system of qualifying teachers is not terribly predictive of effectiveness in the classroom. In other words, one can learn far more about teachers from their in-service performance than from the pathway that brought them into service.[31] Does this mean policymakers should not focus on improving state licensure systems? This is a judgment call, but the above findings suggest that improved evaluation systems for in-service teachers would better inform human-capital decisions.

Variation in Effectiveness of Graduates of Traditional Teacher Preparation Programs

Policymakers have recently begun to focus more attention on the role that teacher training may play in influencing student achievement.[32] Several studies have explored the possibility that the teacher education program in which a teacher is prepared may be an important signal of future effectiveness. For the most part these studies mirror the findings of research on pathway into the teaching profession, finding that the variation in effectiveness of teachers who graduate from the same program is far greater than the than variation between different programs.[33] But while the studies suggest that training programs explain only a small portion of the variation in teacher effectiveness, this does not mean that there are no educationally relevant differences between some training programs. For instance, some studies show that

the larger differences between programs can be roughly equivalent in magnitude to the size of the achievement gap between students eligible for free or reduced price lunch or the difference between a novice teacher and one with five or more years of experience.[34] The size of these differences suggest that hiring officials could, in at least some cases, use the program a student graduated from as a meaningful signal about the prospective teacher's future effectiveness. They also suggest that states might want to use the estimates as a trigger to look more deeply at the practices (selection and training) of different programs,[35] but holding teacher preparation programs accountable for the student outcomes of the teachers they produce is conceptually complex.

There are at least three issues that arise when it comes to thinking about how to interpret differences in the value added effectiveness of teachers who graduate from different programs. First, it is quite difficult to definitively separate the impact of selection into training programs from the impact of the preparation itself. Some programs, for example, have grade point average requirements for admission that are much higher than those of other programs. This distinction between selection and training effects may not be relevant for school districts, who may care about differences in the effectiveness of teachers from different programs but not whether it is due to selection or training. But the distinction is critical if we wish to learn about how to improve teacher training. Second, it can be difficult to separate the impact of teacher training from the schools into which teachers from different programs tend to work. In some cases, there are strong labor market feeder patterns where graduates from particular teacher training programs end up employed in particular schools or districts. Interestingly, however, the nature of feeder patters appears to differ across states. For example, studies in Florida and Washington state show significant differences in the extent to which training institutions feed primarily into local teacher labor markets (this appears to be much more common in Florida than Washington).[36]

Finally, many programs produce so few teachers in a year that it is not possible to assess with much precision how effective they are based only on recent graduates. This is a particular challenge when it comes to thinking about program accountability, because differences in the amount of data available for larger versus smaller programs might influence how they appear in some accountability metrics. I would argue that, given these conceptual challenges, states need to think carefully about how the value added information that could be used for training program accountability purposes can be balanced against other information about training programs. In essence, the situation when it comes to training programs is not that different from performance evaluations of individual teachers (discussed in more detail below): we know that there may be important differences between programs but value added along does not provide much information about the source of those differences (e.g. is it that some program simply enroll more talented students or is it about the training?), nor does it typically provide the kind of nuanced information that might aid in program improvement. The flip side, however, is that today's system for accrediting training program does not appear to be very rigorous.[37] Hopefully the availability of value added information about programs helps to drive policymakers toward a more thorough vetting of the nation's teacher training infrastructure.

Inequities in the Distribution of Teachers for Low-Income and Minority Students

A large body of evidence shows that readily quantifiable teacher qualifications—such as experience, degrees, licensure status, and test scores—are not equitably distributed across schools or students.[38] Poor and minority students, and those who tend to do less well academically, are far less likely to be taught by more-experienced, credentialed teachers, due to the sorting of teachers both among schools and among classrooms within schools.

Some might argue that the unequal distribution of teachers based on observable qualifications is not a big concern, given that the characteristics and credentials used by most states and school systems to determine employment and compensation are not strongly related to students' test outcomes.[39] This is the wrong conclusion for two reasons. First, there is very clear evidence that early-career teacher experience is one of the few qualifications that does predict effectiveness, so the fact that inexperienced teachers are not equitably distributed across students does imply that poor and minority students do not have equitable access to effective teachers. Second, it is quite likely that if these observable qualifications are unevenly distributed, then unobserved, but educationally important, teacher qualities are also unevenly distributed. There is in fact some evidence to support this notion. Several recent papers using direct, value added measures of teachers shows that teachers in higher-poverty schools tend to be somewhat less effective than those in lower-poverty schools, and that the average differences in effectiveness at the school level are driven by a greater variation in effectiveness (with more ineffective teachers) in high-poverty schools.[40]

The clear implication of the findings on teacher distribution is that the teacher labor market, as it currently operates, does not equitably distribute teachers. It is rare, for instance, for school districts to explicitly compensate teachers for the difficulty of their jobs, and teaching disadvantaged students is likely, in many cases, to be a more difficult job than teaching students who face fewer educational and life challenges.[41] Given this, it is not surprising to see public-policy efforts, such as the federal Teacher Incentive Fund, that have an explicit goal of promoting more equitable distribution of teacher quality. Evidence shows that pay differentials designed to keep teachers in disadvantaged schools have an impact on teacher retention.[42]

Teacher Recruitment and Selection

There is relatively little quantitative evidence of the efficacy of different recruitment and selection practices. Most data are derived from surveys, or focus on individuals who are already employed as teachers (meaning we do not observe people who do not choose, or are not chosen, to teach). The scant evidence that exists presents a mixed picture of whether school systems hire the most-capable applicants, at least based on the academic attributes of the applicants.

But while there is only sparse quantitative literature investigating the potential link between schools' recruitment and selection practices and teacher quality, there is a growing consensus in public- and private-sector management research that recruitment and selection practices play an important role in the quality of employees and, ultimately, organizational performance.[43] Research comparing practices in public education to those in the private sector suggests that recruitment and selection policies (as well as other human-resource practices) in school districts could be vastly improved.[44]

Some of the shortcomings of common district practices have been well documented. Teacher labor markets tend to be of limited geographic scope—school systems typically do not look far and wide for teaching talent.[45] School systems are also rather unsophisticated when it comes to screening teachers. Older literature, based on school surveys on hiring criteria from the late 1980s and early 1990s, finds that when making hiring decisions, localities place an emphasis on education-specific credentials.[46] Some school systems use packaged selection tools, such as the Gallop Teacher Insight Assessment or Haberman's Star Teacher Selection Interview, but these types of assessment tools have not been proven to predict teacher effectiveness.[47] One recent study did assess teacher-selection instruments, and found a small positive relationship to student achievement (the researchers collected information on both commonly used and nontraditional teacher traits and characteristics, including teaching-specific content knowledge, cognitive ability, personality traits, feelings of self-efficacy, and scores on a teacher-selection instrument). But it is important to put this in context: collectively, all of the teacher traits and characteristics considered in this study account for only about 10 percent of the variation in teacher effectiveness.[48]

Another recent study analyzes the association between measures collected by Teach For America in the application and interview processes associated with making admission decisions and the future achievement of students in TFA classrooms.[49] This study finds a fairly strong statistically significant relationship between some pre-service measures (linked to measures of achievement, leadership, and perseverance) and student achievement gains in math (the study finds a positive but not statistically significant relationship for student achievement in reading). These findings are encouraging and suggest there may be benefits associated with collecting and selecting on more nuanced traits of teacher applicants, but it is important to keep in mind that TFA applicants are quite different, on average, from typical pre-service teachers so these findings may not be generalizable.

Finally, the timing of district hires may affect the quality of the workforce. Research by the New Teacher Project finds that many districts lose high-quality candidates to more-nimble competitors because they hire late (many large urban districts, in particular, push hiring into the late summer).[50] This finding is confirmed by recent research showing that a significant proportion of teachers (33 percent in a four-state sample) are hired after the school year has already begun.[51]

Late hiring is driven both by state policy and by collective-bargaining agreement constraints. For instance, in 46 states, the state's budget for the coming fiscal year is not completed until June, raising considerable uncertainty at the local level about state resources available to fund teaching positions. Provisions allowing for very late notice of intent to retire or resign, and seniority transfer rights for incumbent teachers, are hardwired into many collective-bargaining agreements. Uncertain budgets and late vacancy notices likely make districts reluctant to hire. And while seniority transfer rights do not directly impact the aggregate number of positions available in a district, they do mean that districts may not be able to tell prospective teachers in which school they would be employed. This may not be a big issue in a small suburban district, but could play a large role in whether teachers accept an offer in a large urban district, where teachers would like to know where they would be teaching when evaluating a job offer.

Late hiring is far more common than one might guess. Estimates suggest that it is not atypical for a significant proportion of teachers to be hired after the school year has already begun.[52] There are few quantitative studies on the impact of late hiring, but the emerging literature buttresses the notion that late hiring is problematic. For example, a recent study of teachers in Michigan, for instance, finds that teachers that are hired after the school year ("late hires") are substantially less likely to remain in the same school after one year, and substantially more likely to leave the teaching profession.[53] And, moreover, the late hiring phenomenon is likely to be more prevalent in low-income and low-performing schools.

A second study on late hiring finds that the student achievement of late hires is substantially lower than the achievement of students in the classrooms of other newly hired teachers who were employed before the beginning of the school year.[54] And, the evidence suggests that the lower level of effectiveness of late hires is due both to late hires being less effective teachers and because there is a disruption effect of being hired late, i.e. there is some evidence that late hiring is related not only to the productivity of those hired late but also to the productivity of other teachers in the schools where late hiring occurs.[55]

Improving the quality of the teacher workforce through recruitment and selection policies and practices is a promising area for reform. Although there is not much in the way of quantitative evidence on different recruitment and selection processes, the recent evidence on late hiring and its causes identifies some low-hanging fruit, especially for large urban systems. Moreover, changes to state policies could allow districts to plan better and begin recruiting teachers earlier. To some extent, changes in recruitment and selection practices might just mean a reshuffling of

where teachers end up employed, since school districts are competing with one another, but better recruitment and selection strategies may also help to reduce the public schools' loss of potential teachers to other employers that hire earlier. Because high-poverty urban districts tend to have less effective recruitment and selection policies and practices than other districts, improvement in this area may lead to more equitable distribution of effective teachers for low-income and minority students.

Teacher Evaluation

Most early-career (pre-tenure) teachers receive a yearly evaluation, while more-experienced (post-tenure) teachers are typically evaluated at least once every three years—but many of these evaluations are worthless. They fail to differentiate teachers and usually provide them with little substantive feedback on classroom practices.

The existing infrastructure for in-service evaluation is weak, with teacher evaluations mainly consisting of quick classroom visits (sometimes referred to as "drive-bys") by principals or other school administrators.[56] Many performance-rating systems are rudimentary, and insensitive to differences in the contexts (subjects, types of students, grade level, etc.) in which teachers work. They typically use only a binary scale whereby teachers are judged to be either "satisfactory" or "unsatisfactory."

The results of the perfunctory evaluations that exist in most school systems are documented in *The Widget Effect*, by The New Teacher Project.[57] This widely cited publication finds that formal teacher evaluations provide very little information about teachers. The reason is that evaluations almost universally suggest that teachers are the same.[58] For example, more than 99 percent of teachers in districts with a binary evaluation system are rated "satisfactory."[59] There is no national teacher-evaluation database indicating how many school systems employ a binary rating system, but the data from *The Widget Effect* and other reports suggest that binary systems are quite common.[60] A binary measure obviously does not enable evaluators to make nuanced judgments about performance. However, an evaluation instrument that allows for more differentiation is in itself not sufficient, given political and cultural constraints. Even in districts with a broader range of rating options, most teachers (94 percent) receive one of the top two ratings.

It is not surprising, given the widespread perception that teacher evaluation is currently broken, that policymakers are actively pursuing evaluation reform.[61] Ultimately, the point of changing evaluation systems is to influence teacher behavior—but we currently know very little about how teachers will respond to changes in the evaluation system, and we are unlikely to know more until those changes have actually been implemented.

The great majority of evaluations today—probably well over 90 percent—are based on teacher observations.[62] As discussed above, current observation-based evaluations do not in general differentiate teachers from one another. However, there is a growing literature on the ability of formal, "high quality" classroom observations (based on a specific observational rubric) to predict gains in student achievement.[63] For example, when students have a teacher with a one-point-higher rating—equivalent to moving up one step in the rating categories of "unsatisfactory," "basic," "proficient," and "distinguished"—they see significant improvement in their achievement.[64]

Value Added and the Potential Challenges to Using It

Policy interest is growing in using direct measures of student-achievement growth as a component of teacher evaluations. These measures have been shown to predict student achievement better than other readily observable credentials and trait.[65] Also driving policymaker interest is the recognition most traditional evaluation systems fail to adequately differentiate teachers. Student-growth measures may also have an important "honest broker" role when it comes to providing feedback to teachers about their performance.[66] For instance, principals may be more comfortable having tough conversations about performance with their teachers when student-growth information confirms their impressions about teachers' classroom skills.

Student-growth metrics come in a variety of forms, from simple year-over-year growth in the achievement of individual students, to more-complex models that adjust growth for what students are expected to achieve given their background (race/ethnicity, or free/reduced-price lunch status), typically referred to as a value added model (VAM) approach. (For more information on the way that different models handle the translation of student growth into measures of teacher effectiveness, see a recent paper by Dan Goldhaber, Brian Gabele, and Joe Walch.)[67]

Different measures or models may be appropriate for different purposes or goals. Simple growth measures, for instance, may provide quick feedback to teachers that can be used to inform the types of professional development they receive. But schools might want to use more-complex value-added measures, utilizing multiple years of data, to help inform tenure decisions. The existing evidence comparing growth measures, such as the "Colorado Growth Model," with VAMs shows a significant positive correlation between the two, but this does not necessarily

mean that the measures would not diverge for teachers serving particular types of students (for example, high-poverty kids or English-language learners).[68]

Regardless of the form, the use of student assessments for teacher evaluation is controversial. Critics of using value added point to three concerns associated with the *measure* itself: 1) the potential that value added measures of teacher performance are biased (unfair) to teachers because they do not fully account for the way in which students are sorted into teachers' classrooms;[69] 2) the potential that the value added measures of teachers will fail to capture important ways that teachers contribute to different measures of students' cognitive or social growth; and 3) the potential that value added measures "fade out" over time because they reflect only short run teacher efforts to improve test scores (e.g. "teaching to the test.

Recent studies speak to all of these issues. Several papers use sophisticated experimental and non-experimental tests to address the potential that value added measures produce biased (i.e. not valid) estimates of true teacher effectiveness. These generally confirm that VAMs, when properly specified (i.e. include adequate controls for differences in students' backgrounds) are not biased, for instance, in favor of teachers with more gifted students.[70] It is worth noting, however, that the validation of value added is generally based primarily on data from elementary and middle schools, and there is evidence that identifying teacher effectiveness at the high school level may be more difficult due to the fact that students are more likely to be tracked at the high school level.[71]

There is a fair amount of evidence on whether VAM measures fail to capture different measures of students' cognitive growth. This is inferred from studies that assess cases where the same students in teachers' classrooms take different standardized achievement tests. And these studies do find that VAM measures of teacher effectiveness are somewhat sensitive to the student tests that are used, some of which may pick up the content on which teachers focus better than others. In general, teacher effectiveness ratings are similar across different student assessments and on different growth models, but there may be important differences for individual teachers, and policymakers should be mindful of these trade-offs.[72]

There is less information on the extent to which teachers contribute to non-cognitive outcomes (e.g. disruptive behavior), shown to be an important predictor of later life success. But here too there is some (very nascent) evidence that value added measures are far from perfect when it comes to measuring a teacher's contribution to student social growth. Specifically, a recent study finds that teachers contribute to both cognitive and non-cognitive student outcomes, and that there is that there is a positive correlation between a teacher's contribution to cognitive and non-cognitive student outcomes.[73] But while the estimated correlation is positive, it is relatively weak, meaning that focusing exclusively on measuring value added would run the risk of missing teacher contributions to student non-cognitive outcomes. This is one of the reasons why one might not want to rely overly much on value added alone as an indicator of teacher performance (others are described in the next section).

Perhaps the biggest concern about value added is that it has been found to "fade out" over time. A number of studies have shown that a significant proportion of the test score gains produced by teachers in one grade seem to dissipate in the next couple of grades, such that estimates find less than half of the value added associated with one teacher is detectible two years after having that teacher. "), This "fade-out" finding might imply that teachers judged to be more effective according to VAM might not actually produce student gains to benefit their students in the long run.

Research on the potential causes of teacher fade out is in its infancy. It is certainly conceivable that some of the fade out is associated with teachers narrowly focusing on students' test score gains,[74] but there are also more benign explanations having to do with the way tests are scaled across grades and the content they cover.[75] There is also some good news for those who believe that focusing on teacher effectiveness is the right policy emphasis. A recent study that finds evidence of value added fade out over a couple of grades also finds that VAM measures predict student outcomes later in life, such as college-going and labor market earnings.[76] In other words, despite short-run evidence that teacher effects fade out according to test gains, there is longer-term evidence that they matter in terms of the outcomes we really care about.

There are also issues of concern when it comes to the use of VAM for high-stakes purposes. For instance, VAM critics raise concerns that high-stakes teacher evaluations based on student growth will lead to the misclassification of teachers.[77] In the case of misclassification, there are two related but distinct concerns: the validity of the estimates produced by VAMs and their stability/reliability. "Validity" refers to the extent to which the measure, in this case value added, reflects the truth about an underlying concept—that is, teacher quality. More simply, if we had lots of information (e.g. many years of data) about the achievement of students with particular teachers, would value added provide an accurate estimate of teachers' contribution to student learning? If the answer is yes, we would infer that value added is valid. As discussed above, there is a growing empirical literature (e.g. the Chetty et al. paper) on the validity of VAM measures. "Reliability" refers to the extent to which repeated measures of a concept will yield the same result. For instance, if we have value added for teachers across different years or classroom sections, does it tend to tell us the same thing about teacher quality (whether or not it is an accurate measure of the true concept). Several new studies address the stability/reliability of VAMs.[78]

A measure will lead to classification errors (teachers being judged as effective when they are not, and vice versa) if either the validity or the reliability of the measure is low. In general the number of classification errors will be reduced if there is more information about teachers; this may entail more years of data and/or information from different sources (e.g. value added combined with classroom observations and student perceptions).[79] But, the bottom line is that there may well be tradeoffs between validity and reliability. Our current evaluation system yields very reliable information, in the sense that teacher performance is judged to be the same year after year, but it is not necessarily valid. While there is disagreement about how or whether school systems should incorporate VAMs into teacher evaluations, a recent Brookings Institution

report concludes that VAM effectiveness estimates should at least be used to inform teacher evaluations because, while they may seem at first glance to be unstable, they are in fact about as stable as measures used for high-stakes purposes in other sectors of the economy. More important, they tend to better predict student achievement than other measures like experience and credentials that districts typically use to determine employment eligibility and compensation.[80] Moreover, as I describe more thoroughly in the next section, the predictive power of value added if also far higher than that of classroom observations if one is trying to predict student achievement on state assessments.

When it comes to using student growth, there is also no right answer about what model or measure is most appropriate or how much weight should be given to student growth–based measures.[81] We are ultimately concerned with the behavioral responses to the way teacher evaluations are used. In other words, in the absence of policy experimentation, we cannot know how teachers will respond to a new evaluation system and the policies linked to that system. This is, if anything, an argument for state and local policymakers to implement different models and carefully evaluate their results to build our knowledge base and ultimately move toward more effective systems.

Any educator evaluation system will involve trade-offs between transparency and accuracy. For example, a transparent, easily understood performance measure that teachers may be more likely to trust could result in greater teacher effort. A more accurate model that is less transparent may be easier for teachers to dismiss as simply statistical mumbo jumbo. That said, policymakers may wish to exercise caution with some methods of translating student growth into teacher-effectiveness estimates that do not measure confidence in the system (that is, no standard errors are associated with the estimate of teacher performance), as this may be legally problematic in cases where the performance measure is used as a factor in making high-stakes personnel decisions.[82] In other words, such a methodology may show that two teachers are different (they get a different "score" or ranking for improving student achievement), despite the fact that we do not know much about whether the estimated differences are statistically meaningful. As an example, we would not want to draw strong inferences about two teachers with class sizes of five or 10 students, but we would be on much firmer statistical ground if the two teachers were teaching classes in the 30-to-35-student range. This same idea applies to the number of years of data that are used to assess teacher effectiveness: more years and more students generally mean more confidence in the teacher measure.

Of course, only a slice of the teacher workforce can be evaluated based on state assessments in English language arts and math, so states and localities enacting evaluation reform necessarily must struggle with how to evaluate teachers in non-tested subjects (as well as what types of additional assessments are appropriate in tested subjects). Unfortunately, information is lacking about many of the other forms of teacher evaluation, including, peer, self-, and/or student ratings

of instruction. But given the limitations of any one evaluation method,[83] there are benefits to using multiple methods.[84]

Other Teacher Quality Measures

Given the increased interest and use of value added measures of achievement, it is natural to wonder how these compare to the increasingly diverse set of other measures that states and districts are using to make judgments about teacher effectiveness. A number of new studies speak to this issue. The most prominent among these studies is the Measure of Effective Teaching (MET) study, funded by the Bill & Melinda Gates Foundation. This comprehensive project is designed to assess the various means of evaluating teachers, whether they are valid, and the extent to which they correspond with one another and predict student achievement singly or in combination.[85]

A recent MET report shows the relationship between various methods of evaluating teachers—ratings based on videotaped observations, student surveys, and value-added assessments —and student achievement on standardized tests (in math and reading/English Language Arts).[86] The findings suggest that a number of well-known teacher-observation instruments (Charlotte Danielson's Framework for Teaching, the Classroom Assessment Scoring System, etc.) are moderately positively correlated with student achievement, and value added measures of teacher effectiveness.[87] The finding that value added is correlated with observation protocols is confirmed at the middle school level by a second recent study that focuses on student achievement in English Language Arts (ELA).[88] And one of the striking findings in this study is that a teacher's use of "Explicit Strategy Instruction" appeared to be particularly important for predicting her value added.[89] This finding is important not only because finding interventions that work in the case of student achievement at the middle school level, in ELA in particular, has been challenging, but also because few teachers tend to use Explicit Strategy Instruction.

The use of student surveys to assess teachers is far newer than classroom observation as a means of teacher assessment, but the MET study also shows that these assessments (measured by the Tripod survey) are even more strongly correlated with value added than classroom observations.[90] And finally, some evidence indicates a correlation between another means of assessing teachers—student-growth objectives (which are currently in place in some states and districts) and student achievement. One study, for instance, finds that a large share of teachers who set student growth objectives (also commonly referred to as student learning objectives), around 70 percent, but also that teachers who are successful in meeting a student growth objective (and where those who meet this objective receive a financial bonus) are more effective

in raising student achievement on a state assessment in both math (by about 6 percent of a standard deviation of student achievement) and reading (by about 3 percent of a standard deviation of student achievement).[91]

Not surprisingly, as the MET project shows, measures of teaching effectiveness are generally best predicted by a comparable baseline measure (e.g. if the objective is to predict teacher classroom observations scores, then a prior measure of a teacher's performance on a classroom observation rubric is better than student perceptions survey results or value added). Thus, when it comes to predicting student achievement on state assessments, nothing works as well as value-added measures. Interestingly, however, composite measures of teacher effectiveness that combine value added with teacher observation and student survey measures significantly outperform value added measures alone when the objective is to predict student achievement on assessments other than the state test.[92]

The bottom line from the MET study is that using multiple measures provides some additional information about teacher effectiveness.[93] Is the cost of using multiple methods worth the added information they provide? The answer to this question involves making a value judgment, but it is worth considering the limitations of VAMs: They cannot be used for the majority of teachers whose students do not have state assessment data and, importantly, they probably cannot be used to provide teachers with timely or concrete feedback about their teaching practices. These limitations may suggest benefits to supplementing value-added measures with additional well-designed and research-supported measures.

In-Service Policies to Increase Teacher Effectiveness: Professional Development, Mentoring, and Feedback

Professional development is a nearly universal strategy for improving teaching, but there is little evidence that professional development, as currently devised, discernibly affects student achievement. The "as currently devised" caveat is an important one. We do not know whether professional development efforts appear to be unsuccessful because the efforts themselves are insufficient to elevate teacher performance, or because teachers have insufficient incentives to derive much benefit from the existing efforts (or because of a combination of these factors).

Research on professional development has focused both on the form of delivery and on the content of the training. Most of the many studies on these issues are small-sample case studies, or they focus on training's impact on teacher attitudes or instructional practices, or they are methodologically weak. Still, reviews of the literature suggest little evidence that professional development improves teaching.[94]

Unfortunately, new well-designed large-scale experimental studies confirm these literature reviews. The American Institutes for Research analyzed the influence of a one-year content-focused teacher-institute series on teacher knowledge, instructional practices, and student achievement.[95] Although there were positive impacts on teachers' knowledge of scientifically based reading instruction and on one of the three instructional practices promoted by the study, the institute series—even with additional coaching—did not result in higher student test scores at the end of the year. No differences in measured teacher or student outcomes were apparent. More recently, AIR reported findings on the impact of providing a professional-development program on rational-number topics to seventh-grade mathematics teachers.[96] In the second and final evaluation, AIR found that the program had no statistically significant impact on relevant student or teacher outcomes.

One might also consider mentoring and induction (a more structured form of mentoring that is instructionally focused and delivered by full-time, trained mentors) to be variants of professional development. Research on this strategy shows somewhat more promising results. For instance, a recent review of 15 empirical studies of induction finds that most show positive (though not always statistically significant) effects on three types of outcomes: teacher retention, teacher classroom practices, and student achievement.[97] A well-designed randomized-experiment study of induction,[98] in which treatment teachers received significantly more support during their comprehensive induction (some for one year, others for two) than teachers in the control group, found that additional support did not translate into impacts on classroom practices in the first year. Nor did the study find that teachers who received two years of comprehensive induction had higher levels of student achievement. But the study did find a positive and statistically significant impact on student achievement in teachers' third years, after they had completed the mentoring and induction program.[99]

One type of professional development that might benefit teachers is the feedback they receive about their *individual* teaching.[100] Recent evidence suggests that more targeted feedback about teaching might be more efficacious. Specifically, a recent study of Cincinnati Public Schools found that a comprehensive feedback system that relied on classroom observations based on the Danielson Framework for Teaching, combined with conferencing about what the observation implied for teacher classroom preparation and lesson planning led to improvement in the effectiveness of mid-career teachers in math instruction.[101] Note, however, that this feedback is likely only meaningful if the teacher observation system meaningfully differentiates teachers. As is evident from the discussion of *The Widget Effect* above, most evaluation systems do not differentiate teachers, meaning it is not possible to provide them with significant feedback about their individual needs.

Clearly one might draw the conclusion that the findings on professional development, mentoring, and feedback are generally not effective teacher improvement strategies. It is true that

the empirical evidence on the efficacy of efforts in these areas does not present a promising picture. But it is also important to consider the fact that they are carried out within a larger context of incentives. Right now, for instance, teacher participation in professional development is often compliance-driven, in that teachers are required to participate or are rewarded for "seat time" regardless of the development's impact. This creates little incentive for teachers to seek out the most-effective forms of professional development. This is unfortunate, because upgrading the skill set of incumbent teachers is probably a more politically palatable workforce-improvement strategy than changing the mix of people in the workforce.

In-Service Policies to Increase Teacher Effectiveness: Incentives

One of the primary arguments for creating a rigorous evaluation system is that teacher evaluations should be used in conjunction with monetary or career-path incentives to influence teacher behavior. There is growing interest in pay reform, as various localities depart from the single salary schedule.[102] There are good arguments for reforming the way teachers are paid. Existing "steps and lanes" salary schedules do not reflect labor-market realities of supply and demand, and they reward teacher characteristics (such as years of experience and postgraduate credentials) only weakly related to effectiveness.[103]

While there has been stronger experimental evidence that pay for performance has beneficial effects in a number of developing countries,[104] the empirical evidence of performance pay's effectiveness in the United States is mixed. Part of this may be related to the types of incentives that have been used. Most of the U.S. teacher performance pay experiments have been short-term, bonus-based incentives, teachers may be more responsive if they felt like the incentives were going to last and were built into base salaries. Nevertheless, while there may be some connection to student achievement,[105] recent evidence from several well-designed field experiments calls into question whether pay-for-performance is an effective school-reform strategy.[106]

An interesting exception to the above findings is a recent experiment that tests two kinds of incentives: the traditional type whereby teachers who have students meeting achievement objectives receive performance incentives and a second type whereby teachers are provided with incentives *in advance* and will lose the incentive if there students do not meet specified achievement objectives.[107] Like the research specified above, the study finds no statistically significant results from the traditional type of incentive, but large significant effects associated with the potential that teachers lose an incentive already received. This finding that "loss aversion" seems to affect teachers more than the potential for gain is consistent with theory about behavioral responses to incentives, but it is not clear how such an incentive could be made to work in public schools more generally.

A possible explanation for these mixed findings is that the studies focused on different types of effects of pay-for-performance. The evidence clearly shows that traditionally structured large financial incentives (up to $15,000 per teacher) to raise student test scores alone do not lead to changes in teacher behaviors that have detectible effects.[108] But changing the behavior of incumbent teachers is only one of the potential avenues through which pay-for-performance might influence the quality of the teacher workforce. Research on the private sector, for instance, finds that performance-pay systems influence both the behavior of incumbents and the composition of the workforce—that is, who opts in and out of the workforce.[109] Experiments are carried out over short time periods (e.g. the POINT experiment lasted for three school years) and consequently they do not focus on workforce-composition effects; other research may be capturing the potential for changes to both behavior and the workforce.[110]

A second explanation is that some performance-pay systems are better-aligned with teacher-evaluation and performance-feedback systems than others. Recent evidence on the pay reform implemented in Denver, known as ProComp, provides a more promising picture of the efficacy of pay reform.[111] But the changes in Denver were enacted alongside broader changes designed to build capacity around data, human-resource functions, professional development, and performance evaluation.[112] For instance, a key component of ProComp is that it requires teachers to sit down with their principals to set student-growth goals, for which they are held accountable.

As with professional development, it is entirely possible that pay-for-performance is not an effective strategy alone, but only works when implemented in conjunction with other systemic changes. Unfortunately, given the aforementioned fact that performance-evaluation systems generally do not differentiate teachers and that very few school systems deviate from the single salary schedule, little can be said about best practices when it comes to linking teacher evaluation (or other human-resource systems) with pay-for-performance.

Shaping the Teacher Workforce Through Deselection Policies

The difficulty of identifying teacher effectiveness at the point of hire, combined with discouraging findings about programs for improving in-service teachers, has led some to advocate opening up the teacher labor market to a broader assortment of prospective teachers.[113] Then, once districts have better information about actual classroom practice, they can be much more selective about which teachers are allowed to stay in the workforce.

The numbers behind the notion that careful teacher deselection could be a key means of improving teacher quality are intriguing. Eric Hanushek, for instance, makes an empirically based argument that more-extensive teacher deselection would have a powerful impact on both student achievement and U.S. international competitiveness.[114] But Hanushek's calculations are based on what we know about the impact of teacher effectiveness on student achievement, not on actual policy variation. Strategic deselection of teachers is exceedingly rare.[115] Such a policy could have far-reaching and unintended consequences on the makeup of those who opt to enter the teacher labor force and on how in-service teachers behave.[116]

Recent research provides more information on early-career de-selection efforts. Chetty et al., for instance, estimate that the impact of replacing a teacher who is in the bottom 5 percent of the value-added distribution with an average teacher would increase the (present) value of the collective lifetime earnings of students in the affected classroom by more than $250,000.[117] While this is evidence of the potential *value* of a de-selection policy, other studies have analyzed how likely it is to be successful in practice. Three recent reports have found that, while teachers tend to improve in their first few years on the job, poor performing teachers do not, on average "catch up" with higher performers and districts may be better off taking their chances with new teachers.[118] Only two studies examine actual teacher de-selection policies. One investigates a change in the collective-bargaining agreement for Chicago Public Schools that allowed principals to fire non-tenured teachers for any reason, without having to provide documentation go through a typical dismissal hearing process.[119] The study finds some evidence of an increased likelihood of teacher with frequent absences or lower value added being fired.[120]

Among teachers in Washington state who received layoff notices following the recent economic downturn (in 2008–09 and 2009–10), seniority was the overwhelming driver in layoff determination.[121] Not surprisingly, the set of teachers who received a layoff notice were quite different from the teachers who would have been targeted, had the school districts been using value added as a measure of teacher effectiveness[122]—and that difference translates to about two and a half to three and a half months of student learning.

Explicit deselection policies are rarely used in public schools, so any conclusions about their efficacy would be premature. The empirical evidence that preservice credentials are not strong predictors of teacher effectiveness, combined with the very good evidence that we can learn quite a lot about the effectiveness of in-service teachers, validates the notion that school systems might greatly increase the effectiveness of their workforce through the use of more judicious deselection policies.

Conclusions

This paper begins by stressing the oft-cited refrain that teacher quality is the key schooling variable when it comes to influencing student achievement, and goes on to review what is known about the ways to affect the quality of the teacher workforce. What does the existing evidence suggest for individuals or groups wishing to know the "right" set of teacher policies to promote? Unfortunately, academic research rarely yields findings that are so definitive as to provide specific policy guidance, but it does suggest some broad implications. First, it is clear that improvements to the quality of the teacher workforce have the potential to radically improve the performance of America's schools. While we cannot easily predict which teachers will be successful with students based on their credentials, we do know that there is significant variation among teachers and that the differences in teacher effectiveness have meaningful effects on student outcomes.

Second, and closely related to the point above, state-regulated licensure systems do not, in general, appear to be an effective means of screening for teacher quality. Definitive evidence indicates that there is far greater variation among teachers who hold similar preservice credentials (the pathway into the classroom, the college from which teachers graduated, certifications held, etc.) than among teachers with different credentials.

Third, existing research suggests that investments in the incumbent workforce through better professional development, mentoring, or incentives have not generally resulted in significant improvements in teacher effectiveness. Much of the research suggests that the quality of the workforce may be more dependent on getting the right people "on (and off) the bus" than on helping the people already on the bus to improve.

But it is important to understand that this is not necessarily how things would play out in a radically different educational context. It is conceivable, for instance, that the effectiveness of a professional-development program might be quite different if teachers had a more direct incentive (that is, if they were rewarded for performance in some way) when making professional-development decisions and receiving training. Likewise, it is possible that incentives could make a difference if teachers were provided with higher-quality professional development options. The bottom line here is that reforms are unlikely to produce big workforce-productivity gains unless they are based in a coherent theory of action. The most pressing teacher-policy issue is the fact that, in most school districts, evaluation systems are broken.

Arguably, the starting point for all human-capital policies in all school systems should be performance evaluations. But meaningful action is nearly impossible when current evaluation systems give roughly 99 percent of teachers the same rating. Inserting more rigor into the process is probably the first step, and likely the most important one toward insuring a high-quality teacher workforce.

[1] See, for example, Eric A. Hanushek, "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro-Data," *American Economic Review* 61, no. 2 (May 1971): 280–88.

[2] See, for example, D. Goldhaber, D. Brewer, and D. Anderson, "A Three-Way Error Components Analysis of Educational Productivity," *Education Economics* 7, no. 3 (1999)**:** 199–208; S. G. Rivkin, E. A. Hanushek, and J. F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73, no. 2 (2005): 417–58; D. Aaronson, L. Barrow, and W. Sander, "Teachers and Student Achievement in the Chicago Public High Schools," *Journal of Labor Economics* 25, no. 1 (2007): 95–135; B. Nye, S. Konstantopoulos, and L. V. Hedges, "How Large Are Teacher Effects?," *Educational Evaluation and Policy Analysis* 26, no. 3 (2004): 237–57.

[3] Moreover, differences in teacher effectiveness ("effectiveness" here is used to refer to value-added estimates of teacher effects) are generally found to be far larger than other readily quantifiable educational investments, such as reductions in class size. Rivkin, Hanushek, and Kain, for instance, find that a standard-deviation increase in teacher quality raises student achievement in reading and math by about 10 percent, an effect comparable with reducing class size by 10 to 13 students.

[4] Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "The Long-term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," Working Paper No. 17699 (Cambridge, MA: National Bureau of Economic Research, 2011).

[5] See, for instance, http://www.edweek.org/ew/issues/teacher-quality/.

[6] "Deselection" refers to policies associated with the removal of teachers from the workforce: firing, layoffs, and non-tenuring.

[7] As described later in this paper, studies examining the relationship between teachers' performance on standardized tests and student achievement tend to find a positive, albeit weak, correlation between the two.

[8] M. Bacolod, "Do Alternative Opportunities Matter? The Role of Female Labor Markets in the Decline of Teacher Quality," *Review of Economics and Statistics* 89 (2007): 737–51; M. Bacolod, "Who Teaches and Where They Choose to Teach: College Graduates of the 1990s," *Educational Evaluation and Policy Analysis* 29 (2007): 155–68; S. Corcoran, W. Evans, and R. Schwab, "Women, the Labor Market, and the Declining Relative Quality of Teachers," *Journal of Policy Analysis and Management* 23 (2004): 449–70; E. A. Hanushek and R. R. Pace, "Who Chooses to Teach (and Why)?" *Economics of Education Review* 14, no. 2 (1995): 101–17; C. Hoxby and A. Leigh, "Pulled Away or Pushed Out? Explaining the Decline in Teacher Aptitude in the United States," *American Economic Review* 94 (2004): 236–40.

[9] Corcoran and Schwab.

[10] Hoxby and Leigh.

[11] Corcoran and Schwab

[12] E. A. Hanushek and S.G. Rivkin, "Teacher Quality," in: E. Hanushek and F. Welch, eds., *Handbook of the Economics of Education* (Amsterdam: North-Holland, 2006), 1,054–78.

[13] Hoxby and Leigh; Goldhaber and Liu find that differences in the individual characteristics rewarded between teacher and non-teacher labor markets means that those with stronger academic backgrounds or technical training face greater opportunity costs for being a teacher. Dan Goldhaber and Albert Liu, "Occupational Choices and the Academic Proficiency of the Teacher Workforce," in William Fowler (Ed.), *Developments in School Finance 2001-02* (Washington, DC: NCES, 2003).

[14] D. Ballou, "Do Public Schools Hire the Best Applicants?," *Quarterly Journal of Economics* 111, no. 1 (1996): 97–133.

[15] Drew Gitomer, Andrew S. Latham, and Robert Ziomek, "The Academic Quality of Prospective Teachers: The Impact of Admissions and Licensure Testing," (Princeton, NJ: Educational Testing Service, 1999).

[16] Robin R. Henke, Xianlei Chen, and Sonya Geis, "*Progress through the Teacher Pipeline: 1992–93 College Graduates and Elementary/Secondary School Teaching as of 1997,"* (Washington, DC: U.S. Department of Education, National Center for Education Statistics, Office of Educational Research and Improvement, 2000).

[17] D. Goldhaber, "Everyone's Doing It, But What Does Teacher Testing Tell Us about Teacher Effectiveness?," *Journal of Human Resources* 42, no. 4 (2007): 765–94; Hanushek and Pace; M. Podgursky, R. Monroe, and D. Watson, "The Academic Quality of Public School Teachers: An Analysis of Entry and Exit Behavior," *Economics of Education Review* 23, no. 5 (2004): 507–18; T. R. Stinebrickner, "A Dynamic Model of Teacher Labor Supply," *Journal of Labor Economics* 19, no. 1 (2001): 196–230; T. R. Stinebrickner, "An Analysis of Occupational Change and Departure from the Labor Force: Evidence of the Reasons That Teachers Leave" Journal of Human Resources 37, no. 1 (2002): 192–216.

[18] By Mona Mourshed and Michael Barber, How the Best-Performing School Systems Come Out on Top (McKinsey & Company, September 2007). http://mckinseyonsociety.com/downloads/reports/Education/Worlds_School_Systems_Final.pdf
[19] Goldhaber, "Everyone's Doing It."
[20] Donald Boyd et al., "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools," *Journal of Policy Analysis and Management* 27, no. 4 (2008): 793–818.
[21] Bryon Auguste, Paul Kihn, and Matt Miller, "Closing the Talent Gap: Attracting and Retaining Top-Third Graduates to Careers in Teaching" (Washington, DC: McKinsey & Company Social Sector Office, September 2010).
[22] For more-exhaustive reviews of the theory behind teacher licensure and what we know about its impacts, see Donald Boyd, Dan Goldhaber, Hamilton Lankford, and James Wyckoff, "The Effect of Certification and Preparation on Teacher Quality," *The Future of Children* 17 (1), Spring 2007, 45-68 and Dan Goldhaber "Licensure: Exploring the Value of this Gateway to the Teacher Workforce," in *Handbook of the Economics of Education*, Vol. 3, edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann (315–39). (Amsterdam: North Holland, 2011).
[23] D. D. Goldhaber and D. J. Brewer, "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement," *Educational Evaluation and Policy Analysis* 22, no. 2 (2000).
[24] Some recent research has attempted to address this knowledge gap by focusing on whether any teacher-training-program characteristics (for example, timing of student teaching) predict program effectiveness. See, for example, D. Boyd et al., "Teacher Preparation and Student Achievement," *Educational Evaluation and Policy Analysis* 31, no. 4 (2009): 416–40; Rockoff, Staiger, and Kane.
[25] The number of individuals entering the teaching profession through alternative routes has risen considerably, from roughly 20,000 in 2001 to almost 60,000 in 2006. (See Paul E. Peterson and Daniel Nadler, "What Happens When States Have Genuine Alternative Certification," *Education Next* 9:1, November 20, 2008.) The extent, however, to which individual states utilize alternative licensure as an important source of new teachers varies greatly.
[26] Paul T. Decker, Daniel P. Mayer, and Steven Glazerman, "The Effects of Teach for America on Students: Findings from a National Evaluation" (Princeton, NJ: Mathematica Policy Research, 2004); Thomas J. Kane, Jonah Rockoff, and Douglas Staiger, "What Does Certification Tell Us about Teacher Effectiveness?: Evidence from New York City," *Economics of Education Review* 27, no. 6 (2008): 615–31; Z. Xu and C. Hannaway, "Making a difference: The Effects of Teach for America in High School," (Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, 2007).
[27] Note that the study by Decker, Mayer, and Glazerman is quite convincing on this, as it is based on a randomized experiment. Also, note that, while less definitive, my read of the research evidence also suggests that TFA teachers are more effective in math than in reading.
[28] Decker, Mayer, and Glazerman; J. E. Rockoff, D. O. Staiger, and T. J. Kane, "Photo Finish: Certification Doesn't Guarantee a Winner," *Spring* 7, no.1 (2007).
[29] J. Constantine et al., "An Evaluation of Teachers Trained through Different Routes to Certification, Final Report" (NCEE 2009-4043, Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2009).
[30] D. Boyd et al., "Teacher Preparation and Student Achievement," *Educational Evaluation and Policy Analysis* 31, no. 4 (2009): 416–40; Rockoff, Staiger, and Kane.
[31] Boyd et al., "Teacher Preparation and Student Achievement"; J. E. Rockoff, Brian Jacob, Thomas Kane, and Douglas Staiger, "Can You Recognize an Effective Teacher When You Recruit One?," *Education Finance and Policy* 6, no. 1 (Winter 2011): 43–74.
[32] Several states – Colorado, Louisiana, Texas, and Tennessee – for example have passed legislation that judges teacher preparation programs based, in part, on the student test results of the teachers they produce.
[33] Donald Boyd, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff, "Teacher Preparation and Student Achievement," *Educational Evaluation and Policy Analysis*, 31 (2009): 416-440; Dan D. Goldhaber, Stephanie Liddle, and Roddy Theobald, "The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement," *Economics of Education Review*, 34 (2013): 29-44; Cory Koedel, Eric Parsons, Michael Podgursky, and Mark Ehlert "Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs?," University of Misourri Working Paper No 12-04 (2012); Kata Mihaly, Daniel McCaffrey, Tim R. Sass, and JR Lockwood, "Where you Come From or Where You Go? Distinguishing Between

School Quality and the Effectiveness of Teacher Preparation Program Graduates," *Journal of Education Finance and Policy* (Forthcoming)

[34] These findings were found to vary based on the state in the study. The magnitude of the effects, for instance, were relatively large in a study based on teachers from New York City (see Boyd et al., 2009) and relatively large based on a sample of teachers from Missouri (see Koedel et al., 2012).

[35] The Boyd et al. (2009) study suggests that there are important features of teacher training that may contribute to the effectiveness of in-service teachers.

[36] See Goldhaber et al. (2013) and Mihaly et al. (Forthcoming).

[37] See, for instance, Edward Crowe "Measuring What Matters: A Stronger Accountability Model for Teacher Education." Washington, DC: Center for American Progress, 2010.

[38] Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor, "Who Teaches Whom? Race and the Distribution of Novice Teachers," *Economics of Education Review* 24, no. 4 (2005): 377–92; Dan Goldhaber, Hyung-Jai Choi, and Lauren Cramer, "A Descriptive Analysis of the Distribution of NBPTS-Certified Teachers in North Carolina," *Economics of Education Review* 26, no. 2 (2007):160–72; Hamilton Lankford, Susanna Loeb, and James H. Wyckoff, "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis," *Educational Evaluation and Policy Analysis* 24, no. 1 (2002):37–62.

[39] For a review of the literature linking observable teacher qualifications to student outcomes, see D. Goldhaber, "Teacher Pay Reforms: The Political Implications of Recent Research" (CEDR Working Paper 2010-4, University of Washington, 2010).

[40] Tim R. Sass et al., "Value Added of Teachers in High-Poverty Schools and Lower-Poverty Schools" (Calder Working Paper No. 52, 2010); Dan Goldhaber, Brian Gabele, and Joe Walch, "Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments," *Statistics, Politics, and Policy* (Forthcoming).

[41] D. Goldhaber, K. Destler, and D. Player, " Teacher Labor Markets and the Perils of Using Hedonics to Estimate Compensating Differentials in the Public Sector," *Economics of Education Review* 29, no. 1 (2010), 1–17.

[42] Charles T. Clotfelter, Elizabeth Glennie, Helen F. Ladd, and Jacob L. Vigdor, "Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina," *Journal of Public Economics* 92, no. 6 (2007): 1,352–70.

[43] Y. Hou, S. Selden, P. Ingraham, and S. Bretschneider, "Decentralization of Human Resources Management in the States—Driving Forces and Implications, *Review of Public Personnel Administration* 20, no. 4 (2007): 9–23; S. C. Selden, S. Ammar, R. Wright, and W. Jacobson, "A New Approach to Assessing Performance of State Human Resource Management Systems: A Multi-Level Fuzzy Rule Based System," *Review of Public Personnel Administration* 20, no. 3 (2000); S. C. Selden, P. W. Ingraham, and W. Jacobson, "Human Resource Practices in State Governments: Findings from a National Survey," *Public Administration Review* 61, no. 5 (2001).

[44] Michael DeArmond, Betheny Gross, and Dan Goldhaber, "Look Familiar? Charters and Teachers" in *Hopes, Fears, and Reality: A balanced look at American charter schools in 2007*, Robin Lake (ed). (Seattle: Center on Reinventing Public Education, 2007).

[45] Districts tend to hire teachers who either attended or graduated from schools in that same district. It appears that districts hire "simply those [people] the district knows best, their own graduates" (p. 405). (See: R. P. Strauss, L. R. Bowes, M. S. Marks, and M. R. Plesko, "Improving Teacher Preparation and Selection: Lessons from the Pennsylvania Experience," *Economics of Education Review* 19, no. 4 (2000): 387–415.) This is consistent with findings that teachers in New York were twice as likely to work at a school within five miles of their hometown (more precisely, where they graduated from high school) than at a school within 20 miles of their hometown. See: D. Boyd, H. Lankford, S. Loeb, and J. Wyckoff, "Analyzing the Determinants of the Matching of Public School Teachers to Jobs" (Albany, NY: Rockefeller Institute of Government, State University of New York, 2002).

[46] D. Goldhaber, "Teacher Pay Reforms."

[47] S. Metzger, and M. J. Wu, "Commercial Teacher Selection Instruments: The Validity of Selecting Teachers through Beliefs, Attitudes, and Values. *Review of Educational Research* (2008). While their study was limited in that it focused on in-service teachers, Rockoff, Jacob, Kane, and Staiger surveyed new math teachers in New York City, collecting information on both traditional and nontraditional teacher traits and characteristics (for example, teaching-specific content knowledge, cognitive ability, personality traits, feelings of self-efficacy, and scores on a teacher-selection instrument). They found that few of these characteristics are predictive of teacher effectiveness. Collectively, the characteristics explain only about 10 percent of the variation in teacher effectiveness in the workforce.

[48] The study was somewhat limited in that it focused on new math teachers already in service (that is, they had already been selected to teach), and the assessments were based on surveys of those teachers. It is conceivable that the relationship would be stronger when examining all potential teachers (Rockoff, Jacob, Kane, and Staiger).
[49] See Will Dobbie, "Teacher Characteristics and Student Achievement: Evidence from Teach For America," Available from: http://www.people.fas.harvard.edu.
[50] J. Levin and M. Quinn, "Missed Opportunities: How We Keep High-Quality Teachers out of Urban Classrooms" (New York: The New Teacher Project, 2003).
[51] Edward Liu and Susan M. Johnson, "Staffing to the Test: Are Today's School Personnel Practices Evidence Based?," *Educational Evaluation and Policy Analysis* 33 (December 1, 2011): 483–505.
[52] The extent of late hiring appears to vary considerably by state. Nationally the figure is estimated to be just over 10 percent, but it can be over 30 percent in some years in states with growing school populations like California and Florida. See Mimi Engel, "The Timing of Teacher Hires and Teacher Qualifications: Is There an Association?," Teachers College Record (Forthcoming); and Edward Liu and Susan Moore Johnson, "New Teachers' Experiences of Hiring: Late, Rushed, and Information-Poor," *Educational Administration Quarterly, 42*(3) (2006): 324-360.
[53] Specifically, about 21 percent of teachers hired before the school year left before the start of the next year as compared to 36 percent of those hired after the school year; the corresponding figures for leaving teaching (in Michigan) were found to be about 8 and 16 percent. See Nathan D. Jones, Adam Maier, Erin Grogan, and Barbara Schneider, "The Extent of Late Hiring and its Relationship with Teacher Turnover: Evidence from Michigan," Working paper (2012).
[54] Specifically, the achievement of students in classrooms with late hired teachers is estimated to be about 4 percent of a standard deviation lower on both math and reading tests than the achievement in classrooms staffed by other newly hired (not late) teachers. See John P. Papay, Matthew A. Kraft, Julia Bloom, Kate Buckly, and David Liebowitz, "Missed Opportunities in the Labor Market or Temporary Disruptions? How Late Teacher Hiring Affects Student Achievement," Paper presented at the 2013 annual meeting of the Association for Education Finance and Policy (March 2013).
[55] In other words, late hires may be less effective because of their own skill set, but also because they do not have time to acclimate to a new school or are a poor match with their colleagues, and this issue of match may lead to spillover effects that impact other teachers.
[56] Tom Toch and Robert Rothman, "Rush to Judgment: Teacher Evaluation in Public Education," (Washington, DC: Education Sector, 2008).
[57] D. Weisburg, S. Sexton, J. Mulhern, and D. Keeling, "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness" (The New Teacher Project, 2009).
[58] Note that the value-added literature discussed above, however, suggests that teachers are in fact quite different from one another in terms of their impacts on student learning.
[59] And moreover, these evaluations do not reflect more critical private assessments by principals or teachers themselves (See: Pamela D. Tucker, "Lake Wobegon: Where all teachers are competent (or have we come to terms with the problem of incompetent teachers?", *Journal of Personnel Evaluation in Education,* 11(1)**,** 1997, 103-126 and Daniel Weisburg, Susan Sexton, Jennifer Mulhern, and David Keeling, "The widget effect: Our national failure to acknowledge and act of differences in teacher effectiveness*,"* (Chicago: The New Teacher Project, 2009). In fact, some research (See: Brian A. Jacob and Lars Lefgren, "What Do Parents Value in Education? An Empirical Investigation of Parents Revealed Preferences for Teachers," *Quarterly Journal of Economics,* vol. 122, no. 4, 2007, pp. 1603–1637) shows that principals can accurately (as reported by their *private* assessments of teachers) identify teachers at the top and bottom of the value-added performance distribution.
[60] Survey data from school districts in Washington state found that three-fourths were using binary evaluation systems in 2010-11. See Chad Aldeman, "The Evergreen Effect: Washington's Poor Evaluation System Revealed," (Washington, D.C.: Education Sector, 2013).
[61] Much of this has been spurred on by the federal government's Race to the Top grant competition.
[62] K. S. Loup, J. S. Garland, C. D. Ellett, and J. K. Rugutt, "Ten Years Later: Findings from a Replication of a Study of Teacher Evaluation Practices in Our 100 Largest School Districts," *Journal of Personnel Evaluation in Education* 10, no. 1 (1996): 203–26.
[63] What constitutes "high quality" is discussed in detail elsewhere (for example, Toch and Rothman), but it is worth noting that the quality of the observation will depend on the instrument used, the number of times teachers are observed, and the nature (announced or unannounced) of the observation, etc.
[64] J. H. Tyler, E. S. Taylor, T. J. Kane, and A. L. Wooten, "Using Student Performance Data to Identify Effective Classroom Practices," *American Economic Review* 100, no. 2 (2010): 256–60.
[65] D. Goldhaber and M. Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions," *American*

Economic Review: Papers and Proceedings 100, no. 2 (2010): 250–55; Douglas N. Harris, "Clear Away the Smoke and Mirrors of Value-Added," *Phi Delta Kappan* 91, no. 8 (2010): 66–69.
[66] D. Goldhaber, "Teacher Pay Reforms."
[67] Goldhaber et al. (Forthcoming).
[68] For a discussion of how student growth percentile measures in one school district compare with value-added measures of teachers, see Goldhaber et al. (Forthcoming).
[69] The heart of the issue is whether the type of variables that are typically available in administrative datasets can be used to separately identify what students "bring to the table" when they show up in a teacher's class from the contribution that teachers make toward their learning (as is generally measured by standardized tests). For more on this potential problem, see Jesse Rothstein "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics,* vol. 125, no. 1, February 2010: 175–214.
[70] Two studies focus on this issue by assessing whether VAM estimates of differences between teachers appear to be consistent in an experimental period, when teachers are assigned to their classrooms randomly, and in a non-experimental period, when they are not. See Thomas J. Kane and Douglas O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation: Working Paper No. 14607," (Cambridge, MA: National Bureau of Economic Research, 2008); and Thomas J. Kane, Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger, "Have We Identified Effective Teachers?: Validating Measures of Effective Teaching Using Random Assignment," MET Project Report (2013). An alternative methodology, used by Chetty et al. (2011) tests for bias relying on the fact that parents are unlikely to follow teachers when they leave one school for another so the change in school value added associated with an incoming teacher can provide a gage of the validity of the value added measure.
[71] See, for instance: Dan Goldhaber, Pete Goldschmidt, and Fannie Tseng, "Teacher Value Added at the High School Level: Different Models, Different Answers?," *Educational Evaluation and Policy Analysis,* (in Press); and Kirabo C. Jackson, "Teacher Quality at the High-School Level: The Importance of Accounting for Tracks" NBER Working Paper No. 17722 (2012).

[72] Several studies also compare value-added measures across statistical models. In general these studies find a strong correlation between various models, implying that different models, such as the "Colorado Growth Model" and value added models that account for prior student scores and student covariates, such as receipt of free and reduced price lunch, tend to produce similar teacher rankings. However research also shows that two models that generally result in similar rankings can still produce very different results for teachers who serve classrooms of students that are very different from the average classroom (e.g. very high poverty), depending on how they handle student background differences. For more detail on the implications for teacher rankings of using different tests or different value added models, see Goldhaber et al. (Forthcoming);
[73] See C. Kirabo Jackson, "Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina," NBER Working Paper No. 18624 (2012).
[74] For more on teaching to the test, see James W. Popham, "Teaching to the Test," *Educational Leadership* 2001 58(6): 16-20. For more on cheating, see Brian A. Jacob and Steven D. Levitt, "Rotten Apples: an Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, 118(3), 2008: 843-877.
[75] See, for instance, Elizabeth U. Cascio and Douglas O. Staiger, "Knowledge, Tests, and Fadeout in Educational Interventions," NBER Working Paper No. 18038 (2012).
[76] See Chetty et al. (2011).
[77] Or indeed to outright cheating, a problem that clearly needs to be addressed as the recent scandal in Atlanta illustrates.
[78] On the reliability/stability of estimates, see: Dan Goldhaber and Michael Hansen, "Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance," *Economica.*(Forthcoming)..
[79] For more detail on classification errors, see Dan Goldhaber and Susanna Loeb, "What Do We Know About the Tradeoffs Associated with Teacher Misclassification in High Stakes Personnel Decisions?" (Carnegie Knowledge Network, Carnegie Foundation for the Advancement of Teaching, April 2013).
[80] See Steven Glazerman et al., "Evaluating Teachers: The Important Role of Value-Added" (Washington, DC: The Brookings Institution, November 2010).
[81] For a more thorough discussion of this, see D. Goldhaber, "Teacher Pay Reforms."
[82] Note, however, that these do not exist in the case of evaluation methods commonly used today either.
[83] D. Goldhaber, "Teacher Pay Reforms."

[84] Toch and Rothman.
[85] There are a number of different aspects to the study and MET project reports, all of which can be accessed from www.metproject.org.
[86] Thomas J. Kane and Douglas O. Staiger, "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains" (Measures of Effective Teaching, Gates Foundation Report, 2012).
[87] Importantly, findings from the MET study also suggest that the reliability of the classroom observations depends a great deal on teachers' being observed multiple times by trained observers.
[88] See Pam Grossman, Susanna Loeb, Julia Cohen, Karen Hmmerness, James Wyckof, Danoald Boyd, and Hamilton Lankford, "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value Added Scores," *American Journal of Education*, Forthcoming.
[89] This is a measure of whether teachers focus on providing students with step-by-step explicit strategies they can employ to, for instance, analyze grammatical errors, interpret literary text, or make a compelling argument.
[90] Kane and Staiger (2012).
[91] These are results from Denver, Colorado. See Dan Goldhaber and Joe Walch, "Strategic Pay Reform: A Student Outcomes-Based Evaluation of Denver's ProComp Teacher Pay Initiative,". *Economics of Education Review*. 31, no 6 (2012): 1067–1083.
[92] The MET study administered supplemental assessments to participating students which were designed to assess more complex skills than those measured by state assessments. See Kata Mihaly, Daniel F. McCaffrey, Douglas O. Staiger, and J.R. Lockwood, "A Composite Estimate of Effective Teaching," (Measures of Effective Teaching, Gates Foundation Report, 2013).
[93] This is assuming that one considers non-student test outcomes as legitimate measures.
[94] D. K. Cohen and H. C. Hill, "Instructional Policy and Classroom Performance: The Mathematics Reform in California," *Teachers College Record* 102, no. 2 (2000): 294–343; M. Kennedy, "Form and Substance in In-Service Teacher Education" (research monograph 13 (Madison, WI: University of Wisconsin, 1998).
[95] American Institutes for Research, "Development Interventions on Early Reading Instruction and Achievement" (NCEE 2008-4030, U.S. Department of Education, 2008).
[96] American Institutes for Research, "Middle School Mathematics Professional Development Impact Study: Findings after the Second Year of Implementation" (NCEE 2011-4024, U.S. Department of Education, 2011).
[97] Richard Ingersoll and Michael Strong, "The Impacts of Induction and Mentoring Programs for Beginning Teachers: A Critical Review of the Research," *Review of Education Research* 81:2, 201-233.
[98] Steven Glazerman et al., "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study" (NCEE 2010-4027, U.S. Department of Education, 2010).
[99] Exposure to neither one nor two years of comprehensive induction had a positive impact on retention or other teacher-workforce outcomes such as job satisfaction or feelings of preparedness.
[100] Generally, however, this is not how professional development is delivered. Rather it tends to take the form of workshops or in-service programs, and is not tailored to the needs of individual teachers.
[101] Interestingly, there was no statistically significant effect of the feedback system on value added measures of teachers in reading. See Eric S. Taylor and John H. Tyler, "The Effect of Evaluation on Teacher Performance," *American Economic Review*, 102(7) (2012): 3628-3651.
[102] Michael Podgursky and Matthew G. Springer, "Teacher Performance Pay: A Review," *Journal of Policy Analysis and Management*, 26:4.
[103] The single-salary schedule rewards teacher experience and degree level. Good evidence indicates that teachers become more effective in their first few years in the classroom, but it also shows that this relationship levels out beyond the first three to six years of experience (Charles T. Clotfelter, Helen Ladd, and Jacob L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources* 41 (2010): 778-820). But, outside of a few narrow exceptions, teacher degree level appears to be completely unrelated to effectiveness. (For more detail, see Dan Goldhaber, and Dominic Brewer, "Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity," *Journal of Human Resources*, 32(3) (2007):505-523.)
[104] See, for example,Duflo and Hanna, 2005; Muralidharan and Sundararaman, 2011.
[105] David N. Figlio and Lawrence W. Kenny, "Individual Teacher Incentives and Student Performance," *Journal of Public Economics* 91, no. 5–6 (2007): 901–14.
[106] Fryer, Roland, "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." Journal of Labor Economics, forthcoming; Matthew G. Springer, Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R.

Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher, "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching," (Nashville, TN: National Center on Performance Incentives at Vanderbilt University, 2010). For other recent evaluations of pay-for-performance systems, see also Dale Ballou "Pay for Performance in Public and Private Schools," *Economics of Education Review*, 20 (2010): 51-61; S. Glazerman and A. Seifullah, "An Evaluation of the Teacher Advancement Program (TAP) in Chicago: Year Two Impact Report," (Princeton,NJ: Joyce Foundation by Mathematica Policy Research, 2010); Dan Goldhaber and Joe Walch (2012); M. G. Springer, M.A. Winters, "New York City's School-Wide Bonus Pay Program: Early Evidence from a Randomized Trial" NCPI Working Paper 2009-02, (Nashville, TN: Vanderbilt University, 2010); and L. Taylor and M.G. Springer, "Optimal Incentives for Public Sector Workers: The Case of Teacher-Designed Incentive Pay in Texas," NCPI Working Paper 2009-05 (Nashville, TN: Vanderbilt University, 2009).

[107] See Fryer, Roland, Levitt, Steven D., List, John, and Sadoff, Sally (2012). "Enhancing the Efficacy of Teacher Incentives through Loss Aversion: A Field Experiment." NBER Working Paper No. 18237.
[108] The study on loss aversion, however, suggests there is a potential for incumbent teachers to become more effective.
[109] Edward P. Lazear, "Speeding, Terrorism, and Teaching to the Test," *Quarterly Journal of Economics* 121, no. 3 (2006): 1,029–61.
[110] Figlio and Kenny.
[111] Goldhaber and Walch, (2012).
[112] Proctor et al., "Making a Difference in Education Reform: ProComp External Evaluation Report 2006–2010" (Denver, CO: The Evaluation Center, University of Colorado Denver, 2011).
[113] R. J. Gordon, T. J., Kane, and D. O. Staiger, "Identifying Effective Teachers Using Performance on the Job" (Washington, DC: Brookings Institution, 2006).
[114] Specifically, were 5 to 10 percent of the least effective teachers (2 or 3 teachers in a school of 30 teachers) to be removed from the workforce and replaced with average teachers, achievement of U.S. students (which is below average on international tests) would rise toward the top in international comparisons. Eric A. Hanushek, "Teacher Deselection," in Dan Goldhaber and Jane Hannaway, eds., *Creating a New Teaching Profession* (Washington, DC: Urban Institute Press, 2009), 165–80.
[115] Weisburg, Sexton, Mulhern, and Keeling. "The Widget Effect."
[116] Today, teaching is a relatively secure occupation, and changes to the security of the occupation might shift the number or quality of prospective teachers.
[117] Chetty et al.
[118] See, for example, A. Atteberry, S. Loeb, and J. Wyckoff, "Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness" (Calder Working Paper No. 90, 2013); Goldhaber and Hansen (2010), and "The Irreplaceables: Understanding the Real Retention Crisis in America's Urban Schools" (The New Teacher Project, 2012).
[119] B. A. Jacob, "Do Principals Fire the Worst Teachers?" Working Paper 15715, (Washington, DC: National Bureau of Economic Research, 2010).
[120] He also finds that teachers from less competitive colleges, teachers who failed a licensure test, older teachers, and male teachers are more likely to be fired.
[121] There is, however, some evidence that teachers in high-demand subjects (math and science) are slightly less likely to receive a layoff notice. This subject effect is swamped by the effect of seniority. See: D. Goldhaber and R. Theobald, "Managing the Teacher Workforce: The Consequences of 'Last in, First out' Personnel Policies," *Education Next* 11, no. 4 (Fall 2011).
[122] The Goldhaber and Theobald findings are strikingly similar to simulations conducted using data from New York City on the student-achievement effects of "last in, first out" versus value-added effectiveness layoff systems (see Donald Boyd, Hamilton Lankford, Susannah Loeb, Matthew Ronfeldt, and James Wyckoff, "The Role of Teacher Quality in Retention and Hiring: Using Applications-to-Transfer to Uncover Preferences of Teachers and Schools," *Journal of Policy Analysis and Management* 30 (2010):88-110).

ABOUT DAN GOLDHABER

Dan Goldhaber is the director of the Center for Education Data & Research at the University of Washington Bothell.

ABOUT STAND FOR CHILDREN LEADERSHIP CENTER

Stand for Children Leadership Center is a 501(c)(3) nonprofit that provides leadership development and training to everyday citizens. Our mission is to ensure that all children, regardless of their background, graduate from high school prepared for, and with access to, college and career training. To make that happen, we:

- Educate and empower parents, teachers, and community members to demand excellent public schools.
- Advocate for effective local, state and national education policies and investments.
- Ensure the policies and funding we advocate for reach classrooms and help students.

Learn more at **www.stand.org**