The George W. Bush Institute's Education Reform Initiative Presents

The Productivity for Results Series

Teacher Effectiveness Research and the Evolution of U.S. Teacher Policy

NO. 5
JAN 2015

DAN GOLDHABER, DIRECTOR, *National Center for Analysis of Longitudinal Data in Education Research, and Vice President, American Institutes for Research*

THE BUSH INSTITUTE
AT THE
GEORGE W. BUSH
PRESIDENTIAL CENTER

Empirical research has shown that teacher quality is the largest in-school factor contributing to student achievement, but that observable teacher characteristics such as education level and certification status are poor proxies for teacher quality. This chapter describes the import of teachers for student achievement, and the extent to which the variation in teacher effectiveness is associated with the credentials that have traditionally been used to determine their employment eligibility and compensation. It goes on to discuss the prospects for policies designed to act on the differences that exist between teachers, and concludes with thoughts on possible pathways for improving teacher quality.

INTRODUCTION

"…Research on measures of teacher credentials and tested ability has demonstrated only very limited associations with student learning outcomes… *policy efforts over many decades to improve teacher credentials may have so restricted the range of differences among teachers that the remaining variability is irrelevant to the student learning outcomes…[e.g.,] the range of variability on the degree credential is restricted to the difference between baccalaureate and master's degrees… variables other than teacher quality, such as school policies, student readiness, and parents' expectations, may be so strongly associated with student learning that the contribution of the remaining variability in teacher quality is obscured."* (Boe and Gilford, 1992, p.40).

The above excerpt is from a National Research Council summary of the proceedings of a 1991 National Center for Education Statistics conference focused on teacher supply, demand, and quality. It clearly suggests that there may be little variation in teacher quality for teachers who hold the same teaching credentials. But, the summary also suggests that another explanation for the "limited associations" is that "available measures of teacher quality may have little relevance to the quality of teaching practice." As it turns out, the measures of teacher quality that existed back in the 1980s and 1990s do appear to have little relevance for student test achievement, but today we know that *there are significant differences in the contributions of teachers to student achievement amongst teachers who appear similar based on their credentials.*

The significant advance in our knowledge about teachers was made possible by administrative datasets that track teachers and students over time and link the two, and include year over year measures of student achievement, permitting the estimation of value-added models.[1] Such datasets became much more widely available subsequent to the passage of the *No Child Left Behind* (NCLB) Act in 2002 because NCLB required states to administer annual academic assessment tests in reading and math for every child in grades three through eight by the 2005-06 school year.[2] And NCLB data ushered in a wave of teacher research that has subsequently led to a policy environment today focused on reforming the human capital policies that govern the teacher workforce.

This chapter focuses on what research has shown about the impact of teacher quality on student achievement, the connection between quality and observable teacher characteristics and traits, and what all this might mean for teacher policy. The most important impacts of teacher quality are never directly observed – that is, we care how teachers affect students along both cognitive and non-cognitive dimensions, and we can only assess how students receive teachers' instruction using secondary measures. There are a variety of secondary measures that might be used: principal assessments, observations of teachers, student growth/value-added measures, student assessments of teachers, etc. There is a growing body of research on the degree to which these different measures of teachers align with one another (Grossman et al., 2010; Hill et al., 2011; Jacob and Lefgren, 2008; Bill and Melinda Gates Foundation 2010b; Ferguson 2011; Kane et al., 2010; Rockoff et al., 2011). The cited research generally suggests that different measures of teachers tend to agree with each other (albeit somewhat weakly in cases), but this is certainly not always true.[3]

1 Such datasets only became available for research at the state level in the late 1990s in select states (e.g. Tennessee and Texas), but are now exist in a great many more states (see www.dataqualitycampaign.org).

2 Prior to the existing of these administrative datasets it was much more common for research on teachers to be based on data collected by the U.S. Department of Education. Datasets like the National Education Longitudinal Survey of 1988 did link teachers and students, but they were samples of each so were not well-suited to address questions of how much variation in teacher effectiveness exist in the teacher workforce.

3 Non-alignment may be due to imprecision in the measures or because they are identifying different dimensions of teacher quality.

This chapter concentrates on research that focuses on student achievement on standardized tests, and value-added measures of teacher effectiveness (henceforth I use the terms "teacher quality," "teacher effectiveness," and "teacher performance" interchangeably). This is partially for the sake of parsimony, but there are also three compelling arguments for focusing on student achievement.[4] First, our schooling policies in this country are constructed significantly around test-based accountability; through the democratic process it has been decided, rightly or wrongly, that student and schools are to be held accountable for achievement on state assessments.

Second, value-added measures of teachers have been shown to be predictive of both student achievement in the future (e.g. Goldhaber and Hansen, 2010, forthcoming; Jacob and Lefgren, 2008; Kane and Staiger, 2008) as well as a variety of later life outcomes (Chetty et al., 2011). Finally, there is a matter of practicality: there is generally very little documented variation in teacher evaluation scores (Weisburg et al. 2009); as described below, new policies are being implemented across the country that are designed to change this.[5]

THE VARIATION IN TEACHER EFFECTIVENESS AND TEACHER CREDENTIALS

The import of value-added assessments of individual teacher quality began to be widely discussed in the late 1990s, based at least in part on a paper written by William Sanders and June Rivers (1996), and given wide distribution by the Education Trust.[6] This paper, for instance, notes that:

> "… students benefiting from regular yearly assignment to more effective teachers (even if by chance) have an extreme advantage in terms of attaining higher levels of achievement. (The range of approximately 50 percentile points in student mathematics achievement as measured in this study is awesome!!! Differences of this magnitude could determine future assignments of remedial versus accelerated courses.)" (Sanders and Rivers, 1996, p. 7).

Sanders is often credited as the "father of value-added" (e.g. Ewing 2011), but the use of education production functions to assess individual teachers actually predated the Sanders paper by a number of years. Eric Hanushek, for instance, noted in a 1992 article that:

> "… the estimated differences in annual achievement growth between having a good and having a bad teacher can be more than one grade-level equivalent in test performance." (Hanushek, 1992, p. 107).

Regardless of the academic origination, it is clear that the teachers matter message has entered into the mainstream, as influential writers such as Nicholas Kristof (2010) and Malcolm Gladwell (2008) have recently addressed the importance of focusing on the magnitude of teacher quality effects. And there are good reasons for this. Teachers are likely the most important resource over which schools have direct control (Goldhaber et al., 1999), swamping the impact of other investments like reductions in class size (Rivkin et al., 2005). Finally, differences in teacher quality affect more than just student test scores. An important new paper by Chetty et al. (2011) suggests that differences in the value-added of teachers explain such outcomes as college attendance, teen pregnancy and labor market earnings. In fact, the authors calculate that, on average, a one standard deviation improvement in teacher value-added in a single grade raises earnings by about 1 percent at age 28.

4 Value-added is an estimate of teacher contributions to student learning on standardized tests. For a more detailed argument for using value-added measures in the context of teacher evaluations, see Glazerman et al. (2010).

5 When all you have is the hammer of value-added, everything looks like a nail; in other words, using student achievement-based measures to assess teachers becomes a necessity if there is no variation in other measures of teacher quality

6 See, for instance, the Education Trust publication, "The Real Value of Teachers," Thinking K-16, vol. 8, issue 1.

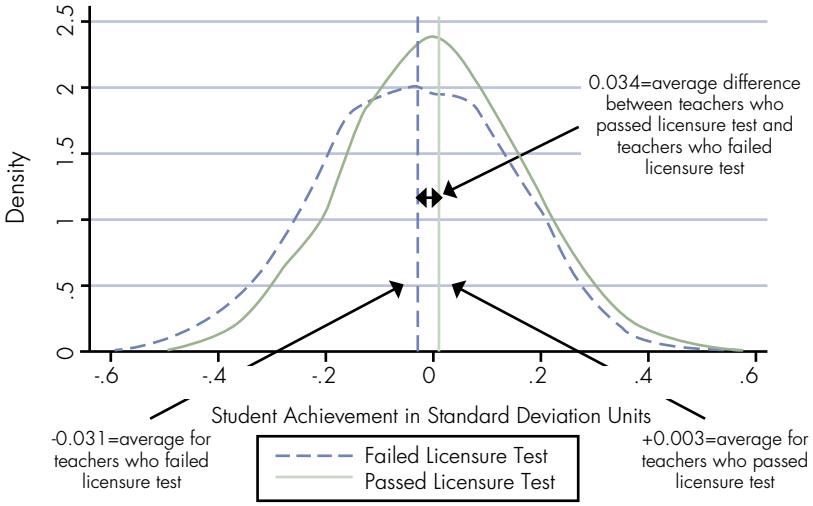
Statisticians discuss the importance of teacher quality in standard deviation terms, so, for the discussion that follows, it is useful to have a frame of reference for thinking about the importance of differences between teachers. Empirical estimates of the effect size—i.e. the estimated effect of a 1 standard deviation change of teacher effectiveness on student achievement—are in the neighborhood of 0.10 and 0.25 (e.g. Aaronson et al., 2007; Goldhaber and Hansen, forthcoming, Hanushek and Rivkin, 2010; Nye et al., 2004; Rockoff, 2004).[7] To put these effect sizes into perspective, a one standard deviation change in teacher effectiveness is equivalent to the difference between having a teacher at the 31st percentile of the performance distribution versus the 69th percentile, or the 50th versus the 84th.[8] And, as students move from one grade to the next, they typically gain about 1 standard deviation in math and reading achievement in the lower elementary grades, half a standard deviation in the upper elementary grades, a third of a standard deviation in middle school, and a quarter of a standard deviation in high school (Bloom et al., 2008). And, the average gap between black and white students, or economically advantaged and disadvantaged students, is on the order of magnitude of 0.7 to 1.0 standard deviations (Hanushek and Rivkin, 2010). Putting this all together, this means that having a highly effective teacher (i.e., in the 84th percentile) rather than an average teacher, likely makes a difference of 10 to 50 percent of a grade-level's worth of achievement in elementary school, and/or could cut achievement gaps down by 10 to 35 percent.[9]

While the statistics cited above show the importance of *individual* teacher effectiveness, teacher human capital policies are typically built around three teacher *credentials*: licensure, experience, and degree level.[10] For instance, teachers are required in all states to be licensed in order to be eligible to participate in the teacher labor market. The specific requirements differ from one state to the next, but typically entail graduation from an approved teacher training program, and passing one or more licensure tests (Goldhaber 2011a).[11] And, once teachers are in the profession, their salaries are generally determined by their degree and experience levels (Strunk and Grissom 2010; Podgursky 2011). A decade ago it was controversial to suggest that these credentials are not correlated, or only weakly correlated, with student achievement.[12] Today we know that even when there is good empirical evidence that a credential is associated with student achievement across educational context (e.g. grade levels, subjects, etc.), as is the case with early career teacher experience, the differences that exist between teachers who hold the same credential swamp the differences between teachers holding different credentials.[13]

7 These estimates vary by grade level, subject, and whether the measure is based on the estimated total variation in teacher effectiveness, or the within school variation alone.
8 I am assuming a normal distribution.
9 Importantly, knowing that this is the case is not the same as being able to act on it. In particular, it is only known how effective a teacher is, in value-added terms, after a class has been taught. Thus policies relying on value-added would need to use past predictions, which may not always be a good indicator of future teacher performance, either because of bias in the original estimates (Rothstein, 2010) or because the estimates are not very reliable (Goldhaber and Hansen, 2013; McCaffrey et al., 2009). I discuss these issues briefly in the following section.
10 For a more comprehensive review of the relationship between these, along with other teacher credentials and characteristics, and student achievement, see Goldhaber (2008).
11 Most states now allow for some alternative pathways into the profession, and these often permit teachers to teach for a specified period while obtaining formal teacher training. For more discussion, see Boyd et al. (2007) or Goldhaber (2011).
12 See, for instance, the debate about the evidence on teacher licensure (Darling-Hammond et al., 2001; Goldhaber and Brewer 2000a, 2000b) or degree and experience levels (Greenwald et al., 1996; Hanushek, 1997).
13 Of course we know the relationship between teacher credentials and student outcomes more definitively today than was the case several decades ago, but it is worth noting that doubts about the relationship between teacher characteristics and student outcomes were clearly voiced back in the mid-1980s (Hanushek, 1986).

The saying "a picture is worth a thousand words" applies in the case of making clear the relative importance of differences between individual teachers who hold a common credential versus differences between teachers who hold different credentials. Figure 9.1 is derived from analysis of the ability of licensure tests to predict students' math achievement at the elementary level in North Carolina (Goldhaber, 2007). It shows the estimated distribution of effectiveness—the x-axis, which is measured in student test terms—for teachers who passed required licensure exams (the solid line) and those who failed them (the dashed line).[14]

Figure 9.1. Teacher Effectiveness in Mathematics by Licensure Test Passing Status.



One might expect teachers that pass licensure exams to be more effective than those that do not, and, in fact, that is what is found. The horizontal distance between the means of each distribution, which is about 3 percent of a standard deviation, represents the statistically significant differential between those who passed and those who failed. But what is telling is the degree of overlap in the distributions, 92 percent, of those who passed and those who failed the test.

A couple of comparisons help put the above finding in perspective. First, a one standard deviation change in teacher effectiveness of the effectiveness distribution (in North Carolina at the elementary level) is estimated to correspond to an increase of student achievement of about 0.2 standard deviation of student achievement (Goldhaber and Hansen, forthcoming). This means that the differential between average (50th percentile) and effective (84th percentile) teachers who are in the same category (i.e. they either passed or did not) is roughly six times as large as the typical difference in performance between those who pass or fail the required licensure tests.[15]

Second, in Figure 9.1, all the teachers lying below the dashed line but to the left of the average effect of teachers who failed the test passed the required licensure tests but were *less* effective than the average teacher who failed to achieve the standard, whereas the teachers below the solid line but to the right of the average effect of test passers failed the standard, but were *more* effective than the average teacher who passed. As is apparent from the figure, there are a significant number of teachers who perform poorly on licensure tests, but turn out to be quite effective in the classroom and vice versa.

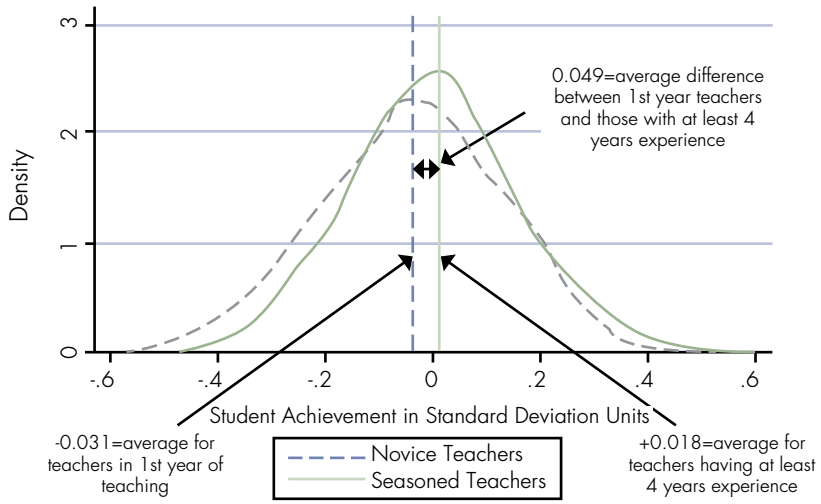
14 At the time of the study elementary teachers in North Carolina were required to pass two Praxis licensure tests. For more detail, see Goldhaber (2007).

15 Nor is this an isolated finding with respect to licensure. Kane et al. (2007), for example, examine teachers in New York City and estimate that the typical gap in teacher effectiveness within a licensure category (between top and bottom quintile teachers) is about ten times larger than the average gap between categories.

The findings on the relationship between a teacher's licensure status, and the components of licensure, are bound to vary depending on state, subject area, and grade level because licensure requirements also vary along these dimensions. The findings for teacher experience, by contrast, are much more robust across educational context. *Early career* teaching experience is the one teacher credential that is typically found in empirical studies to be predictive of teacher effectiveness (Clotfelter et al., 2010; Goldhaber, 2007; Goldhaber and Hansen, forthcoming; Harris and Sass, 2011; Hanushek and Rivkin, 2007; Master et al. 2012; Rice, 2010; Rivkin et al., 2005; Rockoff, 2004).

But, while early career teaching experience may consistently predict teacher effectiveness, it is not necessarily an accurate predictor in the sense that knowing a teacher's experience level provides a lot of information about the teacher as an individual. Here again, a picture, based on existing empirical evidence, is useful to illustrate the point. In this case, I draw upon research at the elementary level in math from Washington State (Goldhaber et al., 2012). Figure 9.2 plots the value-added distributions of effectiveness for 1st year teachers as compared to teachers at least 4 years of experience.[16]

Figure 9.2. Teacher Effectiveness in Mathematics by Experience Level.



The above figure is quite similar to Figure 9.1 on licensure in that, as expected, more seasoned teachers (represented by the solid line) tend to be more effective than novices (represented by the dashed line), but there is still considerable overlap in the distributions. The Washington study finds that, all else equal, students with novice teachers score about 5 percent of a standard deviation lower on achievement tests in math than students with teachers with 4 or more years of experience. And while 5 percent of a standard deviation difference is not trivial, it is swamped by the typical differences between teachers who fall into the same experience category. In particular, the overlap in the two distributions is about 91 percent and the differential between average (50th percentile) and effective (84th percentile) teachers who are in the same category is between 3 and 4 times the average difference between these two experience groups.[17]

16 Four years was chosen because much of the research, cited above, on teacher experience suggests the productivity gains associated with additional years of experience tend to level out after a teacher's first 4 or 5 years in the classroom.

17 This is based on a finding in Washington that a 1 standard deviation change in teacher effectiveness is associated with a change of 0.18 standard deviations of student achievement in math.

Educational attainment, usually holding a master's degree, is the other teacher credential that is typically used in salary determination. But, unlike teacher experience, there is very little empirical evidence suggesting that a teacher holding a *generic* advanced degree is predictive of student achievement (Aaronson et al., 2007; Clotfelter et al., 2010; Goldhaber and Brewer, 1997; Harris and Sass, 2011; Podgursky and Springer, 2007; Rivkin et al., 2005).[18] Given this, it is pointless to show distributions like those in Figures 1 and 2 for there would be nearly total overlap between the distributions of effectiveness for teachers with and without a master's degree. And, the master's finding is so pervasive in the literature that Education Secretary Arne Duncan was willing to go out on a limb by noting that:

> "Districts currently pay about $8 billion each year to teachers because they have masters' degrees, even though there is little evidence teachers with master's degrees improve student achievement more than other teachers–with the possible exception of teachers who earn master's in math and science." (Duncan, 2010)

Does the above evidence suggest that credentials-based teacher policies should be abandoned? That is obviously a matter in the eye of the beholder, and researchers have reached very different conclusions about this. Gordon et al. (2006), for instance, argue that it makes sense to be less restrictive about who is allowed to enter the teacher workforce and more rigorous, based on information about in-service performance, about who is allowed to stay in it. Clotfelter et al. (2010), by contrast, believe that there are educationally meaningful differences in student achievement associated with having teachers with different combinations of credentials (e.g. licensure plus experience plus competitiveness of college attended, etc.).[19] Tempering this assessment, however, is the fact that our policies today are not based on combinations of credentials. And, moreover, new research points to some limits of what we may be able to learn about teachers before they are actually teaching. Rockoff et al. (2011) assess the degree to which a comprehensive collection of quantifiable information about prospective teachers, including measures of cognitive ability and content knowledge, personality traits, reported self-efficacy, and scores on teacher selection instruments, can predict teachers' value-added. They find this information explains only a small proportion of the variation in effectiveness, just over 10 percent of the variance of the expected value-added distribution of teacher effects.

If there were strong causal evidence about the connection between teacher credentials and student achievement, it would be relatively trivial to adopt teacher policies that greatly improve the effectiveness of the teacher workforce. Again, it is a matter of opinion, but my view is that acting on the differences we observe amongst in-service teachers—the topic of the next section—is a much more promising area of reform.

18 An exception to this finding is a degree in math at the high school level when teaching math (Goldhaber and Brewer, 1997), and there is some suggestive evidence that a degree in subject may also be predictive of effectiveness at the middle school level (Harris and Sass, 2011). For a more thorough review, see Hanushek and Rivkin (2004).

19 For instance, they estimate that students with teachers with a weak combination of credentials would score about 0.23 standard deviations lower than students with teachers with strong combinations of credentials.

POLICIES DESIGNED TO ACT ON TEACHER DIFFERENCES

The confluence of two findings has led to a major change in the way we think about in-service teacher evaluation. The first is the topic of the above section: the empirical evidence on the importance for students of the variation in in-service teacher effectiveness. The second, juxtaposed against the first, is the now widespread finding that there is virtually no variation in teacher performance ratings based on in-service evaluations. *The Widget Effect* (Weisburg et al., 2009), a study of evaluation practices in 12 school districts across 4 states, showed that while the frequency and methods of teacher evaluation varied, the results of evaluations were almost always the same. In particular, over 99 percent of teachers in districts with a binary evaluation system are rated as satisfactory, and over 94 percent of teachers in districts with a multi-tiered rating system receive one of the top two ratings, with less than one percent of teachers receiving an unsatisfactory rating. The findings from *The Widget Effect* quantified what was already widely believed: teacher in-service evaluation is not a very rigorous process.[20] Not surprisingly these findings have led to calls to reform teacher evaluation processes so that they are made more rigorous and recognize individual teacher differences (e.g. Chait and Miller, 2010; Toch and Rothman, 2008).

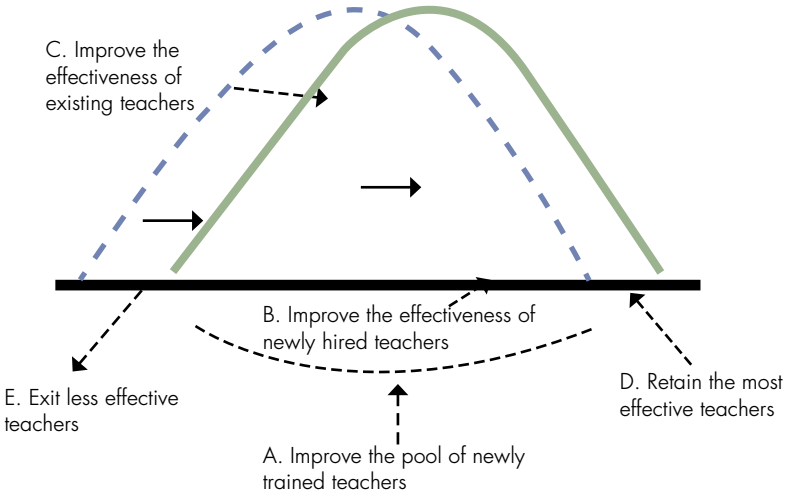
And reform has come. We are in the midst of an ongoing revolution in teacher policy in a great many states today. Jump-started by federal programs like the Teacher Incentive Fund and Race to the Top (RttT), states are implementing new systems designed to radically change policy by requiring outcome-based (including student achievement on standardized tests) measures be used as a component of evaluations, and requiring that the evaluations themselves be used for high-stakes purposes such as licensure, compensation, and tenure.[21]

In theory, most, if not all school human resource policies and systems could key off of in-service evaluations. To explore the theory of action around the use of performance evaluation and changes to the quality of the teacher workforce, it is useful to return to the now familiar teacher effectiveness distribution figure. The ways in which knowing more about the effectiveness of in-service teachers, *and acting on teacher differences*, might translate into a more effective teacher workforce are illustrated in Figure 9.3.

20 While The Widget Effect is probably the most well-known research buttressing the lore that teacher evaluations do not adequately differentiate teachers, it is by no means the sole evidence. See, for instance: Bridges and Gumport (1984); Education Research Services (1988); Lavely et al. (1992); Tucker (1997).

21 For an overview, see Goldhaber (2010a).

Figure 9.3. Improving the Quality of the Teacher Workforce: Theory of Action.



The dashed line in the figure represents the hypothetical distribution of quality of the workforce now and solid line represents a hypothetical shift of the distribution to the right, representing an improvement in workforce quality. What are the pathways through which new in-service evaluations might affect quality? First, the hope is that the pool of potential teachers will be improved, Pathway A, given that recognizing and rewarding success in the teacher labor market might provide feedback to institutions preparing prospective teachers and/or that more able individuals will see teaching as a desirable profession knowing that they are more likely to be rewarded for success. Connected to this, we anticipate, Pathway B, that the more able of the pool of potential teachers are hired into school systems. Of course this means that school systems would probably need to develop better teacher selection mechanisms than exist today since existing evidence suggests only a modest relationship between pre-service credentials and characteristics and teacher effectiveness (Rockoff et al., 2011).

It is expected that teachers who are being rewarded for improving student learning—either through increased compensation or by maintaining the eligibility to participate in the labor force —to focus their instruction on this effort. Moreover, to the degree that new evaluation systems improve performance feedback and professional development, existing teachers might be expected to become more productive.[22] Thus, the effectiveness of existing teachers would be expected to increase, Pathway C.

Systems that reward teacher success should encourage the retention of successful teachers, Pathway D. Finally, new teacher human capital systems are designed to identify teachers who are very ineffective and exit them from the workforce (e.g. firing, non-tenuring, etc.), often referred to as "teacher de-selection," Pathway E.

22 This may result from teachers learning about areas of weakness and addressing them through professional development, or through less formalized channels, such as role-modeling effects. Jackson and Bruegmann (2009), for instance, find that teachers perform better when they have more effective colleagues and that these effects are strongest for less experienced teachers. The results suggest that teachers' productivity is influenced through learning and knowledge spillovers that occur among peers.

Unfortunately, while the theory of action linking reform to teacher effectiveness is pretty simple, translating the theory into positive results may turn out to be both logistically and politically difficult. There exists both opposition to the notion of high-stakes accountability for teachers (Strauss, 2012; Wall Street Journal, 2010), and considerable debate about what measures of in-service performance ought to be used in making high-stakes decisions (Boser, 2012; Education Week, 2011).[23]

There exist a handful of studies examining the ability of in-service assessments to predict student achievement. Several studies show that value-added measures in one year predict student achievement in another year (Goldhaber and Hansen, 2010; Jackson and Bruegmann, 2009; Jacob and Lefgren, 2008; Kane and Staiger, 2008).[24,25] But there are other in-service measures that also seem to be predictive of student achievement. Jacob and Lefgren (2008) and Harris and Sass (2009), for instance, both find that principals' private (and not documented) assessments of teachers predict teacher value-added, particularly at the tails of the performance distribution.[26] Rockoff and Speroni (2011) find that subjective assessments of first year teachers by their mentors predicts the achievement gains made by teachers' students in the future. Finally, several studies suggest that structured teacher observations are related to student achievement (Grossman et al., 2010; Hill et al 2011; Kane et al., 2010).

More recently, the Measures of Effective Teaching (MET) project has focused on the relationship between various methods of evaluating teachers—videotaped observations, student surveys, and value added – as well as the ability of these different measures alone or combination to predict teacher effectiveness and student achievement in the future. The MET studies showed that classroom observations (rated according to several different rubrics) and student surveys about their teacher and classroom environment are moderately positively correlated with teacher value added.[27] The MET project also found that these different measures of teacher performance in one period do predict future teacher effectiveness, but that the weight one would apply to the different potential components of a teacher's summative performance measure to optimally predict future teacher effectiveness depends on a subjective assessment of what one values in the classroom. Not surprisingly, measures of teaching effectiveness are generally best predicted by creating a baseline measure of teacher performance that heavily weights the component of effectiveness that is most valued. So, for instance, if the objective is to predict teacher classroom observations scores, then a prior measure of a teacher's performance on a classroom observation rubric should be weighted relatively more heavily in creating a summative measure than student perceptions survey results or value added. That said, however, each component measure of performance tends to add some explanatory power in the prediction of future teacher effectiveness so, for instance, when predicting student achievement on state assessments, prior measures of teachers' value added would be weighted most heavily, but combining value added with teacher observation and student survey measures generally leads to teacher performance measures that outperform value added measures alone.[28]

23 This is particularly true in regards to the extent to which student-growth/value-added measures ought to be employed as part of a teacher's evaluation (Darling-Hammond et al. 2012; Glazerman et al., 2010; Goldhaber and Chaplin, 2012; Rothstein, 2010).

24 And, as I mention in the prior section, Chetty et al. (2011) find that value-added also predicts a number of students' later life outcomes.

25 There is concern, however, that these value-added measures may be biased indicators of true teacher performance. For a discussion of this, see Chetty et al. (2011), Goldhaber and Chaplin (2012), Kane and Staiger (2008), and Rothstein (2010).

26 The stress on private assessment is an important caveat given that the evidence (e.g. Weisburg et al., 2009) is that assessment instruments, or cultural or political constraints cause principals not to reflect the true variation they perceive of teacher performance in documented ratings.

27 There were only minor differences in the correlations between the various classroom observation rubrics and value added; the correlation between the student perceptions surveys and value added was slightly higher. The various MET study reports can be downloaded from www.metproject.org.

28 Value added alone works well if the objective is to predict achievement on state assessments, but the combined measures work significantly better when the objective is to predict student achievement on more demanding assessments that were administered as part of the MET study.

The findings described above suggest there is significant potential to identify, using a variety of methods, which teachers are effective, and the possibility that evaluations and performance feedback may therefore be harnessed to improve the quality of the teacher workforce. As indicated in Figure 9.3, however, improvement could be achieved through different pathways. Upgrading the quality of the teacher workforce through changes to the productivity of individual teachers (Pathway C) already in the workforce is surely less politically difficult than fundamental changes to who is in the teacher workforce (pathways A and E). Unfortunately, current empirical evidence suggests that few interventions make teachers more productive. While professional development (PD) is a ubiquitous strategy used in an effort to improve teaching, and school systems invest substantial sums in this endeavor,[29] and there are a significant number of studies focused on relating both the content and mode of delivery of PD to teacher instructional practices and effectiveness, most research on PD is not rigorous and there are relatively few studies suggesting that it improves teaching.[30]

More recent and convincing research utilizing longitudinal data with detail on in-service participation in PD also shows little evidence that it influences teacher productivity consistently across grades and subjects (Harris and Sass, 2011; Jacob and Lefgren, 2004).[31] The one area where PD does seem to make a difference (at least in one study) is for math teachers at the middle and high schools levels *in math content*.[32] Unfortunately, several large-scale, well-designed, experimental studies tend to confirm the finding that PD has little identifiable impact on student achievement. In particular, a randomized control trial focusing on a one-year content-focused PD program (that was thought to be well designed) showed positive impacts on teachers' knowledge of scientifically based reading instruction and instructional practices promoted by the PD program, but no effects on student test scores (Garet 2008).[33]

The lack of evidence that professional development typically pays off in student achievement may be related to the institutional structure under which teachers typically receive it. A specified number of hours of PD is often a requirement of teachers who wish to maintain their eligibility to teach. And, as Jennifer King Rice put it in a review of the effectiveness of PD, "The crux of the problem appears to be an incentive structure that rewards seat time rather than improved performance. Policymakers need to consider better ways to structure incentives so that teachers engage in professional development directly related to their work with the potential to improve their teaching practices and, ultimately, their effectiveness." (Rice 2009, p. 243).

It is quite likely that the system in which PD is delivered will be different in the future as many of the new teacher policy reforms involve incentivizing teachers in various ways. As noted above, RttT states are implementing systems that attach high-stakes to teacher performance evaluations. This means teachers should have increased incentive in the future for performance given concern about job security, both maintaining tenure and the possible loss of future employment eligibility.[34]

29 Estimates are typically in the $2000-$4000 per teacher per year range (Rice, 2009).

30 See, for instance, Yoon et al. (2007) for a comprehensive review. The research on PD that is most encouraging suggests that focusing on how students learn a content area tends to be more effective than PD emphasizing pedagogy/teaching behaviors or curriculum (Cohen and Hill, 2000; Kennedy, 1998; King, 2009) suggests.

31 In particular, this reseach shows whether the receipt of PD appears to have a human capital impact on teacher productivity as opposed to the possibility that an association between PD and student performance exists because of self-selection into PD participation.

32 In-service hours of PD was not found to be statistically significant at other grade levels, or in reading, or in math for non-math content PD (Harris and Sass, 2011).

33 Mentoring and induction (a more structured form of mentoring that is instructionally focused and delivered by full-time, trained mentors) might also be thought of as a type of professional development. A recent review (Ingersoll and Strong, 2011) of studies of induction suggests its promise, another a recent large-scale RCT investigation of this type of PD (Glazerman et al., 2010) did find some evidence that teachers who received two years of comprehensive induction had higher levels of student achievement in the third year (however, not in the first two years of receiving additional support).

34 There is some evidence that teachers respond to career concerns – see Fama (1980), Hansen (2009), and Holmström (1982) for a discussion of the career concerns literature – by increasing effort (measured based on absences) when trying to impress a new principal (Hansen, 2009). This behavior makes sense when one considers the fact that principals do typically have some discretion about non-pecuniary aspects of a teacher's job, such as choice over classroom assignments (Player 2009).

There are also an increasing number of school districts that are experimenting with explicit monetary incentives for teachers, known as merit pay or pay for performance (PFP).[35] The empirical evidence of PFP as a system to improve student achievement is mixed (e.g. Ballou, 2001; Figlio and Kenny, 2007; Glazerman and Seifullah, 2010; Goldhaber and Walch, 2012; Fryer et al., 2012; Podgursky and Springer 2007; Springer et al., 2010a).[36]

One of the arguments for the mixed evidence of PFP is that many performance plans are not well-designed, and/or that the design of the research on the PFP plan was not appropriate to detect effects (Imberman and Lovenheim, 2013). But several recent randomized control trial experiments challenge this proposition as they provide evidence, across different policy designs, that PFP has little effect on the productivity of the workforce, even in the case of well-designed PFP systems.

A randomized experiment in New York City (Marsh et al 2011) offered school-wide bonuses of $3000 per teacher to schools in the treatment group that met performance standards set by the school district. The study found that bonuses did not improve student achievement in any grade level and had no effect on school progress scores. There are multiple potential explanations for the fact that PFP did not seem to lead to increased teacher productivity. Marsh et al (2011), for example, offer four potential explanations for the lack of change under the school-wide bonus system in New York City. First, the program may have been too new to produce effects. Second, teachers may have misunderstood the program due to poor communication from schools. Third, motivation may simply not be enough to improve schools. Finally, the size of the bonuses may have been insufficient to motivate teachers.

There are also potential psychological explanations. Research on the conditions under which incentives might influence behavior suggests that explicit incentives (e.g. monetary rewards for performance) can "crowd-out" the incentivized behavior (Gneezy et al., 2011). The argument is that an individual's utility may depend not only on the monetary rewards for engaging in an activity, but also the enjoyment of the activity itself, an individual's self-perception of why they are engaging in the activity, or an individual's perception of the reasons that others may believe they are engaging in the activity. It is thus conceivable, on average, that individuals reduce effort because they worry, in the case of teaching, that working hard for the incentive might be perceived as doing it for the wrong reasons (i.e. affecting either their own self-perception or how they think they may be perceived by others if they were to work harder to achieve the monetary reward).[37]

Another possibility is that school-based incentives have little impact in schools. The concern here is that group incentives dilute the influence of the program on individuals and induce a free rider problem (Lavy, 2007; Podgursky and Springer, 2007; Prendergast, 1999).[38] However, another experiment (the "POINT" experiment) that focused on individual level teacher incentives, sponsored by the U.S. Department of Education, offered substantial bonuses to teachers in the treatment group whose students made large test score gains: $15,000 for teachers with student gains at the 95th percentile or higher, $10,000 for teachers at the 90th percentile or higher, and $5,000 for teachers at the 80th percentile or higher (Springer et al 2010a). Despite the magnitude of these bonuses, the researchers found no consistently significant difference between the performance of teachers in the treatment and control groups (the exception was 5th grade teachers in the 2nd and 3rd years of the study). The most encouraging experimental evidence on pay for performance in U.S. schools comes from a recent study by Fryer et al. (2012).[39] This study shows that teachers are more likely to respond to loss aversion

35 PFP is not new, a number of districts implemented PFP in the 1980s following the release of A Nation At Risk report. These were generally short-lived efforts, and deemed to be unsuccessful (Hatry et al., 1994; Murnane and Cohen, 1986).

36 See Goldhaber et al (2008) and Podgursky and Springer (2007) for general reviews of the PFP literature.

37 See Gneezy et al. (2011) for a review of the kind of activities that may be positively or negatively influenced by explicit incentives as well as studies that suggest this crowing out phenomenon has been observed in a variety of contexts.

38 Other recent studies of PFP that find no effect suggest that the program may have weak incentive effects because teachers, when given the choice, opt to distribute rewards equitably amongst all teachers in a school (Springer et al 2010b).

39 There is evidence from international that pay for performance increases teacher productivity.

(i.e. the threat that compensation that is in hand will be taken away if student performance is not good enough) than the possibility of receiving a bonus. It is not at all clear how one might operationalize this sort of incentive in the teacher labor market given political and cultural constraints in public schools, but the Fryer et al. finding is quite important as it shows the potential for policies to impact the effectiveness of the existing teacher workforce.

The above findings may suggest the prospects for PFP changing the quality of the teacher workforce are not promising, but, while it is a judgment call, my view is that it is too early to write off the potential individual teacher productivity effects of incentives. There are several reasons to take this position. First, the structure of rewards and extent to which teachers understand the incentives facing them will clearly matter when it comes to inducing behavioral changes. The POINT experiment described above is a convincing case where large individual bonuses did not appear to systematically encourage teachers to work harder or smarter such that student achievement gains were induced. That does not mean that a differently structured incentive might not work better.

Second, the studies cited above are only designed to test the short-run effects of incentives on individual productivity. It is conceivable that individual teacher productivity may change over time (it may take longer than the 3 years in the studies cited above) as a consequence of incentives as teachers learn how to improve through PD or from each other. Related to this, the effect of incentives may well be dependent on other contextual factors. Some non-experimental evidence on performance incentives suggest more promising findings. Goldhaber and Walch (2012), for instance, in an analysis of Denver's well-known ProComp PFP system, find that student achievement increased after the implementation of the pay reform but conclude that the changes to achievement cannot be narrowly tied to teachers participating in the new incentive system. Rather they suggest that other ancillary systems—such as evaluation and performance feedback—may be the cause of the increased achievement. It is, of course, an open question as to whether changes to these ancillary systems would have taken place in the absence of the need, because of the new PFP system, to evaluate and differentiate teachers in order to assign performance bonuses.

Perhaps more importantly, a relatively new literature on personnel economics (e.g. Lazear, 2000; Ichniowski and Shaw, 2003) suggests that a substantial portion of the benefits of PFP in the broader labor market comes through labor market sorting, i.e. hiring and labor market selection.[40] Specifically, the theory around PFP would suggest that individuals who will be rewarded under a PFP system will be attracted into the teaching profession and district officials will get the right signals about who amongst the applicant pool to hire (pathway E) because the performance assessments of in-service teachers provide a clear signal about what kinds of prospective teachers tend to be effective. The same principle holds for labor market selection. Individuals rewarded for performance through PFP systems would be more likely to stay in the profession on their own accord (pathway D).[41]

Evidence based on Washington D.C.'s human capital reforms (known as "IMPACT"), which include changes to teacher pay, professional development, and retention policies shows that at least some evaluation reforms are functioning as anticipated and having positive effects. Dee and Wyckoff (2013) find that teachers who are deemed by the evaluation system to be highly effective, who become eligible for a large increase in base pay if they are again classified as highly effective the subsequent year, are likely to become more effective. And, at the other end of the teacher performance distribution, they find that teachers judged to be ineffective, and threatened with dismissal if they were to continue to be found ineffective, were far more likely to voluntarily leave the district.[42]

40 Lazear (2000) pegs the figure at roughly 50 percent.

41 For a more comprehensive discussion of these labor market sorting issues, see Eide et al. (2004), Lazear (2003), and Podgursky and Springer (2007).

42 These findings are based on a regression discontinuity analyses so, as the authors note, should not be interpreted as average treatment effects since they only apply to teachers who are near to the high-stakes cut points. Yet this is precisely the way IMPACT is designed to affect the teacher workforce in Washington so the findings are clearly encouraging.

Ultimately the effect of new evaluation systems on the teacher labor market, particularly whether they lead to meaningful labor market sorting effects, depends on the extent to which these new systems do differentiate teachers. It clearly is not enough to simply change the method by which teachers are evaluated or the rating scale as a variety of different systems result in a Lake Wobegon effect where all, or nearly all, teachers are clustered near the top (Goldhaber, 2010b; Toch and Rothman, 2008; Weisburg et al., 2009), and the politics of making evaluations high-stakes are certainly challenging (Goldhaber, 2009 & 2010b). Washington D.C.'s IMPACT system appears to be an outlier in terms of differentiating teachers. Under IMPACT there are significant numbers of teachers who are falling into performance categories that suggest they need improvement, but the early indication is that most state teacher evaluation systems that have been overhauled do not look like they are differentiating teachers much more than the evaluation systems they replaced (Anderson, 2013). The seeming success of the human capital reforms in D.C. provides some cursory evidence that new evaluation systems might have to "have teeth" to induce behavioral changes amongst teachers. More generally, the impact of evaluation reform will also hinge on whether and how the evaluations are used, and the way that individuals respond to what might be a radically different human capital system in teaching.

Changing the mix of who is in the teacher workforce is now part of the policy discussion in the way that it was not a decade ago. One reason for this is that international comparisons of student achievement show U.S. students are, on average, in the middle (or lower end depending on the interpretation of the data) of the pack of tested nations, and, in contrast to the situation in the U.S., many of the countries at the top of the test distribution also have teachers who come from the upper end of the performance distribution (McKinsey, 2009).[43] On top of this, there is speculation (Ingersoll & Merrill, 2012; Cohen & Varghese, 2012) that part of the explanation for the fact that increasing U.S. investments in K-12 education have not, over time, resulted in better student achievement is that this corresponds to a period with long-term decline in the average academic caliber—measured by standardized test scores and/or the selectivity of colleges—of teachers (Corcoran et al., 2004; Hanushek and Pace, 1995; Hoxby and Leigh, 2005).[44] Unfortunately, even if it is desirable to make the U.S. teacher pipeline and training process look more like a country such as Finland (Darling-Hammond and McCloskey, 2008) there are a number of structural barriers to doing so. Chief among these is the fact that teacher training is regulated at the state level and there are over 1,500 training providers, meaning there is no teacher training curriculum and little centralized control over who is granted eligibility to teach (Goldhaber, 2011a).

Several recent studies exploring the potential implications of changing the mix of people in the teaching profession through de-selection policies (e.g. firing, layoffs, non-tenuring) suggest that even small changes could have profound implications for student achievement. Hanushek (2009), for instance, makes an empirically-based case that were 5–10 percent of the least effective teachers (2–3 teachers in a school of 30 teachers) removed from the workforce and replaced with teachers of average effectiveness, achievement of U.S. students would rise toward the top in international comparisons. This in turn would increase U.S. growth by 4.5 percent, an amount that sounds small, but it is large enough that it would still pay for all 2008 national spending on K-12 education (Hanushek, 2009, p 169).[45] And, the more recent study by Chetty et al. (2011), suggests that replacing a teacher whose true value-added is in the bottom five percent with an average teacher would increase students' lifetime income by $267,000 per classroom taught. Similarly, both Goldhaber and Theobald (2013) and Boyd et al. (2010) investigate the differences in the teachers who would be laid off under a seniority-based system versus a value-added effectiveness based system, and conclude that employing one system versus another has educationally meaningful implications for student achievement. In short, teachers laid off under an effectiveness versus a

43 This, of course, does not mean that there are not a great many teachers in the U.S. with very high test scores and/or who graduate from selective colleges and universities, just that on average they do not. There is also new evidence that this situation has changed when it comes to newer entrants into the teacher labor market (Goldhaber and Walch, 2014), though this may simply be one of the consequences of the Great Recession.

44 There are two major, not mutually exclusive, explanations offered for this decline. One is that labor market opportunities for woman and minorities have expanded significantly since the 1960s, reducing what used to be tantamount to a subsidy to education. The second is that the compression in wages in the teacher labor market, possibly due to unionization and collective bargaining, discourages productive people from choosing teaching as a career.

45 The Hanushek estimates are based on empirical evidence on the links between teacher quality and student achievement, and student achievement and country growth rates.

seniority-based system are dramatically less effective (meaning that more effective teachers would be retained under an effectiveness-based system). The average difference between the effectiveness of teachers laid off under the two systems is between 20–30 percent of a standard deviation in student achievement, which is roughly equivalent to a full-standard deviation in teacher effectiveness (e.g. the difference between a teacher at the 16th percentile versus the middle of the teacher performance distribution).

But while calculation like those discussed above suggest intriguing possibilities when it comes to using new evaluation systems for more purposeful de-selection, they are based on what we know about the impact of teacher effectiveness on student achievement, not actual policy variation. De-selection of the sort described above, or really any documented de-selection of teachers, is exceedingly rare.[46] It is conceivable that making teaching a riskier occupation, in the sense that there is greater uncertainty about job security—or compensation in the case of PFP—could have far-reaching and unintended consequences on who opts to enter the teacher labor force and how in-service teachers behave.[47] As I discuss in the next section, this means the policy decision over how to use evaluations is somewhat of a judgment call.

CONCLUSION

As is hopefully clear from the review of research and policy above, the benefits of increasing the quality of the teacher workforce are profound, but things get murkier when it comes to delineating a clear path toward this objective.

The difficulty is two-fold. First, teacher effectiveness is not well-predicted by the teacher credentials—licensure, degree and experience levels—that are now utilized for high-stakes purposes. Second, there is little evidence that in-service professional development leads to changes in teacher practices that result in significant improvements in student achievement. As mentioned above, it is possible that PD might be more effective under a different incentive system. Teachers obviously can improve as is apparent from the well-documented returns to early career experience, but there is simply no evidence that PD delivered under today's system can, for instance, raise the effectiveness level of an average novice teacher to that of a third-year teacher.

Changing who is in the profession is likely to either be slow (through attrition) or politically challenging. Changing teachers is not like substituting one component for another in the production of a product in the private sector. As noted elsewhere (Goldhaber, 2011b), "[t]he transistor may be superior to the vacuum tube, but the vacuum tube doesn't have a position about whether it is employed. Incumbent teachers, on the other hand, clearly do. Moreover, this opinion carries weight. Not only is teaching the single largest profession, but teachers tend to be politically mobilized."

There are three possible options for changing the policies governing employment and compensation in public schools. First, some have argued that technology will disrupt the current system (Christensen et al 2008) and could lead to different teacher career paths and an extended reach of highly effective teachers (Hess, 2009 & Rhim et al, 2007). Highly skilled teachers might have differential career paths – supervising other teachers, for instance – or teach students remotely across schools in or across districts.[48] The arguments for the use of technology to help spread teaching talent to more students are compelling, but they are both untested and likely to face political opposition (Hill, 2009).

46 Space precluded a full discussion of the teacher pipeline, but one critique of teacher training programs is that they too generally do not actively deselect prospective teachers who are working toward obtaining a teaching credential. While there is speculation on this point, there is relatively little in the way of empirical evidence on the efficacy of the quality control processes within teacher training institutions.

47 This point is made empirically in a paper by Rothstein (2013), which provides simulation evidence showing that the estimated effects of high-stakes policies, such as selective tenuring and dismissal, are sensitive to assumptions about the behavioral responses of teachers (and prospective teachers) to the use of such policies.

48 See Hassel and Hassel (2009) for a more comprehensive discussion.

Second, increases in educational resources, which flow down to impact individual teachers, clearly help to ease the political opposition to changing governance structures. This has the potentially dual benefit of changing structure and making teaching a relatively more attractive occupation, which is arguably quite important given the documented long-term slide in relative teacher salaries (Hanushek and Rivkin, 2007). Under this strategy teachers receive a risk premium (i.e. increased salary) for accepting greater employment insecurity in the form of, for instance, pay for performance or giving up tenure. This is essentially the strategy employed by the federal government in the Race to the Top and Teacher Incentive Fund competitions.

Third, policymakers may simply reach the point where they decide that teacher policies need to change to make K-12 education more productive, even in the absence of additional resources. Up until the last few years reform efforts could be described as "same operations with greater intensity" (Hanushek, 2009). The steady increase in real expenditures directed toward public schools (until the onset of the Great Recession)—mostly toward lowering pupil-teacher ratios—but these resources have not resulted in significant increases in student performance over time (Hanushek, 2003). Radical change to current policies to identify differences between teachers and act upon those differences, without an infusion of additional money is politically difficult because it challenges the status quo without any new infusion of resources, but it does seem that there are ways that resources directed to teachers could be spent more effectively. In particular, the evidence is pretty clear that investments in teacher quality (if we can figure out how to target this elusive trait) are far more cost effective than investments in more teachers (and lowered class size). And, moreover, in terms of the way teachers are currently compensated, eliminating the master's pay premium looks like low-hanging fruit for reform. In fact, given the research on the value of generic master's degrees, modifying the bump in salary that teachers receive for getting a master's degree might be a good litmus test for the seriousness of reform efforts.[49]

Ultimately the changes to teacher policies that are likely to yield great benefit come with a non-trivial amount of risk, both politically and because they will, by design, likely lead to teacher behavioral responses that are uncertain. Whether this risk is worthwhile is clearly in the eye of the beholder.

49 I am not suggesting abandoning pay increases for teachers, that would likely have a negative effect on the quality of the workforce, just that the master's pay premium might be used to increase teacher pay in different, more effective, ways. For a more comprehensive discussion of this issue, see Goldhaber (2010b).

RESOURCES

Aaronson, D., Barrow, L., and Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools." Journal of Labor Economics 25, no. 1 (2007): 95-135.

Anderson, J. (2013). Curious grade for teachers: Nearly all pass. New York Times. Published March 30, 2013.

Ballou, D. (2001). Pay for performance in public and private schools. Economics of Education Review, 20:1, pp. 51-61.

Bloom, H., Hill, C., Black, A., and Lipsey, M. (2008). Performance Trajectories and Performance Gaps as Achievement Effect-Size Benchmarks for Educational Interventions. New York: MDRC.

Boe, E. and Gilford, D. (1992). Teacher supply, demand, and quality: policy issues, models, and data bases. Proceedings of NCES conference, National Academies Press.

Boser, U. (2012). Race to the Top: what have we learned from the states so far?: A state-by-state evaluation of Race to the Top performance. Center for American Progress, March 26, 2012.

Boyd, D., Goldhaber, D., Lankford, H., and Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. Future of Children, 17(1): 45-68.

Bridges, E. M. & Gumport, P. (1984). The dismissal of tenured teachers for incompetence (Technical Report). Stanford, CA: Institute for Research on Educational Finance and Governance.

Chait, R. & Miller, R. (2010) Treating different teachers differently: How state policy should act on differences in teacher performance to improve teacher performance and equity. Report for the Center for American Progress.

Chetty, R. and Friedman, J.N. and Rockoff, J. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER.

Christensen, C., Johnson, C., and Horn, M. (2008). Disrupting class: How innovation will change the way the world learns. New York: McGraw Hill.

Clotfelter, C., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: a cross subject analysis with fixed effects"." Journal of Human Resources 45: 655-681.

Clotfelter, C., Glennie, E., Ladd, H., and Vigdor,J. (2008). Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina. Journal of Public Economics, 92:5, pp. 1352-1370.

Cohen, D. K. & Hill, H. C. (2000) Instructional policy and classroom performance: The mathematics reform in California. Teachers College Record, 102(2), pp. 294-343.

Cohen, E. and Varghese, P. (2012). "Teacher quality roadmap: Improving policies and practices in LAUSD." National Council on Teacher Quality.

Corcoran, S., Evans, W., and Schwab, R. (2004). Women, the labor market, and the declining relative quality of teachers." Journal of Policy Analysis and Management 23(3): 449-470.

Darling-Hammond, L., Berry, B., and Thoreson, A. (2001). Does teacher certification matter? Evaluating the evidence. Educational Evaluation and Policy Analysis, 23:1, pp. 57-77.

Darling-Hammond, L. and McCloskey, L. (2008). Assessment for learning around the world: What would it mean to be internationally competitive. Phi Delta Kappan 90:4, pp. 263-272.

RESOURCES

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. and Rothstein, J. (2012). Evaluating teacher evaluation."

Phi Delta Kappan, March, 2012.

Dee, T. S. & Wyckoff, J. (2013). Incentives, selection, and teacher performance: Evidence from IMPACT. NBER Working Paper No. 19529, October 2013, JEL No. I2,J45

Duncan, A. (2010). The new normal: Doing more with less. Remarks to the American Enterprise Institute, November 17, 2010.

Educational Research Service. (1988). Teacher evaluation: Practices and procedures. Arlington, VA.

Education Week. (2011). Teacher quality. Education Week July 8, 2011.

Eide, E., Goldhaber, D., and Brewer, D. (2004). The teacher labour market and teacher quality." Oxford Review of Economic Policy, 20: 230-244.

Ewing, J. (2011). Mathematical intimidation: Driven by the data. Notices of the American Mathematical Society, May 2011, pp. 667-673.

Fama, E. (1980). Agency problems and the theory of the firm. The Journal of Political Economy 88: 288-307.

Ferguson, R. (2011) "How students' views predict graduation outcomes and reveal instructional disparities under Children First reforms." Education Reform in New York City: Ambitious Change in the Nation's Most Complex School System. Ed. O'Day, Jennifer A., Catherine S. Bitter, and Louis M. Gomez. Harvard Education Press, 225-254.

Figlio, D.N. and Kenny, L.W. (2007). Individual teacher incentives and student performance. Journal of Public Economics, 91:5, pp. 901-914.

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.

Garet, M.S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H.S., Doolittle, F., Zhu, P., Sztehjnberg, L. (2008). The impact of two professional development interventions on early reading instruction and achievement." American Institutes for Research and MDRC. NCEE 2008-4030.

Gladwell, M. (2008). Outliers: the story of success. Little, Brown, and Company, New York, NY.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S. and Whitehurst, G. (2010). Evaluating teachers: The important role of value-added. Brookings Institution.

Glazerman, S. and Seifullah, A. (2010). An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report. Mathematica Policy Research.

Gneezy, U., Neier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. Journal of Economic Perspectives 25(4): 191-209.

Goldhaber, D. (2007) Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? Journal of Human Resources, 42(4), 765-794.

Goldhaber, D. (2008) Teachers matter, but effective teacher quality policies are elusive." In Ladd, H. and Fiske, E., ed., Handbook of Research in Education Finance and Policy, New York: Routledge, pp.146-165.

RESOURCES

Goldhaber, D. (2009). Politics of teacher pay reform." In Springer, M, ed., Performance Incentive: Their Growing Impact on American K-12 Education. Washington, DC: Brookings Institution Press.

Goldhaber, D. (2010a). When the stakes are high, can we rely on value added? Center for American Progress, December 2010.

Goldhaber, D. (2010b). Teacher pay reforms: The political implications of recent research." CEDR Working Paper 2010-4.0.

Goldhaber, D. (2011a). Licensure: exploring the value of this gateway to the teacher workforce." In Hanushek, E., Machin, S., and Woessman, L., ed., Handbook of the Economics of Education, Vol 3. Amsterdam: North Holland.

Goldhaber, D. (2011b). Two avenues for change. Education Week Blogs: The Futures of School Reform, April 25, 2011.

Goldhaber, D. and Brewer, D. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. Journal of Human Resources, pp. 505-523.

Goldhaber, D., Brewer, D. & Anderson, D. (1999) A three-way error components analysis of educational productivity. Education Economics, 7(3), 199-208.

Goldhaber, D. and Brewer, D. (2000a). Does teacher certification matter? High school teacher certification status and student achievement. Educational evaluation and policy analysis. 22:2, pp. 129-145.

Goldhaber, D. and Brewer, D. (2000b). Evaluating the evidence on teacher certification: A rejoinder. Educational Evaluation and Policy Analysis. 23:1, pp. 79-86.

Goldhaber, D., Player, D., DeArmond, M., and Choi, H. (2008). Why do so few public school districts use merit pay?" Journal of Education Finance, 33(3): 262-289.

Goldhaber, D. & Hansen, M. (2010) Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Seattle, WA, Center on Reinventing Public Education.

Goldhaber, D. & Theobald, R. (2013). Managing the teacher workforce in austere times: The determinants and implications of teacher layoffs. Education Finance and Policy, 8(4): 494-527.

Goldhaber, D. and Walch, J. (2012). "Strategic pay reform: a student outcomes-based evaluation of Denver's ProComp teacher pay initiative." Economics of Education Review, 31(6): 1067-1083.

Goldhaber, D., & Walch, J. (2014). Gains in teacher quality: Academic capabilities of the U.S. teaching force are on the rise. Education Next. 14(1).

Goldhaber, D. & Chaplin, D. (2012). Assessing the "Rothstein falsification test." Does it really show teacher value-added models are biased? CEDR Working Paper 2012-1.3. University of Washington, Seattle, WA.

Goldhaber, D., Liddle, S., Theobald, R., & Walch, J. (2012). Teacher effectiveness and the achievement of Washington's students in mathematics. WERA Educational Journal, 4(2): 6-12.

Goldhaber, D. & Hansen, M. (2013) Is it just a bad class? Assessing the stability of measured teacher performance. Economica, 80(319): 589-612.

Gordon, R., Kane, T., and Staiger, D. (2006). Identifying effective teachers using performance on the job," Hamilton Project White Paper 2006-01, April 2006.

RESOURCES

Greenwald, R. and Hedges, L.V. and Laine, R.D. (1996). The effect of school resources on student achievement. Review of educational research, 66:3, pp. 361-396.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J. H., Boyd, D. J. & Lankford, H. (2010) Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. CALDER working paper.

Hansen, M. (2009). How career concerns influence public workers' effort: Evidence from the teacher labor market. CALDER Working Paper #40

Hanushek, E. (1986) The economics of schooling - production and efficiency in public-schools. Journal of Economic Literature, 24(3), 1141-1177.

Hanushek, E. (1992) The trade-off between child quantity and quality. Journal of Political Economy, 100(1), 84-117.

Hanushek, E. (1997) Assessing the effects of school resources on student performance: An update. Educational Evaluation and Policy Analysis, 19(2), 141-164.

Hanushek, E. (2003). The importance of school quality." In Peterson, P., ed., Our Schools and Our Future: Are We Still at Risk?. Stanford, CA: Hoover Institution Press.

Hanushek, E. (2009). Teacher deselection. In Goldhaber, D. and Hannaway, J., ed. Creating a New Teaching Profession. The Urban Institute.

Hanushek, E. and Pace, R. (1995). Who chooses to teach (and why)?" Economics of Education Review 14(2): 101-117.

Hanushek, E., & Rivkin, S. (2004). How to improve the supply of high quality teachers. Brookings Institute.

Hanushek, E., & Rivkin, S. (2007). Pay, working conditions and teacher quality. Future of Children, 17(1), 69-86.

Hanushek, E. & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality." American Economic Review 100, no. 2 (2010): 267-71.

Harris, D. and Sass, T. (2009). What makes for a good teacher and who can tell? National Center for Analysis of Longitudinal Data in Education Research. Working Paper 30, 2009.

Harris, D., and Sass, T. (2011). Teacher training, teacher quality, and student achievement. Journal of Public Economics, 95, 7-8: 798-812.

Hassel, Emily A., and Hassel, Bryan (2009). "3X for All: Extending the Reach of Education's Best." Public Impact.

Hatry, H., Greiner, J., & Ashford, B. (1994). Issues and case studies in teacher incentive plans. Washington, DC: Urban Institute Press.

Hess, F. (2009). The human capital challenge: Toward a 21st-century teaching profession. In Goldhaber, D. and Hannaway, J., ed. Creating a New Teaching Profession. The Urban Institute.

Hill, H.C., Umland, K. L. & Kapitula, L.R. (2011). A validity argument approach to evaluating value-added scores. American Educational Research Journal 48, 794-831.

Hill, P. (2009). Consequences of instructional technology for human resource needs in education. In Goldhaber, D. and Hannaway, J., ed. Creating a New Teaching Profession. The Urban Institute.

RESOURCES

Holmström, Bengt. (1982). Managerial incentive problems: A dynamic perspective. In Essays in economics and management in honor of Lars Wahlbeck. Helsinki: Swedish School of Economics. Reprinted in Review of Economic Studies (1999) 66: 169-182.

Hoxby, C. and Leigh, A. (2005). Wage distortion: Why America's top female college graduates aren't teaching." Education Next 5(2): 51–56.

Ichniowski, C. and Shaw, K. (2003). Beyond incentive pay: Insiders' estimates of the value of complementary human resource management practices. The Journal of Economic Perspectives 17(1), pp. 155-180.

Imberman, S. and Lovenheim, M. (2013). Incentive strength and teacher productivity: evidence from a group-based teacher incentive pay system. NBER Working Paper No. 18439. March 2013.

Ingersoll, R. and Merrill, L. (2012). Seven trends: The transformation of the teaching force." Working Paper. Consortium for Policy Research in Education. May 2012.

Ingersoll, R. and Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers. Review of Educational Research 81:2, pp. 201-233.

Jacob, B. and Lefgren, L. (2004). The impact of teacher training on student achievement. Journal of Human Resources, 39:1, pp. 50-79.

Jacob, B., Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective evaluation in education. Journal of Labor Economics 26 (1), 101–136.

Jackson, C. and Bruegmann, E. (2009). Teaching students and teaching each other: the importance of peer learning for teachers." American Economic Journal: Applied Economics 1(4): 85–108.

Kane, T., Rockoff, J. & Staiger, D. (2007) Photo finish: Teacher certification doesn't guarantee a winner. Education Next, 7(1), 60-67.

Kane, T., Rockoff, J., Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review 27 (6), 615-631.

Kane, T. & Staiger, D. O. (2008) Estimating teacher impacts on student achievement: An experimental evaluation. Cambridge, MA, NBER.

Kane, T., Taylor, E., Tyler, J. and Wooten, A. (2010). Identifying effective classroom practices using student achievement data. NBER Working Paper.

Kennedy, M. (1998) Form and substance in teacher inservice education, Research Monograph No. 13 (Madison, WI, National Institute for Science Education, University of Wisconsin-Madison).

Kristoff, N. (2009). Our greatest national shame. The New York Times, published February 14, 2009.

Lavely, C., Berger, N., & Follman, J. (1992). Actual incidence of incompetent teachers. Educational Research Quarterly, 15(2), 11-14.

Lavy, V. (2007). Using performance-based pay to improve the quality of teachers. The future of children, pp. 87-109.

RESOURCES

Lazear, E. (2000). Performance pay and productivity." American Economic Review 90, no. 5: 1346-61.

Lazear, E. (2003). Teacher Incentives. Swedish Economic Policy Review, 10, 197-213.

Marsh, J., Springer, M., McCaffrey, D., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., and Peng, A. (2011). A big apple for educators: New York City's experiment with schoolwide performance bonuses. RAND Corporation.

Master, B., Loeb, S., Whitney, C. and Wyckoff, J. (2012). Different skills: Identifying differentially effective teachers of English Language Learners." CALDER Working Paper 68.

McCaffrey, D., Sass, T., Lockwood, J. & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. Education Finance and Policy, 4(4), 572-606.

McKinsey & Company. (2009, November). Shaping the future: How good education systems can become great in the decade ahead. Report on the International Education Roundtable.

Murname, R. & Cohen, D. (1986). Merit pay and the evaluation problem: Why most merit pay plans fail and few survive. Harvard Education Review, 56(1), 1-17.

Murnane, R., Singer, J., Willett, J., Kemple, J., and Olsen, R. (Eds.). (1991). Who will teach?: Policies that matter. Cambridge, MA: Harvard University Press.

Nye, B., Konstantopoulos, S. and Hedges, L.V. (2004). How large are teacher effects? Educational Evaluation and Policy Analysis 26:3, 237-257.

Player, D. (2009). Monetary returns to academic ability in the public teacher labor market. Economics of Education Review, 28:2, pp. 277-285.

Podgursky, M., & Springer, M. (2007). Teacher performance pay: A review. Journal of Policy Analysis and Management, 26, 909-949.

Podgursky, M.J. (2011). Teacher compensation and collective bargaining. Handbook of the Economics of Education, 3, pp. 279—313.

Reback, R. (2006). Entry costs and the supply of public school teachers." Education Finance and Policy 1 (2), 247-265.

Rhim, L., Kowal, J., Hassel, B., and Hassel, E. (2007). School turnarounds: a review of the cross-sector evidence on dramatical organizational improvement. Public Impact.

Rice, J. (2009). Investing in human capital through teacher professional development. In Goldhaber, D. and Hannaway, J., ed. Creating a New Teaching Profession. The Urban Institute.

Rice, J. (2010). The impact of teacher experience: Examining the evidence and implications for policy. Research brief for the Center for Longitudinal Data in Education Research. Washington DC: The Urban Institute.

Rivkin, S., Hanushek, E. & Kain, J. (2005) Teachers, schools, and academic achievement. Econometrica, 73(2), 417-458.

Rockoff, J. (2004) The impact of individual teachers on students' achievement: Evidence from panel data. American Economic Review, 94(2), 247-252.

RESOURCES

Rockoff, J., Jacob, B., Kane, T. & Staiger, D. (2011). Can you recognize an effective teacher when you recruit one? Education Finance and Policy, Winter 2011.

Rockoff, J. and Speroni, C. (2011). Subjective and objective evaluations of teacher effectiveness: evidence from New York City." Working Paper February, 2011.

Rothstein, J. (2010) Teacher quality in educational production: Tracking, decay, and student achievement. Quarterly Journal of Economics, 125(1), 175-214.

Rothstein, J. (2013). Teacher quality policy when supply matters. NBER Working Paper No. 18419. September 2012.

Sanders, W. and Rivers, J. (1996). Cumulative and residual effects of teachers on future student academic achievement. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Springer, M., Hamilton, L., McCaffrey, D., Ballou, D., Le, V., Pepper, M., Lockwood, JR, Stecher, B. (2010a). Teacher pay for performance: experimental evidence from the project on incentives in teaching. National Center on Performance Incentives.

Springer, M., Lewis, J., Podgursky, M., Ehlert, M., Taylor, L., Lopez, O., Ghosh-Dastidar, B., and Peng, A. (2010b). District awards for teacher excellence (DATE) program: year one evaluation report. National Center on Performance Incentives.

Strauss, V. (2012). Hawaii teachers reject contract in 'blow' to Race to the Top." The Washington Post. January 21, 2012.

Strunk, K. and Grissom, J. (2010). Do Strong Unions Shape District Policies? Educational Evaluation and Policy Analysis 32:3, pp. 389—406.

Toch, T. & Rothman, R. (2008) Rush to judgment: Teacher evaluation in public education. Report for Education Sector (Washington, D.C.).

Tucker, P. (1997). 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 11:103-126.

Wall Street Journal. (2010). Unions v. Race to the Top: states are waiting for Arne Duncan. The Wall Street Journal January 7, 2010.

Weisberg, D., Sexton, S., Mulhern, J. & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York, NY: The New Teacher Project.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. (2007). Reviewing the evidence on how teacher professional development affects student achievement (Issues & Answers Report, REL 2007–No. 033). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southwest. Retrieved from http://ies.ed.gov/ncee/edlabs

The Bush Institute at the George W. Bush Presidential Center
2943 SMU Boulevard | Dallas, Texas 75205 | 214.200.4300
www.bushcenter.org | educationreform@bushcenter.org

THE BUSH INSTITUTE
AT THE
GEORGE W. BUSH
PRESIDENTIAL CENTER