Teacher Effects and Teacher-Related Policies

C. Kirabo Jackson
Northwestern University
kirabo-jackson@northwestern.edu

Jonah E. Rockoff
Columbia Business School
jonah.rockoff@columbia.edu

Douglas O. Staiger
Dartmouth College
douglas.o.staiger@dartmouth.edu

ABSTRACT

The emergence of large longitudinal datasets linking students to teachers has led to an explosion in the study of teacher effects on student outcomes by economists over the last decade. One large literature has documented wide variation in teacher effectiveness that is not well explained by observable student or teacher characteristics. A second literature has investigated how educational outcomes might be improved by leveraging teacher effectiveness, through processes of recruitment, assignment, compensation, evaluation, promotion, and retention. These two lines of inquiry are closely tied; the first tells us about the importance of individual teachers, the latter about how this information can be used in policy and practice. We review the most recent findings in economics on the importance of teachers and on teacher-related policies aimed at improving educational production.

At least since the release of the Coleman Report in 1966, economists have been interested in identifying and quantifying the determinants of educational production. Early work in the economics of education focused on estimating the returns to per-pupil expenditures, class size, and average teacher salaries. However, many recent studies focus on the impact of the individual teacher providing instruction.

The study of teacher effects on student outcomes by economists began with Hanushek (1971) and Murnane (1975). Since then, better data and modern empirical methods have led to a renaissance of this literature. In particular, the emergence of large longitudinal datasets linking students to teachers has allowed researchers to disentangle the contribution of teachers from other factors at the student, classroom, and school levels that also influence student outcomes.[1] The central conclusion of this research is that teachers are not interchangeable production inputs. There is wide variation in teacher effectiveness that is not well explained by observable teacher characteristics such as years of education and experience.

Taking the wide variation in teacher effectiveness as given, a second literature has emerged investigating how educational outcomes might be improved by leveraging teacher effectiveness, through processes of recruitment, assignment, compensation, evaluation, promotion, and retention. These two lines of inquiry are closely tied; the first tells us about the importance of individual teachers, the latter about how this information can be used in policy and practice. We

[1] Such databases became available through the Texas Schools Project, the North Carolina Education Research Data Center, and Florida's K20 Education Data Warehouse, as well as through districts such as Los Angeles USD, San Diego USD, Chicago Public Schools, and New York City.

review the most recent findings in economics on the importance of teachers and on teacher-related policies aimed at improving educational production.[2]

What Are Teacher Effects and How Are They Measured?

A teacher effect, or, as it is often called, a teacher's "value-added," is not a measure of inputs into the production of education. It is a label given to systematic variation in output across students assigned to the same teacher. Broadly speaking, we conceive of teacher quality as the ability to increase students' stock of human capital, however that may be achieved –better communication to students, classroom management, encouragement of greater effort by students or parents, etc. The first-order empirical challenge is that students' human capital is not directly measurable, so economists rely on measures of academic achievement such as standardized tests. Test scores are by far the most common measure used because they are typically the only objective measure of educational output available for large samples of students, teachers, and schools, and because many studies have linked test scores to adult labor market earnings (Murnane, Willet and Levy, 1995; Neal and Johnson, 1996). After discussing how teacher effects are measured, we discuss recent findings on the foundational question of whether a teacher's effect on test scores is a good measure of teacher quality and whether teacher effects on other outcomes might improve our ability to identify high quality teachers.

[2] Readers familiar with this literature might ask what they might find here that has not been covered in other review articles, including our own previous writing (e.g., Hanushek & Rivkin 2010; Staiger & Rockoff 2010; Jackson 2012). While some basic elements of this literature have not changed, important studies have emerged in the past two years which have added greatly to our understanding of teacher effects, particularly on questions of bias, teacher effects on other student outcomes, other measures of teacher job performance, and teacher incentive programs in US public schools.

Being unobservable, teacher value-added is estimated, typically using least squares regression, where identifiers for students' teachers are linked with positive or negative effects on student achievement.[3] An obvious hurdle is that teachers are not randomly assigned to schools, or even to students within schools, so differences in outcomes across teachers could reflect differences in other determinants of student achievement, rather than the contributions of teachers. Accounting for these other factors is the major goal in the estimation of teacher effects. Whether researchers have achieved this goal has been a matter of debate, which we discuss at greater length below.

While there are many variations in the exact set of covariates and choice of regression specification in studies of teacher effects, their essential ingredients are quite similar. To give a basic sense of the methodology, Equation 1 characterizes a general value-added specification, where achievement (A_{it}) of student i in year t is a function of various observable (X_{it}) and unobservable factors (ν_{it}):

$$(1)\quad A_{it} = \beta X_{it} + \nu_{it} \text{ , where } \nu_{it} = \mu_{jt} + \theta_c + \varepsilon_{it}$$

The most important observable factor invariably accounted for in estimation of teacher effects is a student's achievement outside of the teacher's classroom. In the vast majority of studies this is done using achievement in the previous year, though some studies have done so via the inclusion of a student fixed effect (e.g., Clotfelter et al. 2007). Additional controls for other characteristics

[3] Because we never observe a classroom without a teacher, estimates of teacher effects are measures of relative productivity. A negative teacher effect implies that the teacher's students are expected to achieve less than they would if assigned another teacher at random, not that they achieve less than if left to learn without a teacher at all.

of the student and his/her peers that affect achievement growth (e.g., household resources, learning disability, English proficiency, etc.) are also common.[4]

The residual variation in student achievement is attributed to three factors: a teacher effect (μ_{jt}), an idiosyncratic classroom-level effect (θ_c), and an idiosyncratic student-level effect (ε_{it}). All of these effects are unobservable, but only the teacher effect, by definition, persists across classrooms of different students taught by the same teacher. Thus it can be identified by observing a teacher with multiple classrooms of students.[5] The classroom-level effect encompasses a wide variety of factors that influence the achievement level of the entire class, including factors related to the teacher that do not persist over time (e.g., idiosyncratic match quality between the teacher and classroom or school), as well as correlated measurement error (e.g., a dark barking outside during the test). The student-level effect encompasses myriad factors that can influence a child's year-to-year academic growth, as well as measurement error, which is non-trivial on most standardized tests.

We observe teachers with a finite number of classrooms and students, and estimates of teacher effects will therefore contain estimation error. Researchers typically use an empirical Bayes approach to incorporate this into teacher effect estimates (see Kane & Staiger 2008; Chetty et al. 2013a). Given the prior belief that teacher effects are centered on some grand mean with some

[4] Most researchers (but few states or districts) control for peer variables such as percent free lunch or average baseline test score in the classroom or school. Kane et al. (2013) found that the component of the teacher effect that was correlated with peer variables predicted differences in student achievement when teachers were randomized to classrooms, suggesting that one should not control for peer variables. But since this study randomized teachers within school, it does not answer whether one should control for peer differences between schools.

[5] The earliest studies by Hanushek and Murnane, as well as work by Nye et al. (2004), did not distinguish between classroom and teacher effects. This limitation is essentially what separates these early studies from more recent work, beginning with Rockoff (2004), Rivkin et al. (2005), and Aaronson et al. (2007). In nearly all studies in this literature, teacher effects are assumed to be fully persistent or fixed across classrooms. We discuss recent papers that relax this assumption below.

true variance, an empirical Bayes estimate for the teacher effect can be generated by shrinking the initial teacher effect estimate toward this mean (typically zero). This is done based on the signal-to-noise ratio of the initial estimate, with greater "shrinkage" for teachers for whom less data are available.[6] In typical classroom sizes of 20 to 30 students, the true teacher effect will comprise roughly one third of the residual variance at the classroom level, with the remaining residual variance attributable to transitory effects.

Teacher effects, based as they are on residual variance in outcomes, are measured in standard deviations of student achievement. Figure 1 replicates the distributions of value-added estimated by Kane et al. (2008) for teachers in New York City. Estimates of a standard deviation in teacher value added (i.e., the difference between an average teacher and one at the 84th percentile) typically range from 0.1 to 0.2 standard deviations of student achievement. Compared to the effects of other interventions into educational production, the magnitude of teacher effect estimates is relatively large, and one of the main reasons for increased interest in this topic by economists.

A natural question to ask is whether teacher effects are more important in some parts of the educational process than others. Studies have generally found greater variance in teacher effects on achievement in math than in English (or reading) achievement; Figure 1, replicated from Kane et al., displays this feature.[7] There is no clear explanation for this fact. Nevertheless, our

[6] The true variance of teacher effects is estimated using the covariance of combined teacher-class-student effects across classrooms with the same teacher. The student-level effect variance is estimated based on within-classroom mean squared error from the estimation regression, and remaining variance is assigned to classroom-level effects.
[7] Kane et al. (2012) do find much greater variance in teacher effects on an alternative English test that included open-ended writing questions, suggesting that some of the difference may reflect the particular content on state tests, rather than a general difference between math and English.

prior (which we believe is shared by many researchers in this field) is that the difference is due to the fraction of learning taking place in school; while mathematics is almost exclusively learned in the classroom, English and reading are learned to a great extent outside of school.[8] Due to the paucity of testing in areas other than Math and English, there is little evidence on teacher effects in other subjects (such as science or social studies).

Another natural question to ask is whether teacher effects are more important at different grade levels. Nearly all studies of teacher effects have focused on the upper elementary and middle school grades, where annual testing data are most commonly available. Within this grade range, there is little evidence of teachers becoming more or less influential on achievement as students age. The evidence for high school teachers is mixed. Earlier work by Aaronson et al. (2007), Clotfelter et al. (2010), Goldhaber et al. (2013) and Burgess et al. (2011) find high school teacher effects that are similar in size to those from elementary school settings. However, Jackson (forthcoming) accounts for tracking biases specific to the high school context and finds that the variance of persistent teacher effects on test scores may be smaller than previously thought.

Can Teacher Characteristics Explain Teacher Effects?

While research is fairly conclusive that teacher effects account for significant variation in achievement across classrooms, commonly observed teacher characteristics such as level of education, degree granting institution explain little of this variation (see Gordon, Kane and Staiger, 2006, for a review). For example, Figure 1 shows that the distributions of value-added

[8] Consistent with this notion, the relative magnitude of math and English effects has also been found for the impact of highly effective charter schools on student achievement (e.g., Tuttle et al. (2013)).

calculated in Kane et al. (2008) almost completely overlap when teachers are separated into four groups based on their type of teaching certification.[9] However, early studies document benefits to additional years of experience for early career teachers (e.g. Rockoff 2004; Rivkin et al. 2005), and recent studies indicate that teacher effectiveness meaningfully improves well into a teacher's career (Wiswall 2013; Papay & Kraft 2013).[10] We return to this evidence below in our discussion of policies to improve teacher effectiveness, such as mentoring and feedback.

Two recent studies depart from the use of administrative data in search for a stronger link between teacher effects and teacher characteristics. Rockoff et al. (2011) conduct an online survey of new elementary and middle school math teachers in New York City to collect data on characteristics linked to performance in similar occupations, such as general intelligence, personality traits, and beliefs regarding self-efficacy. They find that no single characteristic is a strong predictor of student achievement growth. However, they can predict roughly 10 percent of the variation in new teacher performance using indices that combine these characteristics.

Dobbie (2011) takes a similar approach, using data from Teach for America (TFA), where new teaching candidates are rated by TFA admissions staff on eight criteria used to make program selection decisions. Scores on each criterion (academic achievement, leadership experience,

[9] In their study, traditionally certified teachers obtained certification through a university teacher education program, either at the bachelors or masters level. Teach for America and Teaching Fellows are both alternative certification programs, where individuals without certification are permitted to teach while they take certification courses, and Uncertified teachers simply lack the proper certification but were hired by New York City schools, in violation of state regulations. The use of uncertified teachers was common in New York City up until roughly 2005.

[10] Wiswall (2013) in particular finds that after accounting for the quality of teachers who remain in the profession, the return to later career teacher experience is sizable. He estimates that a teacher with 30 years of experience has over 1 standard deviation higher measured quality than a new, inexperienced teacher and about 0.75 standard deviations higher measured mathematics effectiveness than a teacher with 5 years of experience. In comparison, estimates on the same data sample using previously restricted models suggest that experienced teachers have between 0.1 and 0.2 standard deviations higher quality than new teachers, with almost all of these gains in the first few years of teaching.

perseverance, critical thinking, organizational ability, motivational ability, respect for others, and commitment to the TFA mission) are based on information collected from an on-line application, a phone interview, and an in-person interview. In line with Rockoff et al.'s previous findings, Dobbie finds that few of these TFA criteria are individually significant predictors of student achievement growth, but they strongly predict new TFA teacher performance when averaged into an index. His results suggest that more than half of the variance in value-added of TFA teachers could be predicted based on their admissions scores.[11]

Thus, it appears that some headway may be made in identifying the type of individuals who are likely to succeed in teaching through more intensive and purposeful data collection during the hiring process. Still, it appears that even with much better data, observable characteristics are unlikely to be able to predict most of the variation in teacher effects.

Are Teacher Effect Estimates Accurate?

Are teacher effects an accurate measure of at least some dimensions of teacher quality? This broad question has been the subject of much recent work by economists. For expositional purposes, we discuss separately research on the issues of bias and precision.

The major identification question is whether estimates of teacher effects based on observable data can separate teacher effects from other factors driving variation in student achievement. This question—are teacher effects biased?—has been the subject of a number of studies, but its subtlety is often not fully understood. First, one must distinguish between the accuracy of

[11] Based on authors' calculations and personal correspondence with Will Dobbie.

teacher effects in measuring how well a teacher performed in the past, or how well a teacher is expected to perform in the future. The literature focuses much of its energy on the latter goal. This is due largely to its relevance for policies such as teacher retention.

Second, one can distinguish a lack of bias in the predictive content of teacher effect estimates over a population from the absence of bias in every individual teacher effect estimate. So far, researchers have not adequately addressed the general issue of individual biases; doing so would require a number of strong assumptions in addition to multiple years of random assignment of students to teachers (Chetty et al. 2013a). As we explain below, the question of predictive bias over a population has seen far greater progress.

In two influential papers, Rothstein (2009; 2010) examines bias in several common value-added regression specifications using data from North Carolina. Rothstein finds that students' future teacher assignments predict current achievement, conditional on the value-added controls, indicating that students are sorted to teachers based on characteristics unobservable to researchers. However, several subsequent analyses suggest that Rothstein's test may be driven by correlated measurement error and mean reversion in test scores (Koedel & Betts 2011; Chetty et al. 2013a). The intuition for this is as follows: if a classroom of students has unusually high test score growth this year due to some non-persistent, correlated shock, then they will experience mean reversion next year and, if there is enough persistence in grouping, their future teacher will significantly predict this year's growth. Additional papers present simulation evidence that Rothstein's test performs poorly in small samples (Goldhaber & Chaplin 2012; Kinsler 2012).

More recently, there have been three important studies which examine bias in teacher effects by testing whether they predict student achievement growth under arguably exogenous teacher assignment rules. Kane & Staiger (2008) and Kane et al. (2013) evaluate field experiments where students were randomly assigned between pairs of teachers in the same school and grade level. They find that teacher effects estimated using historical data (under non-random assignment) were unbiased predictors of the within-pair differences in student achievement under random assignment.

In addition, Chetty et al. (2013a) implement a quasi-experimental test for bias, using changes in the mean teacher effect at the cohort level and changes in achievement across cohorts. Intuitively, if a 4th grade teacher with low (estimated) value-added leaves a school and is replaced by a teacher with high (estimated) value-added, we would expect that average 4th grade achievement would rise. Moreover, if the teacher effect estimates are unbiased, cohort-level scores should rise by the difference in the two teachers' effects, multiplied by the share of students they taught.[12] In contrast, if the two teachers were actually equally effective (i.e., their estimated effects were driven purely by sorting), then there should be no cross-cohort change in scores.

Using data from a large urban school district, they find no evidence for this bias: the quasi-experimental cross-cohort variation in teacher effects is an unbiased predictor of cross-cohort changes in achievement. Figure 2 illustrates this finding, reproduced from Chetty et al. (2013a).

[12] For example, if there are four 4th grade teachers, each with one quarter of the students, and a teacher with value-added of -0.1 is replaced by a teacher with value-added of 0.1, then the improvement in scores across cohorts should be 0.2/4 = 0.05 standard deviations.

Panel A is not the quasi-experiment but is presented for purposes of comparison. It shows student achievement residuals in a teacher's current classroom plotted against the teacher's value-added estimate based on different students in other years. There is a very clean one-to-one relationship, as expected based on the construction of value added. As we have discussed, the persistence of VA over time could be driven by bias as well as through causal impacts of individual teachers. The quasi-experimental test is shown in Panel B, which plots *cross-cohort changes* in student achievement against *cross-cohort changes* in teacher value-added. The one-to-one relationship holds up strikingly well; test score changes and value-added changes are tightly linked, even when looking only across adjacent cohorts of students within the same school and grade. In addition to a number of other checks provided by Chetty et al. (2013a), the relationship shown in Panel B of Figure 2 strongly supports the notion that there is minimal predictive bias in their value added measures.

Of course, there can be no guarantee that teacher effects estimated in other samples will be similarly unbiased. Since teachers are not randomly assigned, the properties of teacher effect estimates will depend on the quality of control variables that account for differences across students. However, it appears that the data and methods most commonly applied in this field are able to establish a causal link between teachers and student achievement.

While the evidence on bias is supportive of teacher effects, stability may be even more important in determining accuracy in predicting teachers' future effects. Sizeable student- and class-level error components mean that a teacher effect based on just one or two classrooms can be a noisy indicator of a teacher's future performance, even if it does contain real and potentially useful

information (see Staiger and Rockoff, 2010). The year-to-year correlation of teacher effect estimates has been found to range from 0.2 to 0.7, similar to objective performance measures in other jobs such as professional sports, insurance and security sales, and manual piece-rate production (McCaffrey et al., 2009). Nevertheless, given that many teacher labor contracts involve an "up or out" tenure decision fairly early in a teacher's career, usually after just 2-4 years, the apparent instability of teacher effect estimates raise concerns about using this information in personnel-related policy.

However, Staiger and Kane (2013) have argued that year-to-year stability in annual performance is a misleading statistic. The impact of a retention decision, for instance, rests on the correlation between a single year's performance (or performance to date) and a teacher's remaining *career* performance. It is straightforward to show that the year-to-career correlation is just the square root of the year-to-year correlation, so that a year-to-year correlation of 0.36 corresponds to a year-to-career correlation of 0.6. Using data from several urban school districts which have 6 or more years of data on teacher's value added, they estimate year-to-career correlations in the range of .65-.8 for math, and .55-.7 for English. These imply, for example, that over three quarters of teachers at the 25th percentile of one-year value added have career value added that is below average.

There are several ways in which noise in value-added measures might be reduced, most obviously by using multiple years of data, or generally multiple classrooms, to construct these measures. Lefgren and Sims (2012) combine teacher effects across subjects and find that the optimal weighted average of math and English value added for elementary teachers in North

Carolina substantially improved the ability of these measures to predict future teacher value added in each individual subject. Alternatively, teacher effects could be combined with other sources of information. Mihaly et al. (2013) find estimates of teachers' effectiveness are more stable when they incorporate classroom observations and student surveys, but that these measures did not substantially improve the ability to predict teacher effects on test scores over what was possible using value added estimates alone.[13]

Are teacher effects stable across time and context?

While most of the literature has assumed teacher effects to be fully persistent and fixed, recent evidence suggest that true teacher effects change over time and across different contexts such as the school, grade, and subject being taught. Chetty et al. (2013a) and Goldhaber and Hansen (2013) estimate teacher effects with imperfect persistence and find that roughly half of the short-run persistence in teacher effects across classrooms in adjacent years is present among classrooms seven or more years apart. Similarly, Jackson (forthcoming) estimates teacher effects allowing for a school-specific match component. He finds that roughly half of the persistence in teacher effects observed across classrooms taught within the same school is present among classrooms taught in different schools. Evidence also suggests that there are subject and grade-specific match components to teacher effects, with less persistence in teacher effects across classrooms taught in different grades (Kane and Staiger, 2005) and across different subjects (Lefgren and Sims, 2012; Condie, Lefgren, and Sims, 2012).[14]

[13] The in-class observation indicator they use is a teacher's scores on the Framework for Teaching evaluation rubric (Danielson, 1996), averaged across four lessons. The student survey indicator was the previous year's class average response to questions on the Tripod Student Perception Survey (Ferguson 2009).

[14] In contrast to the differences observed across school, grade and subject, teacher effects appear to be fairly similar across student demographics and ability (Koedel and Betts, 2007; Lockwood and McCaffrey, 2009).

Changes in teacher effects across time and context have a number of important implications. First, estimates of teacher effects taken from a particular year or context will overstate the impact of that teacher in a different year or context – unless one uses estimation methods that allow for these changes as suggested by Chetty et al. (2013a) and Lefgren and Sims (2012). Moreover, using such methods can yield improvement in the accuracy of teacher effect estimates, as discussed earlier. Finally, as Jackson (forthcoming) and Condie, Lefgren and Sims (2012) suggest, the context-specific match component of teacher effects can be used to improve student performance through better matching of teachers to contexts (schools, grades, subjects) in which they are most effective.

Recent studies have found that the impact of being assigned a more effective teacher declines by half or more between end-of-year test scores and test scores two years later (McCaffrey et al., 2004; Jacob, Lefgren, and Sims, 2010; Kane and Staiger, 2008; Rothstein, 2010; Chetty et al., 2013a). This finding has been cited as a drawback of using value-added models to assess teachers: teacher impacts on end-of-year scores may overstate their long-term impact if students are forgetting what they have learned, or if value-added measured something transitory (like teaching to the test).

However, it is not clear what should be made of such "fade-out" effects. Fade-out could reflect changing content of the tests in later grades (students do not forget the content that they learned in prior years, it is no longer tested). Cascio and Staiger (2012) find that roughly 20% of fade-out may be due to non-comparability of test scales across grades. Alternatively, the impact of a good

teacher could spill over to other students in future years through peer effects, making relative differences in test scores appear to shrink – when mixed together in classrooms in subsequent grades, students who had learned more in prior grades must wait while other students catch up. These types of mechanisms could imply that short term value added measures are indeed accurate indicators of teacher effectiveness, despite apparent fade out. Better understanding of the mechanism generating fade out is critically needed before concluding that teacher effects on student achievement are ephemeral.

Teachers Effects and Outcomes Beyond Test Scores

As mentioned above, economists are not interested in teacher effects on test scores *per se,* and improvements in test scores may not necessarily indicate greater human capital or better long-run outcomes. The assumption that test scores reflect the stock of students' human capital is often assumed, but has only very recently been tested empirically.

Using administrative data on individual student outcomes linked to data from United States tax records, Chetty et al. (2013b) find that students in grades 4 to 8 assigned to high value-added teachers in primary school are more likely to attend college, earn higher salaries, live in higher SES neighborhoods, have higher savings rates, and (females) are also less likely to have children born when they are teenagers. Their results strongly support the idea that teacher effects on test scores have real economic content and capture, at least partially, a teacher's ability to raise students' human capital. Similarly, looking at high-school teachers in North Carolina, Jackson

(2013) finds that teacher value added in 9th grade Algebra and English predict effects on dropout, high school completion, SAT-taking, and college plans three years later.

Teacher effects on test scores appear to be an important measure of teacher quality. However, it is still quite possible that teacher effects on outcomes other than test scores can help better predict teacher effects on human capital and, therefore, longer run outcomes. Recent studies suggest that is likely the case.

A series of papers based on the Gates Foundation's Measures of Effective Teaching (MET) Project (Kane and Staiger 2012, Kane et al 2013) explore how various indicators of teacher performance might be combined to build a summative measure which more accurately depicts teachers' impacts on student achievement.[15] The project began by collecting a wide array of data on teachers from a large number of districts, including student surveys, evaluations of teaching practice based on in-class observation, and teacher effect estimates based on student achievement data. Using these data, Mihaly et al. (2013) found that there is a common component of effective teaching shared by all indicators, but there are also substantial differences across measurement modes (value added, classroom observation, and student perceptions) and even across value added based on different tests. Placing heavy weight on value added yields a composite measure that is strongly correlated with teacher effects on the particular test, but is only weakly correlated with teacher effects on other indicators. However, an equally weighted average of all indicators is only somewhat less correlated with teacher effects on the particular test, but much more highly correlated with teacher effects on other indicators (including teacher effects on alternative, low-

[15] These papers are by no means for the first to address this question, but they are by far the most extensive and have has a large impact on research and policy in this area. Other recent work by economists on this topic includes Rockoff & Speroni (2010), Rockoff et al. (2012) and Lavy (2011).

stakes, tests). They argue that because we care about human capital and not about test scores *per se*, a better measure of teacher quality might be a composite measure that summarizes a teacher's effect across a variety of different indicators of effective teaching that reflect different aspects of human capital.

Jackson (2013) supports the idea that teacher effects on non-test score outcomes might convey important additional information about teacher effects on human capital. Using data from 9th grade students in North Carolina, he finds that Algebra and English teacher in 9th grade have causal effects on both test scores and socio-behavioral outcomes such as absences, suspensions, grades, and on-time grade progression. Linking students in 9th grade to their outcomes up to four years later, Jackson finds that teacher effects on a weighted average of these non-test score outcomes (a proxy for students' non-cognitive skills) predicts effects on dropout, SAT-taking, and college plans—above and beyond teachers' effects on test scores. Calculations suggest that while test score value-added does predict meaningful effects on longer run outcomes, teacher effects on non-test score outcomes can increase our ability to identify those teachers who improve human capital substantially.

Teacher-Related Policies Aimed at Improving Educational Production.

Until recently, measures of teacher effectiveness such as those discussed above have played little role in teacher retention, evaluation, and pay decisions (Weisberg et al., 2009). However, as the evidence of wide variation in teacher effectiveness has grown more persuasive, many states and districts across the country have implemented teacher evaluation policies that incorporate value

added estimates, structured classroom observations, student perception surveys, and other methods to evaluate teachers. In the remainder of this section we review recent findings in economics on using measures of teacher effectiveness for teacher selection, mentoring and feedback, and pay for performance.

1. *Teacher Selection*

As we have discussed previously, differences in teacher effectiveness are large and persist over time. While these differences are difficult to predict at hire based on teacher credentials, they can be predicted after observing a teacher's performance in the classroom. Accordingly, the evidence suggests that using measures of teacher effects for tenure or layoff decisions could improve the average effectiveness of the teacher workforce, as compared to the current practice of granting tenure as a matter of course to nearly all teachers and determining layoffs primarily based on seniority.

The potential for teacher selection has been illustrated in a variety of ways in a number of recent papers. Hanushek (2011) simulates the impact that removing the lowest performing teachers would have on student test scores and earnings, making a range of plausible assumptions about the true variation across teachers, the amount of fadeout, class size, and the relationship between achievement and earnings. He finds that replacing the bottom 5-10% of current teachers with teachers who had average effectiveness would raise average test scores across all students by roughly .04 student standard deviations per year of education (equivalent to half of a student standard deviation over the course of their k-12 education). This increase in average teacher

effectiveness is worth approximately $10k-$20k annually per classroom in net present value of earnings generated over the students' lifetime. Gordon, Kane and Staiger (2006) and Goldhaber and Hansen (2013) perform similar simulations using somewhat different assumptions and come to similar conclusions. Chetty et al. (2013b) perform what is probably the most realistic simulation, using causal estimates of teacher value added on subsequent earnings, and also accounting for the imperfect relationship between the current value-added estimates used to identify low-performing teachers and subsequent teacher performance. They find similar results: replacing the bottom 5% of teachers with an average teacher would result in an increase in average teacher effectiveness across all teachers worth approximately $9k annually per classroom.

Similar results have been found in two studies looking at layoffs in New York City (Boyd et al., 2011) and Washington State (Goldhaber and Theobald, 2013). Both studies found that actual layoffs were largely uncorrelated with teacher value added and determined primarily by seniority. In simulations, using value added to determine layoffs would reduce the number of layoffs required to achieve budget targets (because high-seniority teachers with high salaries were more likely to be targeted for layoff) and would have laid off teachers with lower subsequent value added (by roughly 0.2 student standard deviations). In New York City, where 5% of the workforce would have been laid off, using value added to target teachers for layoff would have increased test scores by 0.01 student-level standard deviations (5% of 0.2).

Staiger and Rockoff (2010) and Neal (2012) take these simulations one step further, deriving optimal rules for teacher dismissal when a school district wishes to maximize teacher value

added. These papers quantify the tradeoff between the potential benefits of dismissing low value added teachers that are highlighted in the preceding simulations, against an important cost of dismissing experienced teachers: they will be replaced by rookie teachers who typically have lower than average value added in their first few years of teaching. Using somewhat different modeling strategies, both papers find that it is optimal to dismiss over half of teachers (those with the lowest value added) after their first year of teaching, and then dismiss a smaller fraction over the next few years. Such an aggressive policy would hypothetically result in increased test scores on the order of 0.1 student standard deviations per year of education. The reason for this stark result is straightforward: even one year of noisy value added data identifies differences in teacher effects that are large and persistent relative to the short-lived costs of hiring an inexperienced teacher. Rothstein (2012) extends these analyses to account for increased teacher pay that is likely to be necessary to attract additional teachers and compensate for the increased dismissal risk. He finds that an optimal firing policy would set dismissal rates below 50% and require a roughly 5% increase in pay, but would still yield a 0.04 increase average teacher value added, which would still be cost effective. For example, based on Rothstein's results, Chetty et al. (2013b) estimate that the increase in students' lifetime earnings from dismissing the bottom 5% of teachers is roughly 10 times the additional salary costs necessary to attract teachers.

All of the proceeding analysis is based on simulation of hypothetical policies that have not yet been tested in the field (although many districts are now in the process of implementing such policies). Rockoff et al. (2012) evaluate a field experiment in New York City where a random subset of principals received reports containing teacher effect estimates – a more laissez fare policy that provided teacher effect estimates to school administrators and allowed them to use the

information as they wished. Principals were surveyed regarding teacher performance at the start and end of the school year, pre- and post-receipt of the reports. Beliefs about teacher effectiveness changed for principals receiving teacher effect estimates, with these principals placing a weight on the estimates which was roughly one-fifth of the weight they placed on their own prior beliefs. In addition, larger weights were placed on teacher effect estimates that were more precisely estimated or when principals had known the teacher for a shorter period of time (and presumably had weaker prior beliefs). Importantly, in schools where principals received information on teacher effects, turnover increased for teachers with low performance estimates and there were small test score improvements for students. Thus, even if left completely to their own devices, providing information on teacher effects to principals appears to have an impact.

An important caveat for all of these results is that they are based on the properties of value added estimates from a relatively low-stakes environment, in which teachers were not being evaluated based on value added. Using value added to evaluate teachers may induce behavior that distorts these measures, making them less useful predictors of teacher effectiveness. As Rothstein (2012) has emphasized, the resulting misalignment between the measures and the policy goals could eliminate much of the expected benefit from dismissing teachers. Thus, it is critically important to evaluate the actual benefits of such programs as they are implemented at scale.

2. *Mentoring and Feedback*

As mentioned previously, teacher effectiveness tends to improve as they gain experience. This suggests that teaching ability is malleable and entails behaviors and skills that are learned over

time. Accordingly, creating schooling environments that promote teacher learning and provide high quality professional development for teachers during their careers may be fruitful. This approach differs from (de)selecting teachers based on effectiveness, but rather improves teacher quality by improving the skills of the existing stock of teachers. While there is consensus that teachers acquire skills over time, there is little consensus on how best to facilitate such learning. Teachers acquire skills informally through learning-by-doing and on-the-job learning in addition to formal professional development. We discuss the empirical literatures on both kinds of learning and highlight implications for policy.

a. Informal Learning

It seems likely that much of the return to teaching experience is due to "learning-by-doing," whereby teachers learn classroom management skills, oratory skills, and presentation skills in the process of interacting with students and colleagues, and providing instruction.

In addition to general skills, teachers may also learn specific content-related skills, such as how to sequence the topics for a particular math class, or anticipating the topics that are likely to cause confusion for students. Ost (forthcoming) investigates the rate at which teachers acquire general versus course-specific skills. To do so, he distinguishes between total years of teaching experience (which should include both general and specific skills) and years of experience in a particular grade and subject (which will include course and grade specific skills). He finds positive effects of both kinds of experience such that teachers become more effective more rapidly when they teach similar curricula year after year. As measured by students' math score improvements, grade-specific experience is between one-third to one-half as important as general

teaching experience. Using a similar design, Henry et al. (2012) find that high school math and science teachers gain course specific returns to experience, with the largest gains in advanced science courses such as Chemistry, and Physics.

In addition to learning-by-doing, interactions with colleagues is likely an important source of information about teaching (Spillane et al. 2012) so that informal peer interactions may be important for how teachers acquire skills. Using other teachers at the same school in the same grade as a measure of peers, Jackson & Bruegmann (2009) study the effect of working around more effective colleagues and find that exposure to better teacher peers improves own performance.[16] After two years, a permanent 1 standard deviation increase in mean teacher-peer quality is associated with about 0.075 standard deviation increase in both students' math and reading test scores. They conclude that about 20 percent of teacher effectiveness at a given point in time can be explained by the effectiveness of a teacher's peers over the previous three years, suggesting a significant role for informal peer learning.

Papay and Kraft (2013) find that the rate at which teacher effectiveness improved varied systematically across schools. Teachers improved most rapidly at schools that teachers reported to promote peer collaboration, provided professional development, and used teacher evaluations.

Taken together, the results indicate that teachers may improve most rapidly by structuring schools so that teachers (a) have similar assignments from year to year so that they may gain

[16] This modeling assumption is justified by finding from Daly et al. (2010) and Bakkenes et al. (1999) who use network analysis to find that teachers are most likely to interact with teachers in the same grade level.

mastery of a particular curriculum, (b) have opportunities to interact with high quality colleagues, and (c) provide opportunities for professional development.

b. Formal Training:

In the United States, the vast majority of teachers engage in training related to their main teaching assignment during the previous year (Parsad et al., 2001). Unfortunately, most of the existing research on this type of formal, in-service, professional development is based on samples where teachers and/or schools are self-selected into training, so that it is unclear whether one can credibly compare the outcomes of teachers who undergo training to the outcomes of those who do not.[17] Two studies of teacher training by economists find different results. Jacob & Lefgren (2004) study a low-intensity training program and find no statistically significant effect of receiving additional hours of teacher training on student outcomes in Chicago using a regression discontinuity design. In contrast, Angrist and Lavy (2001) find that a high-intensity in-service teacher-training program in Israel was associated with test score improvements of between 4 and 8 percentile points. This seemingly successful training program was designed to improve teaching skills (rather than provide course content) and involved a mixture of counseling and feedback sessions for teachers, changes in the organization of class time, and training in the use of instructional aids. The difference between these studies echoes the general trend that successful professional development programs for teachers tend to be high-intensity multifaceted programs that involve multiple sessions per year (Pianta (2011)).

[17] Indeed, Yoon et al. (2007) finds only nine studies based on credible experimental or quasi-experimental designs in a review of more than 1,300 studies on the effects of teacher professional development on K-12 student outcomes. Among these studies, only those programs that included more than fourteen hours of training showed a positive and significant effect on student achievement, but formal professional development programs typically last for the equivalent of 8 hours (Lewis et al. 1999; Mullens et al. 1996).

We now discuss the two most common high-intensity professional development programs and discuss those models that are likely to be most effective at improving teacher skills and student outcomes.

Peer mentoring

One common approach to improving teacher skills is mentoring or "induction," which is provided, in a variety of forms, to the vast majority of new teachers in the United States and a majority of states require mentoring to be given to new teachers. Despite their prevalence, only recently empirical evaluations of mentoring programs surfaced based on credible research methodologies.[18]

Rockoff (2008) studies the impact of a peer mentoring program in New York City launched after a state requirement was put into place. Newly hired first year teachers were assigned a mentor who was expected to meet with them on a weekly basis. Mentors were given training and a detailed program to help improve mentee teachers' instructional skills. Using quasi-experimental variation in the assignment of mentors to teachers as an identification strategy, he finds strong relationships between measures of mentoring quality and teachers' claims regarding the impact of mentors on their success in the classroom but weaker evidence of effects on teacher absences, retention, and student achievement. Importantly, teachers assigned to mentors who logged a greater total number of hours meeting with mentees experienced improvement in student

[18] See Ingersoll & Strong (2011) for a recent review of this literature.

achievement in math and English. This underscores the potential importance of high-intensity professional development that is sustained over a relatively long period of time.[19]

The most ambitious study of teacher mentoring is a randomized evaluation conducted by Glazerman et al., (2008, 2010) and Isenberg et al. (2009). They study how two teacher induction programs affect beginning teachers' retention, classroom practices, and student achievement over a three year period.[20] The authors randomly assigned 418 schools to be treatment or comparison schools, and beginning teachers in treatment schools received a high-intensity "comprehensive" induction for one or two years, which included weekly meetings with a full-time mentor, monthly professional development sessions, opportunities to observe veteran teachers, and continuing evaluation of the teachers' own practices. Beginning teachers in control schools received any support normally offered to teachers by the school. There was no effect of the intervention in the first two years either on students, measures of teaching practice, or on teacher attrition. In the third year of the program, the scores of students taught by teachers receiving two years of comprehensive induction improved by about 4 percentile points. However, due to substantial attrition of teachers after three years, this estimate is relatively imprecise and not robust to changes in the empirical specification.

Taken together, these studies suggest that high-intensity teacher induction programs that include sustained peer mentoring from high quality mentors over the course of the entire school year could be an effective policy, but the evidence is still fairly thin. Moreover, these programs also included some evaluation and feedback, which are an important component of many professional

[19] A caveat is that one cannot interpret these estimates strictly as a return to mentoring hours, since mentors who logged more hours overall may have also provided higher quality mentoring per hour.
[20] One of the two was based on the same model as the program studied by Rockoff (2008).

development programs and may have an independent effect on teacher outcomes. We discuss such programs below.

Evaluation programs

Many professional development programs include an evaluation component. Typically, teachers are observed in the classroom by an expert and receive feedback on their teaching with advice on how to improve. Such programs are predicated on the idea that classroom practices associated with better student outcomes have a *causal* effect on student learning, so that professional development programs that target and promote those observed behaviors should be effective.

In principle, evaluation may improve student outcomes through two channels. The act of measuring teacher productivity may create an incentive for teachers to exert more effort in order to secure positive performance evaluations. Alternatively, evaluations may provide teachers with helpful information on where they are deficient and how they can improve, lowering the costs for teachers to improve their teaching practice.

Despite these possibilities, most of the existing evaluation programs are viewed as being poorly designed and therefore ineffective, either because all but a handful of teachers are given top ratings (Weisberg et al., 2009) or because the system is plagued by "*vague district standards, poor evaluation instruments, overly restrictive collective bargaining agreements, and a lack of time [as well as] the absence of high-quality professional development for evaluators, a school culture that discourages critical feedback and negative evaluation ratings*." (Donaldson, 2009 p. 2).

Nevertheless, recent work suggests evaluation systems may not be ineffective *per se*, but just poorly implemented. Allen et al. (2011) study a program designed to promote a specific set of observable behaviors and practices positively correlated with student learning gains.[21]The program included three main components; (1) a video library of actual classrooms and analysis of these classrooms to provide teachers with opportunities to *observe* effective teacher-student interactions, (2) *Skills training* for teachers in identifying effective and ineffective practices, and (3) *individualized feedback on* and *analysis of* one's own classroom practices. Coaching occurred every two weeks and was repeated throughout the school year. In a randomized controlled trial where all teachers had access to the video library, the treatment group, which received additional coaching and evaluation, demonstrated improvements in observable teaching practices and student test scores that were about 5 percentile points (about 0.14σ) higher.

Taylor and Tyler (2012) study the effects of the evaluation system in Cincinnati Public Schools, where teachers were evaluated based on specific criteria linked to higher achievement; teachers were observed in the classroom by peers and experts and received detailed feedback about where they were deficient and how to improve.[22] Finally, the evaluation outcomes were linked to career development such that teachers who did not have strong evaluations had to undergo a year-long process of intensive assistance from a mentor that included another year of evaluation with more frequent observations. These evaluations are done periodically on a predetermined schedule,

[21] This "CLASS Framework" (see Hamre & Pianta (2010)) is motivated by a theory of positive classroom interactions organized into three domains – Emotional Supports, Classroom Organization, and Instructional Supports. These CLASS behaviors and practices are oriented toward a broad range of positive student outcomes (including outcomes such as student engagement, and positive classroom dynamics) and are not focused only on performance on achievement tests. CLASS behaviors were validated using data on thousands of classrooms across various grade levels and content areas.

[22] Teachers are evaluated on dozens of specific skills and practices and scored using a well-known rubric developed by Danielson (1996).

usually every five years, and the authors employ a quasi-experimental design, comparing the achievement of individual teachers' students before, during, and after the teacher's evaluation year. The authors find that students assigned to a teacher in a post-evaluation year score about 0.1 standard deviations higher in math than similar students taught by the same teacher prior to evaluation. The magnitude of this estimate is notable given that the sample being studied comprised mid-career teachers who many may have assumed could no longer acquire new skills. The fact that the performance gains were sustained even after the evaluation year indicates that while incentive effects might have been important, these programs lead to real persistent increases in teacher skills. While this may not be the only successful model of professional development, it is one that has been proven effective.

3. ***Pay for performance.***

Public school teachers in the US have traditionally been paid according to salary schedules based on years of experience and education level, so that teacher pay is largely unresponsive to actual teacher performance (Podgursky and Springer, 2007). In other contexts, worker effort and worker output are found to be higher when workers are paid for performance on the job (Foster and Rosenzweig, 1994; Lazear, 2000). If teaching is anything like other occupations, rewarding teachers for their performance may increase teacher effort and improve student outcomes.

While performance pay is a promising idea, theoretically there are reasons why performance pay may be only weakly related to student achievement growth. First, we know from research on the estimation of teacher effects that student test scores are influenced by a variety of factors that are

outside the control of the teacher. This problem can be reduced statistically by accounting for the influence of student attributes and family influences, much like the value-added approach discussed earlier. However, if these outside influences fluctuate over time in ways that are hard to predict, teachers may perceive a weak link between their effort and their pay, and accordingly not increase their effort. A second problem is that merit pay may not be effective at improving teacher performance if individual teachers do not know what to do to improve their teaching performance (Murnane & Cohen (1986)). If teachers do not know *how* to improve student outcomes, inducing them to exert more effort may simply be a waste of resources.

There are also general problems associated with incentive pay such as multi-tasking and gaming that could easily surface in the teaching profession. There are many student outcomes valued by society and teaching is a complex job. An incentive pay scheme focused on a limited set of measured outcomes may induce teachers to withdraw effort from other valuable dimensions of their jobs and, at worst, could induce actions that raise measured outcomes in ways that are without value. This is a particularly acute problem in education, where short term measures like test scores are only proxies for the development of human capital.[23] Thus, any well-designed pay-for-performance scheme must be based on outcomes that are a good measure of student learning and cannot be easily gamed. These potential problems underscore the importance of looking at effects of pay for performance on unrewarded outcomes and looking for effects that persist over time.

[23] Indeed, gains on awarded tests often do not generalize to low stakes tests of similar material (Jacob 2005; Holcombe et al. 2013), and Jacob and Levitt (2003) finds evidence of teacher cheating on student tests in order to improve school performance under a district accountability regime.

Since the 1990s, pay for teacher performance has been adopted in many nations worldwide and in many districts in the United States, but it remains relatively uncommon.[24] Where there is a close correspondence between teacher effort and the rewarded performance, it is difficult to improve rewarded tasks without increasing student learning, performance is well measured, and teachers know how to improve the rewarded outcomes, a pay-for-performance contract should elicit more effort and improve student outcomes more than a standard wage or salary contract (as most teachers have). However, it is unclear whether these conditions are satisfied in typical settings where performance pay schemes have been tried.

a. Empirical Evidence

On the whole, the empirical evidence on the effects of teacher performance pay suggests that it can, and often does, improve student outcomes, particularly those outcomes on which rewards were based (Neal 2012). Many of the most positive results are based on experimental studies from outside the United States. In contrast, studies using US samples have yielded mixed results. In this section, we review both the evidence from outside of the US and the evidence from US-based studies and then aim to reconcile these two literatures by highlighting the design features that made certain programs (irrespective of the geographic location) more or less likely to succeed.

Evidence outside the US

[24] For international examples of performance pay, the Pay Performance and Management Reform in the United Kingdom, the Victorian Government Schools Agreement in Australia, the Carrera Magisterial Program in Mexico, the National System of School Performance Assessment in Chile.

In an early analysis of a pay-for-performance system, Lavy (2009) analyzes an experimental program in Israel that offered individual teachers bonus payments on the basis of the performance of their classes on high school graduation exams in English and mathematics. The rewards ranged between 6 and 25 percent of the average teacher's salary, and were structured as a tournament, with teachers competing against other teachers of the same subjects in the same school. Lavy finds that the intervention increased overall pass rates by 12 percent and average math scores by 10 percent. Effects were about half the size in English. Using survey data, he finds that improvements were mediated through changes in teaching methods, enhanced after-school teaching, and increased responsiveness to students' needs.

Positive results have also been found for a pay-for-performance scheme in England that rewarded individual teachers at least an 8 percent permanent salary increase for improving average student performance. Atkinson et al. (2009) use the fact that eligibility for the rewards was conditional on having a minimum level of experience to identify the impact of the program; they therefore compare within-teacher trajectories of student achievement growth for eligible and ineligible teachers in the same schools before and after the creation of the program. They find that the payment scheme improved secondary school test scores, and value added increased on average by about 40 percent of a grade per pupil.

Muralidharan and Sundararaman (2011) and Muralidharan (2012) analyze an experimental program in India that provided bonus payments to primary school teachers based on the average improvement of their students' test scores in independently administered learning assessments (with a mean bonus of 3 percent of annual pay). After two years, students in incentive schools

performed significantly better than those in control schools by 0.28 and 0.16 standard deviations in math and language tests, respectively. For students who completed five years of primary school under the program, test scores increased by 0.54 and 0.35 standard deviations in math and language, respectively, and also by 0.52 and 0.3 standard deviations in science and social studies, for which incentives were not provided.

Note that all these programs rewarded individual teachers for individual teacher outcomes and based rewards on average test scores rather than some proficiency cut-off. Taken together, these studies show that individual teacher incentive pay for average test score gains can lead to sizable improvements in student outcomes. Moreover, they demonstrate that this is true both in developing and developed nations.

Evidence in the US

There is substantial new evidence on pay for performance in the US from randomized experiments, but with mixed results. Some studies show little effect of pay for performance, which has led some to speculate that performance pay cannot work in the United States. Other studies find positive impacts on incentivized outcomes, but only under particular design features or with negative spillovers onto non-incentivized outcomes. Thus, while it is possible that there is something different about teachers and students in the United States that renders teacher performance pay ineffective, design features of the programs may explain the differences in results.

Goodman and Turner (2013) and Fryer (2013) analyze a group incentive program in New York City. Under this program, a random sample of schools participated in a bonus pay scheme that involved team incentive pay at the school-level linked to test score growth targets. The bonuses ranged from between $1500 and $3000 per teacher (between 2.5% and 5% of the average teacher salary in New York City). The authors found that the bonus program had little impact on teacher effort, student performance in math and English, or classroom activities.

Goodman and Turner (2013) highlight the free-rider problem in how the program linked incentive pay to school-wide performance goals. They test for a free-rider problem by seeing if the effects are larger in smaller groups where the free rider problem should be less severe, and find that this is indeed the case. This is also consistent with two studies using quasi-experimental methods to study teacher incentive programs in the U.S. as well as international evidence.[25]

Apart from the group incentive structure, this NYC program based rewards on a performance threshold, rather than rewarding general improvement.[26] Where teachers are responsible for average test scores, they have an incentive to improve the outcomes of all students. However, when teachers are responsible for reaching a performance threshold, they only have an incentive to expend effort on those students who can be pushed over this threshold (Neal & Schanzenbach

[25] Lavy (2002) and Muralidharan (2012) also find positive, but significantly smaller effects of group based incentives in their respective settings. Sojourner et al. (2011) compare the effects of different kinds of pay-for-performance schemes in Minnesota and find that districts offering greater rewards for teacher-level goals experienced large gains in reading, whereas those offering rewards based on school-wide goals or subjective evaluations did not. Imberman & Lovenheim (2012) study the impact of a group-based performance pay system in Texas. Groups were defined at the subject-grade level, so the power of incentives directed at individual teachers varied both over time for the same teacher and for the same teacher in the same year across subjects and/or grade levels. They find robust evidence of weakened incentives when rewards are based on the collective performance of large groups of teachers, with ideal group sizes of three to four teachers.

[26] An additional caveat is that it provided the bonuses for the same outcomes that were sanctioned under the district wide accountability system. As such, even the comparison schools had strong incentives to meet the same targets.

(2010)). If the performance threshold is too low, many schools can meet the standard by expending no additional effort, and if the performance threshold is too high, many schools will realize they not meet the standard even if they expend additional effort and will therefore chose not to do so. In the NYC program, almost 90 percent of school earned awards, suggesting that the performance standard may have been too low to induce increased effort.

Another influential US based finding of no effect of performance pay on test scores focuses on the POINT program in Tennessee. Under this system, middle school mathematics teachers voluntarily participated in a controlled experiment and were randomly offered financial rewards for exhibiting "*unusually large gains on standardized tests*". Specifically, teachers could earn rewards of $5000, $10000, or $15000 if their students' scores were at the 80th, 90th, or 95th percentiles, respectively, of the historical distribution of test score gains on the standardized state test.[27] Springer et al., (2010) find that students whose teacher was eligible to receive bonus payments performed at the same level as those whose teachers were ineligible.

Neal (2012) argues that the lack of results may be due to the performance targets being too high. Intuitively, because there is no monetary reward for achieving test score gains below the 80th percentile (even if those gains might be sizable), only those teachers who felt that they could attain above the 80th percentile of test score growth would have been induced to exert more effort. Because the targets were high, it is likely that the standard would have been out of reach

[27] Specifically, each student's score was normalized by subtracting the average score of students with the same prior test score in the previous two school years. These normalized scores were averaged across the teacher, and teachers were assigned a percentile rank in the historical distribution of the teacher average normalized scores, using classrooms from the prior two years. This has the flavor of the regression method used to estimate teacher effects shown in Equation 1, but is someone more transparent, is based on one year of data, and makes no adjustment for classroom and student noise components.

for most teachers. However, surveys of treatment teachers' beliefs of their own probability to achieve the bonus were, if anything, far too optimistic. Many treatment teachers believed their chances of winning were (well) above 50 percent, while objective predictions based on prior year performance suggest that most teachers' chances were (well) *below* 50 percent. Moreover, there appears to be no correlation between treatment teachers' subjective beliefs and their actual performance, and the majority of teachers did not think that the metric used in the POINT program could distinguish between effective and ineffective teachers. Thus, another potential explanation for the lack of findings is that teachers did not know how their teaching practice related to the incentivized metric. It is also possible that, after the first year of awards, they may have been discouraged that the actual probability of winning was far below their expectations. Nevertheless, the POINT experiment provides a strong note of caution to the potential effects of pay for performance programs.

A more positive note is struck by the results of a study by Fryer et al. (2012). In their randomized empirical design, teachers were paid based on their individual performance with each of their students (rather than based on a single performance threshold). The authors also varied payment schemes between one in which teachers are paid in advance and asked to give back the money if their students do not improve sufficiently (the loss aversion treatment) and one in which teacher are paid an the end of the academic year (as is standard in the other studies discussed). The authors find large and statistically significant gains associated with the loss aversion treatment, but small and marginally statistically significant effects for the traditional gains treatment. Thus, there is evidence that a well-designed pay for performance system can work in the United States.

Alignment with Student Objectives

Insofar as student outcomes are a function of both teacher and students effort, and teachers cannot improve student outcomes unless students allow teachers to do so, one might expect pay-for-performance schemes to be more effective when the incentives of the teachers and students are aligned. Studies of programs that combine incentive pay for teachers with incentive pay for students suggest that student effort could be an important determinant for whether teacher performance pay will be effective. In a recent experimental study of student and teacher rewards, Behrman et al. (2012) evaluate the effect on test scores of three different performance incentives schemes using data from an experiment that randomized 88 Mexican high schools into three treatment arms and a control group. One treatment provided individual incentives for performance on curriculum-based mathematics tests to students only, another provided individual incentives for performance on curriculum-based mathematics tests to teachers only, and the third gives both individual and group incentives to students and teachers. The authors find that the effect of the combination of teacher and student incentives is larger than the sum of the effects of the teacher performance pay and student performance pay treatments. This provides further evidence that student effort and teacher effect are complementary and that alignment of student and teacher incentives is important.

Jackson (2010) analyzes the short and long run effects of a high school intervention that includes cash incentives for both teachers and students for each passing score earned on exams in Advanced Placement (AP) courses, teacher training, and curricular oversight. The program increased enrollment in AP courses and improved AP exam performance, doubling in the number of students taking and passing AP exams after 4 years. These effects are larger than

those found for programs that provide monetary incentives to teachers only or to students only, suggesting that the alignment of student and teacher incentives (though the combination of student and teacher rewards for the same outcomes) is important. Importantly, Jackson (forthcoming) also finds that the AP incentive program had long-term impacts on students' educational attainment, leading to increased college enrollment, persistence, completion, and higher adult earnings.

Selection Benefits of Incentive Pay

By tying teacher pay more closely to teacher performance, high-performing teachers will have a greater incentive to stay in the profession, and likewise low-performing teachers will have greater incentives to leave. Selection of individuals into the profession may also be affected. Higher pay for high-performing teachers may be particularly important if these teachers have higher potential outside wages. Chingos & West (2012) find evidence supporting this notion: among teachers leaving for other industries, a 1 standard deviation increase in a teacher's estimated value added is associated with 6 to 9 percent higher earnings outside of teaching but not associated with higher earnings within teaching. This is also consistent with Leigh (2012), who finds that higher teacher salaries induce higher ability students to select education as an undergraduate major.

Because most evaluations of performance-based teacher pay have been on short-run interventions that are unlikely to generate any selection effects, there is still little conclusive direct evidence of the selection effects of performance-based pay and the associated improvement in student outcomes. Goldhaber and Walch (2012) find some evidence of positive

selection by teachers into Denver's voluntary ProComp incentive scheme. Woessmann (2011) finds support for selection effects using cross-country variation in the use of performance-based pay and student achievement on international assessments of math, science, and reading. This is an important direction for future research.

Conclusion

There is now little doubt that there is wide variation in teacher effectiveness, and that the data and methods commonly applied in the field estimate a causal impact of teachers on student achievement. Based on these conclusions, states and school districts are rapidly implementing teacher evaluation policies that go well beyond the existing evidence, and urgently need answers to a new set of questions. The question is no longer are there teacher effects and can we estimate them, but rather how can we better estimate these effects and how should they be used in policy and practice?

To better estimate teacher effects, further work is needed in two broad areas. The first area is developing practical methods to pool information across multiple years and multiple measures to more precisely estimate predicted teacher effects. Large longitudinal datasets are being developed that will provide annual measures on a growing number of measures of teaching, and there is a pressing need for guidance on how to combine these measures to yield stable and reliable value-added measures that will be better able to discriminate between effective and ineffective teaching. The second area in which work is needed is developing and understanding measures of effective teaching that go beyond the average causal effect of teachers on state math

and English tests. For example, how should we estimate teacher effects on non-cognitive outcomes, in untested grades or subjects, or compare teachers across different contexts? How do teacher effects on these other measures relate to long-term economic and social outcomes? As administrators and teachers gain experience with measures of teacher effectiveness, these practical questions will become increasingly apparent.

Equally important is further work to better understand how measures of effective teaching should be used in policy and practice. What combination of teacher selection, mentoring and feedback, and pay for performance will be most successful? What specific features of these reforms and how they are implemented make the most difference? Over the next few years, state reforms of teacher evaluation policies will provide a rich laboratory for understanding how organizational design influences educational productivity.

References

Aaronson, D, Barrow L, Sander, W. 2007. Teachers and Student Achievement in the Chicago Public Schools. *J Labor Economics* 25(1): 95-135

Allen JP, Pianta RC, Gregory A, Mikami AY, Lun J. 2011. An interaction-based approach to enhancing secondary school instruction and student achievement. *Science* 333: 1034-1037

Angrist J, Lavy V. 2001. Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools. *J Labor Economics* 19(2): 343-69

Atkinson A, Burgess S, Croxson B, Gregg P, Propper C, Slater H, Wilson D. 2009. Evaluating the impact of performance-related pay for teachers in England. *J Labor Economics* 16(3): 251-61

Bakkenes I, De Brabander C, Imants J. 1999. "Teacher Isolation and Communication Network Analysis in Primary Schools." Educational Administration Quarterly. 35:166-202.

Behrman J, Parker S, Todd P, Wolpin KI. 2012. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. PIER Working Paper No. 13-004

Boyd D, Lankford H, Loeb S, Wyckoff J. 2011. Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness. Education Finance and Policy. 6(3):439-454.

Burgess S, Davies NM, Slater H. 2011. Do Teachers Matter? Measuring the Variation in Teacher Effectiveness in England. *Oxford Bulletin of Economics and Statistics* October 2012 74(5): 629-45

Cascio, Elizabeth U. and Douglas O. Staiger. 2012. "Knowledge, Tests, and Fadeout in Educational Interventions." NBER Working Paper #18038.

Chetty R, Friedman JN, Rockoff JE. 2013a. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. NBER Working Paper 19423.

Chetty R, Friedman JN, Rockoff JE. 2013b. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. NBER Working Paper 19424.

Chingos, MM, West, MR. 2012. Do More Effective Teachers Earn More Outside the Classroom? *Education Finance and Policy* 7(1): 8-43

Clotfelter CT, Ladd HF, Vigdor JL. 2007. Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review* 26(6): 673-682

Clotfelter CT, Ladd HF, Vigdor JL. 2010. Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects. *J Human Resources* Summer 2010 45(3): 655-681

Condie S, Lefgren L, Sims D. 2012. "Teacher heterogeneity, value added, and education policy." Working paper.

Daly AJ, Moolenaar NM, Bolivar JM, Burke P. 2010. Relationships in reform: the role of teachers' social networks. Journal of Educational Administration. 48(3):359-391.

Danielson C. 1996. Enhancing Professional Practice: A Framework for Teaching. ASCD

Dobbie W. 2011. Teacher Characteristics and Student Achievement: Evidence from Teach for America. Harvard University Working Paper.

Donaldson ML. 2009. So long, Lake Wobegon? Using teacher evaluation to raise teacher quality. Report, Center for American Progress

Elhert M, Koedel C, Parsons E, Podgursky M. (forthcoming). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in Missouri. *Statistics and Public Policy.*

Ferguson, R. 2009. *Tripod student survey, MET project upper elementary and MET project secondary versions*. Distributed by Cambridge Education, Westwood, MA.

Foster AD, Rosenzweig MR. 1994. A Test for Moral Hazard in the Labor Market: Contractual Arrangements, Effort, and Health. *Review of Economics and Statistics* 76(2): 213-27

Fryer, RG. 2013. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *J Labor Economics* 31(2): 373-407

Fryer RG JR., Levitt SD, List J, Sadoff S. 2012. Enhancing the Efficacy of Teacher Incentives Through Loss Aversion: A Field Experiment. NBER Working Paper 18237

Glazerman S, Dolfin S, Bleeker M, Johnson A, Isenberg E, Lugo-Gil J, Grider M, Britton E. 2008. *Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study*. NCEE 2009-4034. Washington, DC: U.S. Department of Education, National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences

Glazerman S, Isenberg E, Dolfin S, Bleeker M, Johnson A, Grider M, Jacobus M. 2010. *Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, June 2010

Goldhaber D, Chaplin D. 2012. Assessing the `Rothstein Falsification Test': Does It Really Show Teacher Value-Added Models Are Biased?" CEDR Working Paper 2011-5. University of Washington, Seattle, WA

Goldhaber DD, Goldschmidt P, Tseng F. 2013. Teacher Value-Added at the High-School Level: Different Models, Different Answers? Educational Evaluation and Policy Analysis. 35:220-236.

Goldhaber D, Gabele B, Walch J. (forthcoming). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics, and Public Policy.*

Goldhaber D, Hansen M. 2013. Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance. Economica, Vol 80(319), pp 589–612.

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.

Goldhaber D, Walch J. 2012. Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. *Economics of Education Review* December 2012 31(6): 1067–83

Goodman SF, Turner LJ. 2013. The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *J Labor Economics* 31(2): 409-420

Gordon R, Kane TJ, Staiger DO. 2006. "Identifying Effective Teachers Using Performance on the Job," *The Hamilton Project* white paper 2006-01, Washington, DC.

Hanushek EA. 1971. Teacher Characteristics and Gains in Student Achievement: Estimation using Micro Data. *American Economic Review* 61(2) 280-88

Hanushek EA. 2011. "The economic value of higher teacher quality." Economics of Education Review, 30:466-479.

Hanushek EA, Rivkin SG. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." American Economic Review, Papers and Proceedings. 100(2):267-271.

Hamre BK, Pianta RC. 2010. "Classroom environments and developmental processes: conceptualization & measurement." In Judith Meece & Jacquelynne Eccles, ed., *Handbook of Research on Schools, Schooling, and Human Development* (New York: Routledge, 2010): 25–41

Henry GT, Fortner CK, Bastian KC. 2012. The Effects of Experience and Attrition for Novice High-School Science and Mathematics Teachers. *Science* 335(6072) 1118-1121

Holcombe R, Jennings J, Koretz D. 2013. The roots of score inflation: An examination of opportunities in two states' tests. In G. Sunderman (Ed.), *Charting reform, achieving equity in a diverse nation* 163-189. Greenwich, CT: Information Age Publishing.

Imberman SA, Lovenheim MF. 2012. Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System. NBER Working Paper No. 18439.

Ingersoll RM, Strong M. 2011. The Impact of Induction and Mentoring Programs for Beginning Teachers: A Critical Review of the Research. *Review of Educational Research* June 2011 81(2) 201-33

Isenberg E, Glazerman S, Bleeker M, Johnson A, Lugo-Gil J, Grider M, Dolfin S, Britton E. 2009. *Impacts of Comprehensive Teacher Induction: Results from the Second Year of a Randomized Controlled Study* (NCEE 2009-4072). Washington, DC: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education

Jacob BA. 2005. Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools. *J Public Economics* 89(6): 761-796.

Jacob BA, Lefgren L. 2004. The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago. *J Human Resources* 39(1): 50-79

Jacob, BA, Lefgre L, and Sims D. 2010. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources*. 45:915-943.

Jacob BA, Levitt SD. 2003. Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating. *The Quarterly Journal of Economics* 118(3): 843-877.

Jackson CK. 2010. A Little Now for a Lot Later: An Evaluation of a Texas Advanced Placement Incentive Program. *J Human Resources* 45(3): 591-639

Jackson, CK. (2012) Recruiting, retaining, and creating quality teachers. *Nordic Economic Policy Review*, Number 1

Jackson CK. 2013. Non-Cognitive Ability, Test Scores, and Teacher Quality: Evidence from 9th Grade Teachers in North Carolina. NBER Working Paper No. 18624

Jackson CK. Forthcoming. Teacher Quality at the High-School Level: The Importance of Accounting for Tracks. *Journal of Labor Economics*

Jackson CK. Forthcoming. Do College-Prep Programs Improve Long-Term Outcomes? *Economic Inquiry*

Jackson, CK. Forthcoming. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers. *Review of Economics and Statistics*

Jackson CK, 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 TJ, Staiger DO. 2005. “Using Imperfect Information to Identify Effective Teachers.” Working paper.

Kane TJ, Staiger DO. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607

Kane TJ, Staiger DO. 2012. “Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.” MET Project Research Paper, Bill & Melinda Gates Foundation, Seattle WA

Kane TJ, McCaffrey DF, Miller T, Staiger DO. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*. Seattle, WA: Bill & Melinda Gates Foundation

Kinsler J. 2012. "Assessing Rothstein's Critique of Teacher Value-Added Models." *Quantitative Economics* 3: 333-362

Koedel C., Betts JR. 2007. Re-examining the role of teacher quality in the educational production function. University of Missouri WP 07-08.

Koedel C, Betts JR. 2011. Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. *Education Finance and Policy* 6(1): 18-42

Lavy V. 2002. Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement. *Journal of Political Economy* 110(6): 1286-1317

Lavy V. 2009. Performance Pay and Teachers’ Effort, Productivity and Grading Ethics. *American Economic Review* 99(5): 1979-2011

Lavy V. 2011. “What Makes an Effective Teacher? Quasi-Experimental Evidence.” NBER Working Paper #16885.

Lazear, EP. 2000. The Power of Incentives. *American Economic Review* 90(2): 410-14

Leigh A. 2012. Teacher pay and teacher aptitude. *Economics of Education Review* 31(3): 41–53

Lefgren L, Sims D. 2012. "Using Subject Test Scores Efficiently to Predict Teacher Value-Added." Education Evaluation and Policy Analysis, 34(1): 109-121.

Lewis L, Parsad B, Carey N, Bartfai N, Farris E, Smerdon B. 1999. *Teacher Quality: A Report on the Preparation and Qualifications of Public School Teachers*. (NCES 1999–080). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Lockwood JR, McCaffrey DF. 2009. "Exploring student-teacher interactions in longitudinal achievement data," Education Finance and Policy, 4(4): 439-467.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton. 2004. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics* 29(1):67-101.

McCaffrey DF, Sass TR, Lockwood JR, Mihaly K. 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy* 4(4): 572-606

Mihaly K, McCaffrey D, Staiger DO, Lockwood JR. 2013. "A Composite Estimator of Effective Teaching," MET Project Research Paper, Bill & Melinda Gates Foundation, Seattle, WA

Mullens J, Leighton M, Laguarda K, O'Brien E. 1996. *Student Learning, Teacher Quality, and Professional Development: Theoretical Linkages, Current Measurement, and Recommendations for Future Data Collection*. (NCES 96–28). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office

Muralidharan K. 2012. Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India. University of San Diego Working Paper.

Muralidharan K, Sundararaman V. 2011. Teacher Performance Pay: Experimental Evidence from India. *J Political Economy* 119(1): 39-77

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

Murnane RJ, Cohen DK. 1986. Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and Few Survive. *Harvard Educational Review* 56(1): 1-17

Murnane RJ, Willett JB, Levy F. 1995. The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics* 77 (2): 251–66.

Neal D. 2012 The Design of Performance Pay in Education. *Handbook of the Economics of Education* Chapter 6 Volume 4.

Neal, Derek A. and William R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences. *The Journal of Political Economy* 104(5): 869-895.

Neal D, Schanzenbach DW. 2010. Left Behind By Design: Proficiency Counts and Test-Based Accountability. *The Review of Economics and Statistics* May 2010 92(2): 263-83

Nye B, Konstantopoulos S, Hedges LV. 2004. How Large Are Teacher Effects? *Educational Evaluation and Policy Analysis* Fall 2004 26(3): 237-57

Ost, B. Forthcoming. How Do Teachers Improve? The Relative Importance of Specific and General Human Capital *American Economic Journal: Applied Economics*

Papay JP, Kraft MA. 2013. Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. Harvard University Working Paper

Parsad B, Lewis L, Farris E, Greene B. 2001. *Teacher preparation and professional development 2000*. U.S. Department of Education, National Center for Education Statistics, NCES 2001-88, Project Officer: Bernard Greene. Washington, DC: 2001

Pianta RC. 2011. *Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training*. Report, Center for American Progress

Podgursky MJ, Springer MG. 2007. Teacher performance pay: A review. *Journal of Policy Analysis and Management* 26(4): 909-50

Rivkin SG, Hanushek EA, Kain JF. 2005. Teachers, Schools and Academic Achievement. *Econometrica* 73: 417-458

Rockoff JE. 2004. The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review* 94(2): 247–52

Rockoff J. 2008. "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City." Manuscript.

Rockoff JE, Jacob B, Kane TJ, Staiger DO. 2011. Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*. 6(1):43-74.

Rockoff JE, Speroni C. 2010. Subjective and Objective Evaluations of Teacher Effectiveness. *American Economic Review* 100(2): 261–66

Rockoff, JE, Staiger DO, Kane TJ, Taylor ES. 2012. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools," *American Economic Review*, 102(7):3184-3213.

Rothstein, J. 2010. Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal of Economics* 125(1): 175–214

Rothstein, J. 2009. Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy* 4(4): 537-71

Rothstein J. 2012. "Teacher Quality Policy When Supply Matters." NBER Working Paper #18419.

Sojourner A, West K, Mykerezi E. 2011. *When Does Teacher Incentive Pay Raise Student Achievement? Evidence from Minnesota's Q-Comp Program*. Society for Research on Educational Effectiveness.

Spillane JP, Kim CM, Frank KA. 2012. Instructional Advice and Information Providing and Receiving Behavior in Elementary Schools: Exploring Tie Formation as a Building Block in Social Capital Development *American Educational Research Journal*

Springer MG, Ballou D, Hamilton L, Le VN, Lockwood JR, McCaffrey DF, Pepper M, Stecher BM. 2010. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University

Staiger DO, Kane TJ. 2013 "Making decisions with imprecise performance measures: The relationship between annual student achievement gains and a teacher's career value-added." Working paper.

Staiger DO, Rockoff JE. 2010. Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24: 97-117

Syverson, C. 2011. "What Determines Productivity?" Journal of Economic Literature, 49(2). 326-365.

Taylor ES, Tyler JH. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7): 3628-51

Tuttle CC, Gill B, Gleason P, Knechtel V, Nichols-Barrer I, Resch A. 2013. KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Mathematica Policy Research. Washington DC.

Weisberg D, Sexton S, Mulhern J, Keeling D**.** 2009. *The Widget Effect.* Brooklyn, NY: The New Teacher Project

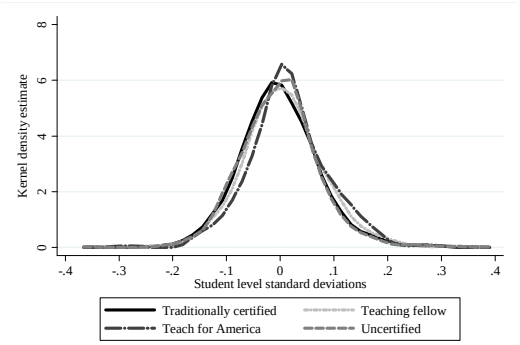
Wiswall M. 2013. The dynamics of teacher quality. *J Public Economics* April 2013, 100: 61-78

Woessmann L. 2011. Cross-country evidence on teacher performance pay *Economics of Education Review* June 2011 30(3): 404–18

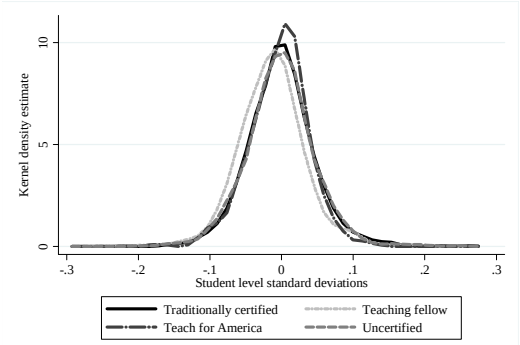
Yoon KS, Duncan T, Lee SWY, Scarloss B, Shapley KL. 2007. *Reviewing the evidence on how teacher professional development affects student achievement.* (Washington: National Center for Education Evaluation and Regional Assistance)

Figure 1: Variation in Value-Added by Teacher Certification in Kane et al. (2008)

Panel A: Math



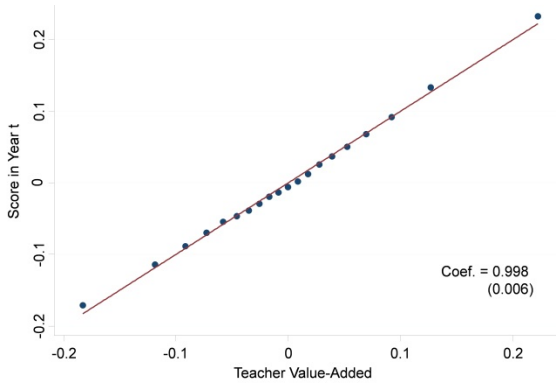
Panel B: English



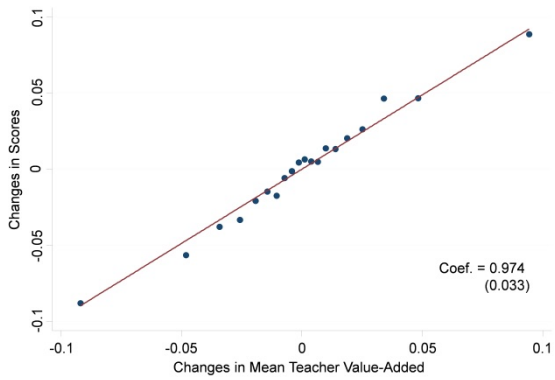
Notes: Panels A and B are reproduced from Figure 3 in Kane et al. (2008) and show, respectively, distributions of math and English value-added estimates for teachers in New York City. Distributions are calculated separately for four groups of teachers, based on their teaching certification.

Figure 2: Quasi-Experimental Testing in Chetty et al. (2013a)

Panel A: Cross-Class Variation



Panel B: Cross-Cohort Variation



Notes: Panels A and B are reproduced from Figure 2a and Figure 4a, respectively, in Chetty et al. (2013a). Panel A shows a binned scatter plot of test score residuals vs. teacher value-added. Panel B shows a plot of cross-cohort changes in mean test scores vs. changes in mean teacher value-added at the school-grade level; these changes are also de-meaned by school year to eliminate secular time trends. Observations are divided into twenty equal-sized bins (vingtiles) based on the x-axis variable (value added in Panel A, change in mean value-added in Panel B), and the mean of the y-axis variable (residual test score in Panel A, change in mean score in Panel B) within each group is plotted against the mean of the x-axis variable within each bin. The solid line shows the best linear fit estimated on the underlying micro data using OLS. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.