IDB WORKING PAPER SERIES No. IDB-WP-438

Theory and Evidence on Teacher Policies in Developed and Developing Countries

Emiliana Vegas
Alejandro Ganimian

August 2013

Inter-American Development Bank
Education Division

Theory and Evidence on Teacher Policies in Developed and Developing Countries

Emiliana Vegas
Alejandro Ganimian

Inter-American Development Bank
2013

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library

Vegas, Emiliana.
Theory and evidence on teacher policies in developed and developing countries / Emiliana Vegas, Alejandro Ganimian.
p. cm. (IDB working paper series ; 438)
Includes bibliographical references.
1. Teacher effectiveness. 2. Teachers—Training of. 3. Teachers—Rating of. 4. Teachers—Recruiting. 5. Teachers—Selection and appointment. I. Ganimian, Alejandro. II. Inter-American Development Bank. Education Division. III. Title. IV. Series.
IDB-WP-438

http://www.iadb.org

Theory and Evidence on Teacher Policies in Developed and Developing Countries

Emiliana Vegas
Alejandro J. Ganimian*

Abstract

The past decade has seen the emergence of numerous rigorous impact evaluations of teacher policies. This paper reviews the economic theory and empirical evidence on eight teacher policy goals: (1) setting clear expectations for teachers; (2) attracting the best into teaching; (3) preparing teachers with useful training and experience; (4) matching teachers' skills with students' needs; (5) leading teachers with strong principals; (6) monitoring teaching and learning; (7) supporting teachers to improve instruction; and (8) motivating teachers to perform. The paper also discusses key concepts and methods in econometrics to understand existing studies and offers some directions for future research.

JEL codes: I250 Education and Development; J450 Public Sector Labor Markets; I210 Analysis of Education.

* The authors would like to thank Susanna Loeb, Pilar Romaguera, Agustina Paglayán, Analía Jaimovich, Andrew Trembley, and Nicole Goldstein, who were key members of the World Bank's SABER Teacher Project team, from which we draw extensively in this paper. Emiliana Vegas is Chief of the Education Division at the Inter-American Development Bank (evegas@iadb.org). Alejandro J. Ganimian is a Doctoral Candidate in Quantitative Policy Analysis in Education at the Harvard Graduate School of Education, and a Doctoral Fellow in the Multidisciplinary Program in Inequality and Social Policy at the Harvard Kennedy School of Government (alejandro_ganimian@mail.harvard.edu).

Introduction

The past decade has seen the emergence of numerous rigorous impact evaluations of teacher policies. These studies differ from previous research in three ways that make them useful to inform policy decisions. First, instead of relying on proxies for student learning and teacher effectiveness, such as student enrollment or teacher certification rates, they capitalize on data on student learning from national and international assessments. Second, rather than identifying associations between policies and outcomes, they employ methods that allow them to distinguish the effects of interventions from other factors that may confound those effects. Finally, while they assess the impact of specific reforms, they also explore how these reforms interact with one another.

This paper distills the main lessons from some of the more recent of these studies on teacher policies. The first section provides an analytical framework to structure the review of the evidence. Using this framework, the second section discusses the economic theory that is driving empirical work on teacher policies and the key findings of the most rigorous studies in this area. The final section concludes by making sense of the theory and evidence on the impact of various teacher policies on student outcomes, and by highlighting the main findings and questions for further research.

Analytical Framework: Teacher Policies as a System

One way to think about the issues that have motivated research on teacher policies is to organize them according to the challenges that education systems face at different stages of teachers' careers—ranging from when individuals seek to enter teaching to when they join a school as teaching staff and become classroom instructors. The main policy issues that education systems face in managing teachers effectively are: (1) setting clear expectations for teachers; (2) attracting the best into teaching; (3) preparing teachers with useful training and experience; (4) matching teachers' skills with students' needs; (5) leading teachers with strong principals; (6) monitoring teaching and learning; (7) supporting teachers to improve instruction; and (8) motivating teachers to perform. While there are other ways to organize the evidence, this framework is reminds us that policies such as certification requirements or evaluations are designed to achieve goals, and should thus be evaluated on the extent to which they reach them.

The degree to which an education system addresses the challenges in one stage of the teacher pipeline can make it easier or harder to tackle the challenges in subsequent stages. Also, the extent to which an education system deals with these challenges for a stock of teachers at a point in time will make it easier or harder for the system to deal with challenges for the incoming flow of teachers. As we argue elsewhere, teacher policies

form a dynamic system—changes in one area have repercussions on all others (Vegas et al. 2011).

Theory and Evidence: Teacher Labor Markets

This section draws on economic theory and empirics to discuss the importance of the teacher policy goals above and review the most rigorous available evidence on interventions involving each one of these goals.

Economics has played an important role in recent rigorous research on teacher policies, including the problems identified and the solutions proposed to address these problems. So while there are many ways to examine the evidence on teacher policies, reviewing the economic theory that has driven empirical work can help us understand why the evidence has developed the way that it has.

There are several aspects that distinguish recent research on teacher policies from previous empirical work. Perhaps the most distinctive feature of recent studies is their capacity to estimate the effects of interventions on student and teacher outcomes. At first, the fact that this has not been the norm may appear surprising, as it would seem that all impact evaluations should aim to study these outcomes given that the objective of interventions is to promote positive changes. However, until recently, most empirical work in education was unable to disentangle the effects of an intervention (known as "treatment effects") from the effects associated with the characteristics of those participating in the program ("selection effects"). This problem, known as "selection bias," has limited the impact of educational research on policy (Murnane and Willett 2011).

Over the past decade, however, education researchers have made remarkable progress in their capacity to address selection bias. In particular, they have gotten better at providing reliable answers to two questions of prime interest to policymakers: (1) how would participants in a program have fared in the absence of the program? and (2) how would nonparticipants have fared in the presence of the program? Answering these questions definitively is impossible because at any point in time an individual is either participating or not participating in a program. Approximating an answer by comparing program participants before and after their participation is not good enough, since several things could have happened to those individuals while they participated in the program that could have influenced the outcomes to be measured. This is known as the "program evaluation problem" (Duflo, Glennerster and Kremer 2008). The innovation in recent research has been to use methods to find an adequate comparison group for program participants and to offer a credible estimate of the two "counterfactual" questions above.

The latest, most rigorous generation of studies on teacher policies in both developed and developing countries uses econometric methods such as randomized control trials, differences-in-differences analysis, instrumental variables estimation, and regression discontinuity designs to offer unbiased estimates of the effects of teacher policies.

This review focuses on rigorous studies that assess the impact of teacher policies on student learning, as measured by standardized tests. The aim is not to imply that increased student learning forms the only outcome of a well-functioning education system.[3] Instead, the focus is driven by the increasing interest of both developed and developing country governments in education policies that raise student achievement, and by a body of evidence that links learning to economic development (Hanushek and Woessmann 2007). Reference is made to less-rigorous studies (using propensity score matching or fixed effects) in two cases: (1) when there are no rigorous studies on a given teacher policy; and/or (2) when the objective is to draw attention to potential differences in the evidence between developed and developing countries.

The boxes in this paper feature brief overviews of the methods employed in the studies as well as some of the commonly used statistical concepts. The aim is to bring policymakers and other stakeholders into the discussion of the evidence and make them active contributors.

The effects of programs are mostly reported in standard deviation units, which are commonly used in statistics to measure how spread out values are within a distribution (e.g., how much students' test scores vary in a class). A standard deviation is calculated through a process that might seem complicated and arbitrary to a nontechnical audience. Yet, standard deviations are used because they offer a common metric to compare the effects of different programs on outcomes such as student test scores. In the social sciences in general, effect sizes of .80 of a standard deviation are considered to be large, effects of .50 moderate, and effects of .20 small (Cohen 1988). As noted by Murnane and Willett (2011), however, even the most successful interventions in education have small effects by these standards.

In interpreting standard deviations, it helps to think about them in one of five ways. One way is in terms of what they say about how much student learning improves, typically in student achievement tests of math and reading. If one assumes that students' skills follow a "normal" distribution so that most students are clustered around an average skill level

[3] Rigorous research has documented the impact of education on important short-term outcomes such as student satisfaction and long-term outcomes such as labor market outcomes, educational attainment, and crime. These outcomes are not always captured by test scores (see Chetty et al. 2011; Deming 2011; Kane and Staiger 2010, 2012; Kemple and Willner 2008).

and there are few at the lowest and highest levels—and the psychometricians who design tests make sure this is the case—then an effect size in standard deviations can be easily translated into a change in a student's percentile rank (Box 1). For example, an improvement by .25 of a standard deviation means that a student previously performing at the 50th percentile now performs at the 60th percentile.

Box 1. What Are Percentile Ranks?

In statistics, percentiles ranks are often used to understand where an individual falls within a distribution. Calculating a percentile rank is straightforward, but it is not necessary to understand how to do so in order to understand what percentile ranks mean. The rankings are simply the percentage of people below that individual. For example, if a test is administered and a student scores in the 10th percentile, that means that 10 percent of the student's peers scored below him or her—or equivalently, that 90 percent of the student's peers outscored him or her. Similarly, an improvement from the 10th to the 15th percentile means that while a student initially outscored 10 percent of his/her peers, the student now outscores 15 percent of them.

Another way to make sense of effect sizes is to compare them to reference points. Many studies in the United States compare the effects of a program to the black-white achievement gap, which is between .65 and .75 of a standard deviation (Reardon 2011). Others compare effect sizes to "grade equivalents"—i.e., how much learning takes place in a year of school. Kane and Staiger (2012) equate .25 of a standard deviation to nine months of schooling in grades 4 through 8. Yet others understand these effects in terms of the tangible benefits that they have for children. Krueger (2003) has estimated that a 1 standard deviation increase in elementary math scores is associated with an 8 percent increase in adult earnings. It is also possible to compare effect sizes of new policies to those of widely known interventions. In the United States, many compare the effects of new interventions to that of a high-profile class-size reduction policy discussed later in this paper, which yielded .20 of a standard deviation effect (Hanushek 2003).

Setting Clear Expectations for Teachers

Economic theory indicates that setting clear expectations is important from the point of view of both recruiting and managing effective teachers. Expectations influence how potential entrants into teaching perceive the profession. As Roy (1951) and Borjas (1987) argued, individuals "self-select" into job markets by choosing the one that is most likely to give them the highest expected earnings, based on their assessment of the skills that they possess and those that are demanded by the job. Professional standards can give employers a mechanism by which to "screen" those who are most likely to succeed (Stiglitz 1975). A profession with low or unclear standards creates an imbalance in the information that employers and potential employees have about the skills of the latter.

This is known in economics as "information asymmetry." It is problematic because it encourages low-skilled individuals to try to enter professions for which they might be unprepared, discouraging high-skilled entrants—a problem that economists refer to as "adverse selection" (Akerlof 1970; Greenwald 1986).

Second, expectations can guide teachers' work. As Baker (1992) and Prendergast (2002) indicated, clear professional standards can minimize "principal-agent problems" in which a principal (e.g., an education system) hires an agent (e.g., a teacher) to perform tasks that are in the interest of the principal but that are costly to the agent and difficult to observe (e.g., raising student learning). Professional standards can mitigate these problems by serving as the basis for a "contract" in which the employer and employee agree on the tasks to be performed and align the ways in which performance will be measured with the employee's compensation.

A body of research focuses on the extent to which expectations for teachers' work are clear, fair, and/or ambitious enough to produce the desired levels of student learning. These expectations include the explicit guidelines for teachers' work in the classroom, such as curricula or learning standards, and characteristics of the system, such as the time allotted for instruction or the number of students per classroom. These guidelines implicitly dictate the conditions under which teachers are expected to conduct their work effectively.

Rigorously evaluating expectations for teachers' work is challenging because the evaluations generally are applied to all students in a national or subnational education system. Therefore, as for other universal policies, it is hard for researchers to find an adequate comparison group. Yet, researchers have made important inroads in this field of study by tinkering with existing expectations and randomly assigning groups of schools within an education system to business-as-usual and new conditions—a method discussed below.

Table 1 summarizes three types of interventions that have been rigorously evaluated: (1) scaffolding teachers' work; (2) increasing instructional time; and (3) reducing class size. The last two types of interventions might not seem as related to setting clear expectations as the first one, but they are included here because they inform our understanding of the adequacy of current expectations for teachers' work.

Scaffolding Teachers' Work

One way in which education systems have sought to improve student and teacher performance is by offering teachers more structured guidelines for their classroom work.

One intervention in the United States that offers scaffolding for teachers is "Success for All" (SFA), a reading program with a highly structured school-wide curriculum that uses novels and basal readers, periodically regroups students across age and grade boundaries, and requires students to engage in reading at home. In 2000, Borman et al. (2007) evaluated SFA using a Randomized Control Trial (RCT) (Box 2). The authors randomly assigned 41 high-poverty schools into a grade K–2 SFA treatment or a grade 3–5 SFA treatment and compared kindergarten and first grade students in these two groups. They found that the program had positive effects on three literacy outcomes. The effect sizes ranged from .21 of a standard deviation on passage comprehension to .33 on the word attack measure. These results suggest that changing expectations in schools that serve disadvantaged children can potentially affect achievement by making teachers' job more manageable.

Box 2. What Is a Randomized Control Trial?

Randomized Controlled Trials (RCTs) have become increasingly popular. Their main benefit is that they allow researchers to obtain a "clean" estimate of a program's average effect by comparing the outcomes of individuals who participated in the program (the "treatment" group) to those of individuals who did not (the "control" group). When a program is randomly assigned, the only factor that distinguishes persons in the treatment and control groups is chance; neither group has characteristics that affect the outcomes of interest. As a result, the threat of selection bias disappears and the effect can be estimated as the difference between the average outcomes of the treatment group minus those of the control group.

There are more rigorous evaluations of initiatives that increase the scaffolding provided to teachers in developing countries, where teacher capacity tends to be much lower. In fact, two recent studies in India suggest these types of interventions can have a considerable impact on teacher pedagogy and student learning. In Maharasthra, He, Linden, and MacLeod (2007) used an RCT to evaluate a program designed to teach English that could be implemented either through a specially designed machine or flashcards. They found that both versions of this program yielded gains in English achievement of about .30 of a standard deviation, and that they proved particularly effective with older and lower-performing students. They also found that the version of the program taught by teachers not only improved students' English scores, but also their math scores. This suggests investing in improving teachers' pedagogy (rather than in replacing it) had some positive spillover effects.

The other study in India suggests, however, that while scaffolding might be effective at producing simple changes in teacher pedagogy, the low capacity of the teaching force in some schools might limit the extent to which more complex changes can be achieved. In 2004, in Mumbai, He, Linden, and MacLeod (2009) randomly assigned students in pre-

schools, primary schools, and stand-alone reading classes either to a program that changed the reading curriculum and provided new activities for teachers, or to no treatment at all. The program produced gains in students' reading scores of .26-.70 of a standard deviation, and it was most effective with pre-school and low-performing students. Yet, the version that was taught out of school was more effective than the one during school time, yielding an additional .24 of a standard deviation effect. The greater impact of the program for out-of-school hours suggests scaffolding might not be enough to produce the types of changes in teacher pedagogy needed to radically improve student learning.

Even relatively simple changes in classroom activities have been shown to improve learning in developing countries. In the Tarlac province of the Philippines in 2009, 5,510 fourth grade students were randomly assigned either to a treatment group that received in-service training for teachers, reading materials, and a reading marathon, or to a group that received no intervention (Abeberese, Kumler, and Linden 2011). One month after the program was implemented, the number of books students read increased from 2.3 to 9.5 and students' reading scores increased by .13 of a standard deviation. Further, the effects persisted with time. Three months after the reading marathon, treated students still read 3.1 books more than those in the control group and had reading scores that were .06 of a standard deviation higher.

Two studies in Madagascar indicate, however, that systemic problems can limit the extent to which improvements in pedagogy can be taken to scale. In 2005, the country provided teachers with tools that specified their duties in detail (e.g., an operations manual). This initiative was evaluated using an RCT: researchers recruited 30 districts for the evaluation and randomly assigned 15 of them to receive this intervention, while letting the others continue to run their schools as they usually would.

In a first study of this intervention, Lassibille et al. (2010) found that after two years schools where the program had been implemented had higher student attendance and lower repetition rates than control schools, but only when coupled with interventions at the district and subdistrict levels that improved workflow. However, even with these interventions, the test scores of the treatment schools were not statistically significantly higher than those of control schools. (See Box 3 for an explanation of statistically significant differences.) These results suggest that improving teacher management might not be enough when there are other systemic obstacles.

Box 3. What Are Statistically Significant Differences?

In simple terms, when a difference in the outcomes for two groups is "statistically significant," it means it is unlikely to have occurred by chance. The phrase has little to do with the size of a difference between two groups. In an evaluation, the group of students who participated in a program might have statistically significant higher outcomes than the group of students who did not participate, but these outcomes might be only marginally higher. In order to determine whether the difference between two outcomes is statistically significant, statisticians have developed tests that calculate whether the difference could have been found by pure chance. If this probability is too high, researchers will be reluctant to claim that differences in outcomes are stable. Defining what it means for these probabilities to be "too high," however, is a somewhat arbitrary process. In fact, researchers in education typically report three levels of statistical significance: 10, 5, and 1 percent. A 10 percent significance level means that there is a .10 probability of encountering a difference like the one observed by chance; that is, that there is a 1 in 10 chance that the difference observed between two groups is an artifact of the sample selected for the study.

In a second study of the same intervention, Glewwe and Maïga (2011) found that the average program effect was the same for civil service and contract teachers. The main difference between these two groups was that the former had tenure while the latter was hired through annual renewable contracts. This is an interesting question because Madagascar has hired a large number of contract teachers, who typically have less training but who may have higher incentives to work hard in order to get their contract renewed. If contract type is a good proxy for teacher experience or aptitude, these results imply that improving the efficiency of inexperienced teachers (who arguably need it most) may not be enough to raise student achievement.

These studies illustrate the potential of interventions that clarify expectations for teachers' work by producing improvements in pedagogy. But they also serve as a reminder that such initiatives might be hard to scale or ineffective in the presence of systemic obstacles.

Increasing Instructional Time

The United States has experimented with policies that expand the school day and year, and the results of these interventions are encouraging. Linden, Herrera, and Grossman (2001) evaluated a program in Washington, DC in 2006 that offered academic instruction, enrichment activities, and mentoring after school and during the summer to middle-school students. They found the program raised students' test scores by .09 of a standard deviation in reading and by .12 in math by the second year. They also found that students participating in the program were more likely to be proactive about enrolling in

high school by seeking information about schools, visiting schools, and talking to adults and peers about schools and where to apply.

In Chicago in 1996, the state government began mandating all students in grades 3, 6, and 8 who had not passed their end-of-year tests to attend a six-week remedial education program during the summer and then retake the exams. Students who still failed were held back a grade. Jacob and Lefgren (2004a) evaluated the effects of this summer program. Students were not randomly assigned to the program, so the researchers compared students who failed their exams by only a few points to those who passed their exams by a few points. Since students around the passing cutoff are likely to have a comparable level of skills—and given that all students who failed had to attend the summer camp while those who passed did not—comparing these two groups of students could yield a reliable estimate of the effect of the program for students around the passing cutoff. This empirical strategy is known as a Regression Discontinuity Design (Box 4).

Box 4. What Is Regression Discontinuity Design?

Regression Discontinuity Designs (RDDs) are commonly used when individuals are assigned to a program according to whether they reach a cutoff in some type of score, such as student achievement tests or teacher certification exams. In these cases, researchers can get an estimate of a program's effects by comparing the outcomes of individuals who just missed the threshold to those who were just above it. If one makes the reasonable assumption that individuals around the cutoff can be expected to have comparable outcomes (e.g., if one assumes that students just passing and just failing a test have similar abilities), then comparing the outcomes of these two groups after the program is enacted will give an estimate of its average effect. Importantly, however, the estimate will only apply to individuals around the cutoff, which is why it is commonly called the "local" average treatment effect.

Third graders who attended the summer camp gained about 20 percent of a year's worth of learning in the first year, with a 25-40 percent fadeout in the second year. Third graders who were held back also scored better than those who were not. Yet, Jacob and Lefgren (2004a) found no effects for sixth graders. These results are intriguing because they do not offer a good explanation of why the program had an effect only at some grade levels. Further investigation is thus required.

Studies such as that of Jacob and Lefgren suggest that increasing time in school can be an effective strategy to improve the learning outcomes of the most disadvantaged students. However, it is important to note that such studies examine targeted programs that serve disadvantaged schools or students and do not necessarily speak to the effectiveness of across-the-board increases in class time.

Reducing Class Size

Interventions reducing class size have been highly contentious. Proponents argue that decreasing student-teacher ratios can make classes more manageable and provide disadvantaged students with the personal attention they need. Opponents contend that merely lowering the number of students in a classroom will not boost learning unless it triggers changes in teacher pedagogy, which are unlikely to occur, and that hiring more teachers per student raises costs considerably.

Studying the effects of lowering class size is challenging because students are hardly ever randomly assigned to classes of different sizes. One way to obtain an estimate of the impact of these interventions is to exploit "class-size caps" that set a maximum number of students per class to conduct an RDD and compare the performance of students in classes on either side of that rule. Angrist and Lavy (1999) took advantage of a longstanding rule in Israel that set the maximum class size at 40 students. If a 40-student class enrolled an additional student, the rule required that the class be broken into two classes of 20.5 students on average. By comparing students on either side of the 40-student cutoff, the authors found that being assigned to a smaller class size had a positive effect on student achievement. This finding, however, was later challenged by Urquiola and Verhoogen (2009).[6]

Fortunately, governments in developed and developing countries have become increasingly interested in whether reducing class size improves student learning and have invested in RCTs to get a definitive answer.

By far the best-known rigorous evaluation of a class size reduction policy in a developed country was conducted in Tennessee in the United States. There, in 1985, the state government adopted a class-size reduction program called the Student/Teacher Achievement Ratio, which came to be known as Project STAR. The program randomly assigned 11,600 students and their teachers to one of three types of classes: (1) small classes (13–17 students); (2) regular-size classes (22–25 students); and (3) regular-size classes with a full-time teacher's aide. Krueger (1999) evaluated the results. He found that, on average, students who attended small classes performed four percentage points higher in their first year (about .20 of a standard deviation) and that their advantage

[6] Urquiola and Verhoogen studied the case of Chile, where a rule stipulates a maximum class size of 45 students. They found that the data followed a pattern close to that in Israel. Yet, they also found that most private schools tried to avoid enrolling the number of students that would require them to act upon a class-size cap—either by adjusting their fees or by not admitting that additional student. (The one exception consisted of schools in which parents valued small classes, which were also schools in which students tended to come from wealthier backgrounds.) Thus, the authors showed that the strategy used by Angrist and Lavy (1999) could lead to biased estimates of the impact of class size reductions, since students assigned to small classes are not comparable to those in regular classes.

increased by about one percentage point per year in subsequent years. These effects were larger for minority and poor students. Teacher's aides, however, had little effect on student achievement. While the magnitude of these effects has been the subject of much debate (Hanushek 2003), the size effect in the study remained a benchmark for other impact evaluations in education. The study is seen as demonstrating the potential of large class-size reductions in a developed country setting.

The main contribution of Krueger's study was that it used an RCT to evaluate a class-size reduction. Yet, what experiments such as Project STAR gain in internal validity they can lose in external validity (Box 5). Many have argued that while Krueger offered a rigorous answer to the question of whether Project STAR boosted learning, it is less clear whether his findings carry on to contexts other than Tennessee.

Box 5. What Is Internal and External Validity?

Internal validity refers to whether a study obtained an unbiased estimate of the effect of an intervention. External validity refers to whether the effect can be generalized to a population of interest. Take the example of a pre-service training program. Internal validity would be concerned with issues such as whether teachers were randomly assigned to the program (selection bias), whether these are novice teachers who would have improved even in the absence of the program (maturation), or whether teachers who left the program before the outcomes were measured were the ones who benefited the least (differential attrition). External validity would be concerned with whether the program or its main beneficiaries are sufficiently similar to others so that an evaluation of the latter would yield comparable results.

Hoxby (2000) contributed to the discussion of the external validity of Project STAR by evaluating the impact of class size in Connecticut using two rigorous methods. First, she compared students assigned to small or regular classes due to minimum/maximum rules using an RDD. Then she exploited changes in population trends. Given that such changes affect student-teacher ratios but are unlikely to influence student learning, Hoxby used such changes to conduct an Instrumental Variables Estimation (Box 6). Using these two methods, she found that class size did not have a statistically significant effect on student achievement. In fact, she was able to rule out even modest effects (.02–.04 of a standard deviation for a 10 percent reduction in class size). While Hoxby could not benefit from random assignment in Connecticut, her study is widely considered to be an important qualification for the implications of the STAR findings.

Box 6. What Is Instrumental Variables Estimation?

Instrumental Variables Estimation (IVE) is often used whenever there has been no random assignment of individuals to a treatment group and there is no discontinuity that determines who participated in a treatment group and who did not. It is based on the observation that variation in the outcomes of a program's participants is partly endogenous (i.e., related to the participants' characteristics) and partly exogenous (i.e., related to the program's characteristics). Thus, one way to evaluate such a program is to obtain data on a variable that would be related to the exogenous variation but not to the endogenous variation in outcomes. One can then estimate the effect of the program by using this variable. It is not easy to identify such a variable (the "instrument") and obtain data on it. Yet, there are a number of commonly used instruments, such as geographic features that influence treatment (e.g., distance between participants and a program), legal institutions that determine the intensity of treatment (e.g., different laws in neighboring similar states), and natural experiments (e.g., birth dates that determine whether a person has access to a treatment group or not), that can be utilized. Finally, random assignment to a treatment group can also be used as an instrument, since it is uncorrelated with individuals' characteristics but influences take-up.

The other question is whether the results of Project STAR apply to developing nations, where class sizes are much larger than in the United States. Duflo, Dupas, and Kremer (2007) evaluated a program in Busia and Teso, Kenya, that randomly assigned 210 primary schools to a program that (1) lowered student-teacher ratios on average from 80 to 46; (2) combined class-size reductions with improved incentives for teachers (by hiring local teachers on short-term contracts or increasing parental oversight); or (3) combined class-size reductions with tracking by students' achievement. Reducing student-teacher ratios, in the absence of other reforms, actually led to lower teacher effort and to small and statistically insignificant increases in test scores. However, combining class-size reductions with improved incentives led to significantly larger test scores increases (.19 of a standard deviation), and combining class-size reductions with tracking led to even larger increases (.25–.31 of a standard deviation). This study draws attention to the importance of policy interactions by suggesting that class-size reductions are unlikely to improve student achievement without incentives for teachers to change their behavior in the classroom.

These studies offer a nuanced answer to the question of whether class-size reductions improve student learning. The answer seems to be that lowering student-teacher ratios can boost achievement, but only insofar as it changes teachers' pedagogy and thus the educational experience of children in the classroom. This does not happen automatically and may depend on the extent to which improvements can be made in classroom management (i.e., initial class size) and on the mix of incentives to produce changes in teacher behavior.

Attracting the Best into Teaching

Economic theory suggests that individuals with outstanding academic and/or professional achievements are more likely to be effective teachers. On one hand, "human capital theory," developed by Becker (1964), Schultz (1963), and Mincer (1962), proposes that employers should hire individuals with more schooling because additional education increases productivity by raising workers' ability to understand and use new information. On the other hand, "market signaling theory," developed by Spence (1974) and Arrow (1963), among others, posits that, even if the education that potential employees acquire does not raise their productivity, high-skilled individuals will pursue it to send a signal to their employers that they possess skills that others lack and deserve to be paid more. While the theories differ on why employers should hire applicants with exceptional credentials, they both imply that attracting individuals with higher credentials will result in a higher-skilled teaching force.

Studies in economics suggest that luring top talent into teaching can also have a multiplier effect. Becker and Murphy (2000) argued that people's actions often influence their neighbor's incentives or information, and Montgomery (1991) showed that this occurs in various ways in labor markets. They call these interactions "social multiplier effects." These models give reason to believe that if teaching is able to attract qualified people, then competitive candidates who had not considered teaching might be drawn to it.

One way in which education systems can attract talented individuals is by offering competitive wages. Not surprisingly, economists have a lot to say about the role of pay. As Akerlof (1982) and Shapiro and Stiglitz (1984) noted, wages serve two functions: they allocate labor and provide incentives for employee effort. This, they argue, is the reason why firms seeking to attract outstanding candidates might want to pay their employees an "efficiency wage," i.e., more than what competitors offer. Lazear and Rosen (1981), however, argued that firms can avoid some of the problems inherent in efficiency wages by paying their employees commensurate with their performance. This idea has recently regained relevance in the education literature. Specifically, Barlevy and Neal (2011) showed that a "pay-for-percentile" approach, in which teachers are rewarded for improvements in the performance of their students relative to peers with similar achievement levels at the beginning of the school year, incentivizes teachers to allocate socially optimal levels of effort to all students.

Many studies have sought to understand the extent to which requirements for entry into the teaching profession motivate talented individuals to aspire to become teachers and identify those who will succeed as classroom instructors.

Table 2 summarizes four types of interventions that have been rigorously evaluated: (1) setting requirements for entry into teaching; (2) relaxing entry requirements for outstanding individuals; (3) rewarding advanced educational qualifications and experience; and (4) increasing teacher pay.

Rigorous studies on these topics are scarce because teachers are not randomly assigned to characteristics (e.g., race or gender) or qualifications (e.g., degrees or experience) and students are hardly ever randomly assigned to their teachers. The studies reviewed here are among the most rigorous in this line of research, but much remains to be done to obtain more conclusive answers to the questions that they pose. To minimize the risk of conflating selection and treatment effects, most research on teacher requirements uses a strategy known in econometrics as "fixed effects" (Box 7).

Box 7. What Are Fixed Effects?

Fixed Effects (FE) are commonly used in research that identifies correlations between two factors. Their main purpose is to account for factors that cannot be observed or for which there are no data. This is done by breaking up the sample into subgroups (e.g., subgroups of students, schools, and teachers) and comparing units in each subgroup among themselves over time. FE control for characteristics that are "fixed" (i.e., those that do not change) over time, but they have no way to account for time-varying factors. Unlike the other empirical strategies reviewed in this paper, FE, regardless of their sophistication, generally cannot conclusively establish cause-and-effect relations. Rather, they are used to increase the reliability of correlations.

Setting Requirements for Entry into Teaching

Virtually all education systems require individuals to meet some qualifications to become a teacher. These requirements are intended to set minimum standards for the profession, but in many education systems, traditional requirements have been called into question for being poor predictors of teacher effectiveness and for making entry into teaching too costly—particularly for top candidates who already face a high opportunity cost for going into teaching.

Most research on this topic has focused on whether certified teachers have students who perform, on average, better than those of uncertified teachers. The evidence suggests that the predictive effect of certification is usually small.

Kane, Rockoff, and Staiger (2006) used data from the 1998–99 to the 2004–05 school years to examine the relationship between student achievement and teacher certification for reading and math teachers in grades four to eight in New York City. This "panel data" allowed the authors to observe a teacher's effectiveness within the same school, grade,

and year over time and use combinations of FE to account for factors that confound the effect of certification. The effects of teacher certification were, at best, small. In fact, in math, students assigned to teachers without a certification performed, on average, no differently than peers assigned to traditionally certified teachers. Yet, these results differed somewhat according to the group of uncertified teachers to which traditionally certified teachers were compared.

A potential explanation for the lack of effects of certification in New York City may be that certification is a better proxy for teacher quality at some levels and in some subjects than in others, or that the quality of teacher certification in places such as New York is not stringent enough to predict teacher effectiveness. This would imply that teacher certification should be made more stringent rather than abandoned altogether.

This is what a study by Clotfelter, Ladd, and Vigdor (2007) in North Carolina suggested. The authors measured the predictive effect of a number of teacher qualifications (including certification) on student achievement at the high school level from 1999–2002 using permutations of student, school, subject, and year FE. Consistent with the hypothesis that certification might play a more important role in teacher effectiveness at the high school level, Clotfelter and his collaborators found that having a teacher with an alternative certification *reduces* student achievement by .06 of a standard deviation, and that having a teacher with a certification that is neither traditional nor alternative negatively affects student achievement by .05 of a standard deviation. The authors also found support for the case for higher certification standards by detecting small but positive effects of having a teacher with a National Board of Professional Teaching Standards (NBPTS) certification, a national certification program widely believed to have higher standards than traditional certification programs in the United States. NBPTS requires teachers to submit classroom videos and write essays and is both labor- and time-intensive.

The mixed evidence on certification led governments to commission an RCT on whether certification is worth the investment. Cantrell et al. (2008) compared the performance of classrooms of elementary students in Los Angeles in 2003 randomly assigned to NBPTS applicants and to non-applicants. Students of NBPTS-certified teachers scored .22 of a standard deviation above those of teachers who applied but were not certified. The students of successful NBPTS applicants also achieved greater learning gains than those of unsuccessful applicants. However, NBPTS-certified teachers were not more effective than teachers who did not apply to become NBPTS-certified. This suggests that while NBPTS might be successful at identifying the most effective teachers among its applicants, it may not be attracting the most effective applicants.

Another entry requirement that education systems are increasingly resorting to is exams, which typically assess teachers' subject-matter knowledge or field-specific pedagogical knowledge. The evidence on the usefulness of these tests is still evolving, but it suggests that they show some promise for identifying individuals who will be successful classroom instructors.

In Peru, Meltzer and Woessmann (2010) took advantage of the fact that students and teachers were tested in the same year on two subjects to determine whether the same student taught by the same teacher in two different subjects performs better in one of those two subjects if the teacher's knowledge is relatively better in that subject. The authors used FE to focus on the within-teacher, within-student variation in student outcomes while controlling for the fixed characteristics of students, teachers, and subjects. They found suggestive evidence that a teacher's subject-matter knowledge affects student achievement: a 1 standard deviation increase in teacher test scores raised student test scores by .10 standard deviation units. This means that if a student switched from having a teacher at the 5th percentile of the distribution of subject-matter knowledge to one at the 50th percentile, the student's achievement would increase by .17 of a standard deviation by the end of the year.

The results of a recent U.S. study differed considerably from those in Peru. Cantrell and Kane (2013) used correlations to examine whether the Content Knowledge for Teaching (CKT) tests, which assess teachers' understanding of how students acquire and understand subject-specific skills in math and reading, correlated with measures of teachers' value-added scores in these subjects in elementary and middle school in six U.S. districts: Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City; and Memphis, Tennessee. To account for the effect of teacher assignment to students, the authors used teachers' CKT scores to predict their average performance on a series of indicators across two different sections. Cantrell and Kane found CKT tests were correlated with subject-specific classroom observation protocols. Yet, they were not correlated with teachers' value-added scores, regardless of whether these were measured using state tests or "audit" tests that measure more complex skills.

Studies on entry requirements thus seem to find slim support for the effectiveness of traditional requirements (e.g., certification) in developed countries, and encouraging results for the use of alternative requirements (e.g., entry exams) in developing countries. Yet, important questions remain about the generalizability and robustness of these findings.

Relaxing Entry Requirements for Outstanding Individuals

It is perhaps this lack of clarity on the ideal set of teacher entry requirements that has prompted some education systems to relax traditional requirements for talented individuals seeking entry into the profession, in hopes of making teaching an attractive career choice for them.

One program seeking to lure talented youth into teaching in the United States is Teach for America (TFA), which recruits outstanding college graduates who majored in a wide variety of fields to teach in high-need schools for two years. TFA's dual mission is to raise the quality of instruction and at the same time serve as a transformative experience for its corps members so that, after their two-year commitment, they become champions of education reform in whatever field or sector they choose to work. Since its founding in 1989, TFA has inspired replicas in 28 other countries.

The most rigorous study on TFA's impact was conducted in 2001 in grades 1–5 in six of the 15 regions where the program places its members (Baltimore, Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta). Decker, Mayer, and Glazerman (2004) randomly assigned students to TFA or control teachers at the same school and grade and compared their achievement after one year. Control teachers included all those to whom students would have been assigned in the absence of TFA teachers (who were neither necessarily certified nor experienced). In fact, TFA teachers were more likely to have attended competitive colleges and to have acquired their education degree, but less likely to have education-specific training and student-teaching experience than control teachers. Yet, overall, students of TFA teachers performed .15 of a standard deviation higher than those of control teachers and .26 higher than those of novice control teachers (with 1–3 years of experience). There was no difference, however, in reading.

Recent evidence suggests that the effect of such programs might be greater in developing countries, where average teacher effectiveness is lower. Alfonso, Santiago, and Bassi (2011) evaluated the impact of an adaptation of TFA in Chile called *Enseña Chile* ("Teach Chile" or eCh) during the 2009–10 and 2010–11 school years. Alfonso and her collaborators were not able to randomly assign eCh teachers to students, so they used Propensity Score Matching (Box 8). Like FE, this technique is an improvement over correlations that try to account for observable measures that might bias estimates of the impact of a particular policy, but it cannot distinguish selection from treatment effects. After a year, students in schools that received an eCh teacher scored .22–.51 of a standard deviation higher in Spanish and .17–.43 higher in math. After two years, students scored .75 of a standard deviation higher in Spanish and .33 higher in math. These estimates seem too large for an educational intervention and hint at the existence of selection bias,

but they offer a good motivation to evaluate these types of programs more rigorously in developing countries.

Box 8. What Is Propensity Score Matching?

Propensity Score Matching (PSM) is an empirical strategy often employed when it is not possible to use any of the rigorous evaluation methods mentioned earlier in this paper. It cannot be used to claim that a given program caused a particular impact. Rather, it is simply a way to mitigate the evaluation problem mentioned earlier. The approach is simple. Researchers first identify the factors that made individuals more prone to self-select into a treatment group and then create a comparison group by matching each "treated" individual in the treatment group with one (or more) "untreated" individuals with comparable observable characteristics. In this way, researchers can compare the outcomes of both sets of individuals after the treatment as an effect of that treatment. The downside of these types of studies is that researchers can only match individuals on what they can observe. If there are unobservable characteristics that made some individuals more prone to select into a program, there is no way for researchers to know that, so their estimate will be biased.

These studies suggest that while it is not known which entry requirements are most effective at predicting teacher effectiveness, neither is there any compelling case for doing away with such requirements altogether. A better understanding is needed of which entry requirements (if any) serve to ensure a minimum "floor" of teaching quality and whether (and if so, how) one can more broadly use entry requirements without relying on other ways of identifying effective teachers that may be more costly but also more effective.

These studies, however, focus on teachers entering the profession through highly selective alternative pathways. The question thus arises of whether the bulk of alternatively certified teachers, who enter the profession through far less selective pathways, will be less effective than regular teachers.

Exactly this question prompted Constantine et al. (2009) to compare the impact of traditionally and alternatively certified teachers on student achievement in the United States in 2003 by randomly assigning students in the same school and grade to one or the other of the two types of teachers. The study was conducted with 2,600 students in 63 schools in 20 school districts. The authors found that the difference between traditionally and alternatively certified teachers might be less clear-cut than policy debates suggest because (1) both varied widely in the hours of pre-service training they had undertaken; (2) on average they did not differ in their scores on college entrance exams, the selectivity of their college, or their educational attainment; (3) neither group was more effective on average in raising student achievement; and (4) neither the amount nor content of the courses that alternatively certified teachers took made a difference in their

effectiveness. These findings suggest that alternatively certified teachers who do not enter the profession through selective programs are not all that different from regular teachers.

Rewarding Advanced Educational Qualifications and Experience

Many education systems pay their teachers more for graduate studies or years of experience. The rationales for these pay differentials are to attract better educated individuals and to retain seasoned instructors. The implicit assumption in these policies is that these are more effective teachers, so research on this topic assesses whether this is the case.

Hanushek et al. (2005) combined the Texas Schools Microdata Panel with data on classroom assignment from an anonymous large urban school district in the state to determine whether certain teacher characteristics, including having a master's degree or experience, predicted higher student achievement. The study used school and student FE. The authors found that students of teachers with master's degrees performed no differently, on average, than those whose teachers did not have such a degree. They also found that first-year teachers had much lower performance on average than other teachers (the achievement of their students was .12–.16 of a standard deviation lower), but that additional years of teaching experience had little impact on student achievement.

Until recently, the external validity on traditional teacher credentials in the United States was also problematic, since most studies focused on a single school district. In 2012, Kane and Staiger published a report that examined whether teachers with more experience or master's degrees have students who perform better on math and reading tests. As in the Cantrell and Kane (2013) study mentioned earlier, the Kane and Staiger (2012) study looked at school districts in Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City; and Memphis, Tennessee.

In the absence of random assignment of teachers to students, the authors tried to account for factors that confound the effect of credentials by using the average performance of students in two different sections taught by the same teacher as their outcome variable. They found that students of teachers with 12 or more years of experience achieved gains that were .01 of a standard deviation higher in math and .02 higher in reading than those of teachers with fewer than three years of experience. Similarly, they found that students of teachers with a master's degree saw gains that were .03 of a standard deviation higher in math and.02 lower in reading.[7]

[7] Given the existing concerns in the United States with the quality of state tests, the authors also examined how well teaching experience and master's degrees predicted achievement gains in "audit" tests of math

The authors benchmark the predictive power of these traditional teacher credentials with a measure of teacher quality that combines a teacher's value added, student survey results, and classroom observation scores from another section (i.e., another group of students). They find that this combined measure does a much better job at predicting student achievement gains: students with a teacher who is in the top 25 percent in this combined measure make gains that are .22 of a standard deviation higher in math and .07 higher in reading in the state tests than their peers with a teacher in the bottom 25 percent in the combined measure.[8]

Together, these studies offer compelling evidence that educational qualifications and experience are very weak proxies for teacher effectiveness and that more thoughtful measures can do a better job at predicting teaching quality. The findings suggest that education systems would be well advised not to rely on pay differentials for educational qualifications and experience to attract better teachers.

Increasing Teacher Pay

Finally, a policy commonly touted as important to attract talented individuals into teaching is to raise teachers' salaries. The effectiveness of high salaries in attracting better teachers is difficult to evaluate because teachers are not randomly assigned to salaries. Hoxby and Leigh (2004), however, noted that unionization compressed wages. They used laws that either helped or hindered teachers' unionization to find out whether the decrease in teacher aptitude in the United States over recent decades can be explained by an increase of opportunities for women outside of the profession or by compressed wages within the profession. This allowed them to use legislation as an "instrument" for pay compression in an IVE.

Pay compression increased the share of the lowest-aptitude female college graduates who became teachers by 9 percentage points and decreased that of the highest-aptitude female college graduates who became teachers by 12 percentage points. Hoxby and Leigh also found that improvements in pay parity reduced the share of women who taught by 3.2 percentage points for the highest aptitude group and 0 for the lowest aptitude groups. These findings suggest that wage compression played a key role in the decline of aptitude

---

and reading. They found that the predictive power of master's degrees was slightly higher with these tests: students of teachers with master's degrees made gains that were .05 of a standard deviation higher in math and .03 higher in reading than their peers with teachers without master's degrees. The predictive effect of teaching experience with the audit tests, however, was more mixed: students of teachers with 12 or more years of experience made gains that were .05 of a standard deviation lower in math and .03 higher in reading.

[8] The results were similar using the audit tests: the first group of students made gains that were .13 of a standard deviation higher in math and .13 higher in reading than the second group.

in the teaching profession. The lesson from this study is not to universally increase wages, but rather to pay better teachers more in order to attract and retain them.

Preparing Teachers with Useful Training and Experience

Theoretical and empirical work in economics has shown that high-skilled workers value training more than low-skilled workers. Salop and Salop (1976) and later Autor (2001) called this phenomenon the "complementarity" between skills and training. Thus, offering training can promote positive self-selection on skills that employers cannot typically observe. This is of interest for teacher labor markets, in which there are very few observable skills that make an individual more likely to be an effective teacher.

If human capital theory is correct and training enables individuals to acquire new skills, education systems should subsidize pre-service teacher training. As Becker's (1964) model suggested, however, this should be true only if the skills that are necessary to be an effective teacher are specific, rather than general. He argued that since training in general skills makes employees more likely to succeed not only at the firm or industry that provides them with the training, but also at others, firms are unwilling to invest in training in general skills. Yet, the incentives for governments to subsidize pre-service teacher training, and for potential entrants to take it, are strong because this training almost invariably leads to a job, avoiding many of the so-called "contracting" problems with specific-skills provision present in private firms (Prendergast 1993).

But as Autor (2001) noted, even if education is pure signaling, training could be a worthy investment as a screening mechanism. Training can elicit otherwise unobservable information about the skills of potential employees, which firms can use to make hiring decisions. This is particularly relevant for teacher policy, given that teaching seems to fit the profile of the occupations described by Terviö (2003) in which initial experience on the job is a fairly accurate predictor of subsequent performance.

Although pre-service training is one of the most common ways in which governments try to increase the capacity of teachers, there is little rigorous research on its impact. This is mostly due to self-selection into training, which makes selection and treatment effects hard to disentangle. The very small body of rigorous research on this issue is composed mainly of empirical work assessing changes to traditional forms of training or interventions early in teachers' careers that might complement pre-service training.

Table 3 summarizes two types of interventions that have been rigorously evaluated: (1) assigning mentors to new teachers; and (2) including a practice component in teacher preparation.

Assigning Mentors to New Teachers

One consistent finding in the teacher effectiveness literature is that teachers are less successful in their first few years on the job. This has led some education systems to assign mentors or coaches to new teachers to accelerate their learning in what are often called "induction" programs. The evidence suggests that such interventions can improve student learning, but that the quality and dosage of mentoring matters in ways that are not easy to anticipate when designing these programs.

Rockoff (2008) evaluated a mentoring program in New York City in 2004. In order to measure the impact of the program, he compared the performance of teachers with prior experience (who were not targeted by the program) to that of new teachers (who were targeted by the program, and therefore were more likely to be assigned to a mentor). This strategy is known as Differences-in-Differences Analysis (Box 9). Rockoff found that novice teachers assigned to mentors were 4.5 percent more likely to complete their first year than teachers without mentors, but just as likely (or, put differently, no more likely) to stay at their schools or raise student achievement.

Box 9. What Is Differences-in-Differences Analysis?

Differences-in-Differences Analysis (DDA) is often used when other rigorous empirical strategies cannot be conducted. It is simple in that it estimates the impact of a program by obtaining the difference in outcomes for a group of individuals who received a treatment before and after the treatment (called the "first difference"), obtaining the difference in outcomes for a similar group that was not exposed to treatment (called the "second difference"), and then subtracting the latter from the former. While one may try to calculate a program's impact by using only the first difference, it is likely that other things happened at the same time of the treatment that would confound its effect. Therefore, using the second difference allows researchers to account for any other events that would affect all individuals (known as "secular trends"). If nothing else changed between the two groups at the time of the treatment, and the composition of the two groups remained the same during the treatment, a DDA offers an unbiased estimate of a program's impact.

Rockoff also used an IVE to look at whether experience, the number of hours spent by mentors with each teacher, and teachers' perceptions of mentors make a difference. He found teachers were less likely to leave a school when they had mentors with prior experience at that school, which he takes to suggest that a key part of mentoring may be the provision of school-specific knowledge. He also found that the number of hours mentors spent with a teacher had a positive impact on student achievement: for every 10 hours of mentoring, students of mentored teachers scored .05 of a standard deviation higher in math and .04 higher in reading.

The importance of the quality and dosage of mentoring programs led Glazerman et al. (2010) to evaluate the impact of such programs in 17 urban districts in the United States. They randomly assigned 418 elementary schools to one of two categories: (1) schools with a "comprehensive" teacher induction program, in which beginning teachers received monthly professional training, opportunities to observe veteran teachers, and a full-time mentor who had access to updated training and materials; or (2) schools whose beginning teachers received the district's usual, less-intensive induction. In 10 of the 17 districts, the services were offered to treatment schools for one year and in the remaining seven districts, services were offered for two years. However, districts were not randomly assigned to the length of the services, so findings for both sets of districts should be interpreted separately.

Ninety percent of teachers in districts assigned to comprehensive induction had a mentor in their first year, compared to 72 percent of teachers in districts with business-as-usual induction. Interestingly, however, teachers who received comprehensive induction were no more effective than their control peers, as measured by classroom observations and student achievement. Further, the achievement of students with teachers receiving two years of comprehensive induction showed no improvements during the teachers' first two years on the job, but by the teachers' third year their students performed .11 of a standard deviation better in reading and .20 better in math than students of control teachers. Neither induction program had a positive effect on teacher retention or satisfaction. This complements the Rockoff study by providing a sobering view of mentoring in which improving quality and dosage is not just a matter of increasing treatment intensity.

Including a Practice Component in Teacher Preparation

Research suggests teacher performance improves considerably during the first few years on the job. Some have attributed this to the importance of first-hand experience in mastering classroom practices, leading some education systems to incorporate "clinical" (practice-based) components in their teacher education programs.

One such initiative is the Boston Teacher Residency (BTR) program, which as stated in its literature recruits "talented college graduates, career changers and community members of all ages" to take an intensive summer training program and a one-year "residency" in a classroom in the Boston public schools. During that year, participants work with a mentor four days a week and attend seminars once a week. Upon graduation, they receive a master's degree and a "dual" license, including one in special education. They are placed in schools with alumni and receive one-on-one induction and professional development. Chicago and Denver adopted similar programs in 2002 and 2004, respectively.

Papay et al. (2011) used a combination of FE to conduct the first independent evaluation of BTR for the school years from 2001–10. BTR graduates were more racially diverse than other Boston public school novices, more likely to teach math and science, and more likely to remain in the profession after their fifth year. BTR graduates were no more effective in the classroom, as measured by their value added in student achievement tests, but they improved faster after their second year, and by their fifth year they outperformed even veteran teachers. These findings suggest that clinical experience might offer teachers tools to improve their performance faster on the job.

Matching Teachers' Skills with Students' Needs

From the perspective of economics, job matching minimizes turnover, which is costly to employers and employees, and maximizes what is called "allocative efficiency" (Hicks 1939; Kaldor 1939). Economic theory offers a framework to examine why a school that is a good match for a teacher can make that teacher more productive. In "search theory" or "matching theory" (Pissarides 1979), economists argue that it is useful to think of jobs as no different from other goods that are consumed: when choosing a job, individuals consider their options and select the one based on their preferences and the constraints that they face.

Yet, there is an important distinction between the consumption of, say, typical household goods and the choice of a job, a distinction that draws on the work of Nelson (1970). Typical household goods are "search goods" that are easy to evaluate before acquiring them (Lucas and Prescott 1974). In the case of search goods, mismatches between people's preferences and constraints and the good that best addresses them are due to imperfect information and can be dealt with by improving information provision. Jobs, by contrast, are "experience goods"—it is not easy for people to evaluate whether a particular job is the most productive fit for them without taking and holding the job for a while (Johnson 1978). This is why economists find that when people switch jobs, they earn more (i.e., because they typically switch to a job that has made them more productive) and why job-switching declines with tenure and experience (i.e., because with time people eventually find the "right" job that best fits them) (Altonji and Shakotko 1987; Bartel and Borjas 1981; Jovanovic 1979a and b; Topel and Ward 1992).

Economists also justify job-matching on the grounds of "job market segmentation." As Reich, Gordon, and Edwards (1973) argued, there is a wealth of empirical work suggesting that workers' skills are not the only determinant of wages, and that workers with similar skill levels get paid different wages in different labor markets. Proponents of segmented labor markets argue that the "law of one price"—which dictates that the same good cannot be sold in two markets for different prices because it would create arbitrage

opportunities that would quickly equalize the price in both markets—does not hold in labor markets. If they are right, then job-matching could be instrumental in correcting these market failures. The precise mechanisms for how matching should take place, however, have been the subject of considerable debate in the field (Lang and Dickens 1987).

Table 4 summarizes three types of interventions that have been rigorously evaluated: (1) offering bonuses for teachers to work in high-need schools; (2) offering bonuses for teachers to teach critical shortage subjects; and (3) improving working conditions.

Offering Bonuses for Teachers to Work in Schools with the Greatest Need

Steele, Murnane, and Willett (2010) evaluated an experiment in 2000–02 in California to offer $20,000 signing bonuses to attract talented, novice teachers to low-performing schools and retain them for at least four years. They took advantage of the fact that this policy was enacted only for two years to conduct an IVE. Bonus recipients would have been less likely to teach in low-performing schools than observably comparable counterparts had the bonus not existed; however, because of the bonus, the probability that its recipients taught in low-performing schools increased by 28 percentage points, and 75 percent of both bonus and nonbonus recipients who began working in low-performing schools stayed in such schools for at least four years.

Offering Bonuses for Teachers to Teach Critical Shortage Subjects

Clotfelter et al. (2008) assessed a program in North Carolina in 2001 that offered a $1,800 annual bonus to retain teachers certified in math, science, and special education in high-poverty or failing schools. They compared the turnover of eligible and ineligible teachers before and after the program to conduct a DDA. The bonus increased year-on-year retention by 10–13 percent. This effect, however, was mainly driven by math teachers: those who received the bonus were 18 percent less likely to leave their schools than those who did not. Science and special education teachers who received bonuses were no less likely to leave than those who did not. The effects were concentrated in middle schools: teachers at this level were 27 percent less likely to leave if they had received a bonus, but bonus recipients in high schools showed no difference in attrition. Considering that in the first year of implementation the bonus was only 4–5 percent of an average teacher's salary, even modest financial incentives seem to influence teachers' decisions to stay in hard-to-staff schools.

Improving Working Conditions

A number of recent papers shed light on the effect of improvements in working conditions on teachers' productivity.

Jackson (2013) used data on third through fifth graders in North Carolina in 1995–2006 and employed year, school, and teacher FE to estimate the changes in teachers' productivity (as measured by their value-added scores) as a result of transferring from one school to another. He found that teachers who switched schools were more effective—by .09 of a standard deviation in math and .07 in reading—than before they moved, which he interpreted to be suggestive of "match effects." These effects accounted for 10–40 percent of what is typically estimated as teacher quality, suggesting a sizable portion is not portable. In fact, Jackson also found certain types of teachers achieved much better outcomes at certain types of schools, such that an optimal matching of teachers to schools could raise outcomes for all students.

While this study suggests matching teachers to schools is important, it does not indicate which factors would raise teachers' productivity. Jackson and Bruegman (2009), however, conducted a study that indicates peer learning (i.e., teachers learning from each other) is an important mechanism through which match quality improves teacher productivity. The authors used the same database as Jackson (2013) and employed student, teacher, school, and year FE. Teachers who experienced a 1 standard deviation improvement in observable peer characteristics (e.g., experience and licensure) were associated with a .10 of a standard deviation increase in the scores of their students in math and reading. Teachers who experienced a 1 standard deviation increase in the value-added scores of their peers had students who performed .05 of a standard deviation better in math and reading.

While there are only a few studies on matching teachers to schools, the message from these studies is clear: the match between teachers and their schools is an important factor in teacher retention and effectiveness, and monetary incentives have shown promising results to allocate teachers where they are most needed. A key gap in the literature is the lack of rigorous studies of these initiatives in developing countries, where they are most needed.

Leading Teachers with Strong Principals

Although economists do not have a lot to say about the importance of administrators, they traditionally justify the assignment of workers to different jobs within a firm on the grounds that each worker has a "comparative advantage"—i.e., he or she is best at doing

a certain task and is therefore better off specializing in that task (Sattinger 1975; Rosen 1978). Thus, economists argue that those teachers with outstanding management skills are best suited for principal posts and those with outstanding classroom effectiveness skills would be best employed as teachers or coaches. Several extensions of this approach also note that whenever jobs must be staffed by a single worker, high-ability workers must be assigned to jobs that value ability more highly (Rosen 1978).

Research on policies affecting principals' work is very recent and there are only a handful of rigorous studies. Yet, this emerging body of evidence is suggestive of the types of policies that might be more conducive to the effectiveness of principals.

Table 5 summarizes two interventions that have been rigorously evaluated: (1) hiring more effective principals; (2) setting requirements for principal positions; and (3) granting principals more authority over staffing decisions.

Hiring More Effective Principals

Research on the effectiveness of principals is new and still evolving, but there are some studies on whether (and if so, how) one can estimate a principal's value added based on the achievement of students at his/her school.

Coelli and Green (2012) exploited the fact that principals in British Columbia, Canada, are rotated across schools by districts to observe the performance of different principals at the same school. Using school, neighborhood, and peer group FE, they found a sizable heterogeneity in school principal quality affecting the English test scores of 12th grade students. Getting a principal who is 1 standard deviation better in the principal-effects distribution implied that graduation rates were 2.6 percentage points higher (or, roughly, .33 of a standard deviation of the cross-school graduation rate distribution) and English exam scores were 2.5 percentage points higher (roughly equivalent to 1 standard deviation).

Branch, Hanushek, and Rivkin (2012) introduced longitudinal data to the estimation of principal effects. They used data from Texas for 1995–2001 to estimate the productivity of principals using principal, school, and year FE. They found the annual impact of having an effective principal was .05–.21 of a standard deviation, depending on the method used. They also explored a potential mechanism through which effective principals can improve student achievement: removing ineffective teachers. Consistent with this hypothesis, they found teachers who left schools with the most successful principals were much more likely to have been among the least effective teachers in that school than teachers leaving schools run by less successful principals. The authors also

looked into whether principals who switched schools were more effective than those who stayed, and found no relationship between principal quality and probability of transfer. Principals in the lowest performance quartile were least likely to remain in their position and most likely to leave public schools entirely. By contrast, principals in the highest performance quartile were more likely to remain in their position and least likely to leave public schools entirely than their peers in the lowest quartile, but were more likely to leave their position and less likely to leave public schools than those in the middle two quartiles.

As Grissom, Kalogrides, and Loeb (2011) argued, however, different models produce estimates that span a wide range. The authors used data from Miami-Dade County, Florida for 2003–10 in grades 3–10 to show that, while some models identified principal effects as large as .15 of a standard deviation in math and .11 in reading, others found effects as low as .02 of a standard deviation in both subjects for the same principals. Interestingly, the most conceptually unappealing models, which attributed too large a share of school effects to principals, aligned more closely with nontest measures than approaches that more convincingly separated the effect of the principal from the effects of other school inputs. The main take-away from the study is that while there is evidence that principal quality matters, precisely measuring the effectiveness of principals will require further research.

Setting Requirements for Becoming a Principal

While there is a consensus that effective principals matter, how to identify effective principals is less clear. There are two studies on the relationship between principal characteristics and effectiveness, and both were done in the United States, which raises questions about their external validity.

Clark, Martorell, and Rockoff (2009) explored this question using employment and student achievement data and school FE from New York City for 1998–2006. They found little evidence of a relationship between school performance and the selectivity of a principal's undergraduate or graduate institution. They also found little relationship between performance and a principal's prior work experience. However, among very inexperienced principals, school performance is higher among those that were previously assistant principals at their current school. There was also a positive relationship between principal experience and school performance: principals with three years of experience had students whose math scores were .04 of a standard deviation higher and principals with five or more years of experience had students whose math scores were .06 of a standard deviation higher than those of students with principals in their first year. Finally,

there was mixed evidence on the relationship between principal training and professional development programs and school performance.

Grissom and Loeb (2011) used data for 2008 from Miami-Dade County, Florida, including surveys of principals and assistant principals, grades assigned by parents to their children's schools, and administrative data (e.g., school performance ratings and state tests). Principals with stronger organizational skills were in charge of schools with greater gains in student achievement: for all schools, a 1 standard deviation increase in principals' self-rating in "organization management" was associated with a .12 point increase in school accountability performance in the original scale, or with .10 of a standard deviation. Principals' self-rating on their organizational skills was also related to parental ratings of their children's schools, but was not consistent with teacher satisfaction. Finally, ratings given by assistant principals of principals' organizational skills were also associated with math and reading gains.

Granting Principals More Authority over Staffing Decisions

The most rigorous study on granting principals greater authority over staffing decisions was conducted in New York City by Rockoff et al. (2011). They assessed the impact of a pilot program in the 2007–08 school year in which principals were selected from a group of volunteers to receive "value-added" performance measures for teachers at their schools and training on the methods used to construct these measures. Because principals were selected at random, the authors could carry out an RCT. Principals' prior beliefs about teacher effectiveness matched value-added measures and they updated their views when they received new performance data. In fact, after data were provided, principals were more likely to keep teachers with higher value-added scores and less likely to keep teachers with lower scores, leading to small improvements in productivity at their schools. Yet, importantly, the provision of value-added measures did not "crowd out" information about teacher effectiveness that principals collected through classroom observations. These results suggest that principals hold accurate views about the effectiveness of their teachers and they can use performance data wisely to make staffing decisions.

A program in Chicago in 2004 gave principals authority to use an expedited process to dismiss probationary teachers (i.e., those with less than five years of experience) for any reason. Jacob (2010a, b) used a DDA to assess this reform by comparing changes in teacher absences before and after the policy for probationary and tenured teachers. He found that the policy reduced absences among probationary teachers by 10 percent and reduced absences among teachers with 15 or more absences by 20 percent. He also found evidence that the policy increased student achievement at the elementary level.

Interestingly, these changes were only partly explained by the reduction in teacher absenteeism, suggesting the policy affected teacher behavior in class. Finally, principals considered teacher absences and value-added measures in choosing whom to dismiss. This suggests that, when given flexibility to dismiss teachers, principals can make sound decisions.

These studies, while tentative, offer clear implications. Principals with management experience are most effective. In fact, principals (at least in developed countries) seem ready to take on additional authority and responsibility to improve their schools' performance.

Monitoring Teaching and Learning

Economists see monitoring as useful for two purposes: keeping employees motivated and identifying talented employees for promotion. Discussions of performance monitoring in economics have been typically framed within a principal-agent problem. Originally, it was a problem of inducing worker effort, in which the employer knew what to ask employees to do in a context of asymmetric information (Ross 1973). Then, economists acknowledged that employers do not always know how their employees could be most productive, so it became a problem of first figuring out adequate employee behavior and then inducing it (Baker 1992). More recently, the field has noted that for some jobs it is best to monitor performance through output-based rather than input-based measures (Prendergast 2002). This might be of relevance to teaching, where there is considerable debate over the teacher actions that work best to raise student learning.

Economists also see performance monitoring as useful for hiring. Whenever credentials or screening devices are poor predictors of productivity, firms can minimize hiring costs while maximizing worker productivity by creating "internal labor markets"—that is, by hiring employees for lower-ranked positions, observing their performance, and deciding which should stay where they are or be promoted (Williamson 1975; Osterman 1984). The way in which many education systems promote their teachers is deeply embedded in this approach, which suggests there is a strong belief that experience is the best way for teachers to become principals and for principals to obtain school-specific knowledge.

Table 6 summarizes four types of interventions that have been rigorously evaluated: (1) increasing parental and community involvement in school affairs; (2) grading and/or ranking schools based on student achievement; (3) monitoring teacher effort; and (4) monitoring teacher performance.

Increasing Community and Parental Involvement in School Affairs

Initiatives seeking to increase parental and community involvement in education have been rigorously evaluated in several contexts. In Uttar Pradesh, India, villages were randomly assigned to a control group or to one of the following: (1) an initiative that explained the role of village committees to their members; (2) a program that trained volunteers to administer and report on reading, writing, and math assessments; or (3) a project that trained local volunteers to provide remedial reading classes after school. None of these interventions affected community involvement in public schools or student achievement. However, the third intervention had a positive effect on youth involvement, getting young people to volunteer to teach, and subsequently improved students' reading test scores by about .02 of a standard deviation. The authors interpreted their findings as suggesting that individuals face considerable constraints to improving the quality of their public schools, even when they care about it and are willing to get involved.

However, evidence from a similar experiment in India suggests that the format and context of these initiatives matter. Pandey, Goyal, and Sundararaman (2009) evaluated an initiative in Karnataka, Madhya Pradesh, and Uttar Pradesh in 2006 that randomly assigned villages to either a campaign that disseminated information to communities about their state-mandated roles and responsibilities in school management or to no intervention. The information campaign increased the number of meetings of the village education committees and member participation by 25 percent in Uttar Pradesh, and it increased the percentage of parents who talked to teachers about the quality of education in Madhya Pradesh. The campaign also raised teacher attendance by 11 percent in Uttar Pradesh and classroom activity by 30 percent in Madhya Pradesh. However, the campaign led to only modest increases in reading achievement. These findings suggest that while information campaigns can raise teacher and community effort, their prospects to raise learning are limited.

These results seem to echo those of a similar study in a developed country. Avvisati et al. (2008) used an RCT to evaluate an initiative that convened meetings with parents to discuss helping their children with homework, grades on report cards, and other school issues. The aim was to increase parental involvement at school and at home. The children of treated families developed more positive behavior and attitudes in school and had fewer literacy problems. In fact, there were large spillovers on the classmates of treated families. These findings lend further support to the promise of interventions seeking to increase community participation in order to influence the behavior of families, although they beg the question of the extent to which these changes raise learning.

Grading or Ranking Schools Based on Student Achievement

Another initiative that has been subject to rigorous evaluation is the grading or ranking of schools according to student achievement. The only RCT of this policy, conducted in Punjab, Pakistan in 2004, evaluated the provision of school- and student-level report cards. Andrabi, Das, and Khwaja (2009) found villages in which report cards were provided had test scores that were .10 of a standard deviation higher than those in which none were distributed, but that the learning gains were concentrated in low-performing schools. In fact, report cards produced .34 of a standard deviation increase in test scores in schools that initially scored below the median baseline, but had no effects on those initially above the median. Finally, report cards decreased private school fees by 18 percent. Given that few students switched schools, providing report cards generated competitive pressures on all schools to increase their quality to make it commensurate with the price that they charge.

A recent study sheds some light on other often overlooked consequences of efforts to rank or grade schools. When the state of Florida suddenly changed the way it graded its schools in 2002, Feng, Figlio, and Sass (2010) took the opportunity to conduct a DDA. They found that teachers in schools that received a lower-than-expected grade were 11 percent more likely to leave their school, and those in schools that got a higher-than-expected grade were 2.3 percent less likely to leave. These effects were clearest in schools that not only received a lower-than-expected grade, but were downgraded to a failing grade. Teachers at these schools were in fact 42 percent more likely to leave their schools and 67 percent more likely to move to a school in the same district. Importantly, however, the positive and negative effects of this policy on teacher mobility left the composition of teachers at downgraded schools unchanged: teachers who remained behind increased their effort, raising student achievement by .05 of a standard deviation, and teachers leaving these schools were also on average more effective. These findings suggest that school accountability can sometimes affect teacher mobility without improving student learning.

A similar policy was instituted in New York City in 2007 to grade schools on an A-to-F scale and to tie these classifications to rewards and consequences, including possible school closure. An evaluation indicated that many of the findings about such policies depend on the specifics of the policies themselves. Rockoff and Turner (2010) used an RDD to assess the effect of school grades under the New York initiative that were arguably released too late in the school year to prompt any significant changes in school activities or personnel. However, they found that schools that received Ds or Fs raised math achievement and that those that received Fs also improved English scores. Additionally, they found that parental evaluations of school quality improved in schools

that received D or F ratings. These findings suggest school accountability policies that go beyond providing information may successfully spur improvements in school quality.

Again, however, there is evidence that the details of these types of interventions matter. Mizala and Urquiola (2007) evaluated an initiative in Chile that since 1996 has ranked schools according to their performance, adjusted for students' socioeconomic status. The program offered a monetary incentive to schools if they performed above a certain threshold. The authors took advantage of this threshold to assess the effects of the program using an RDD. The policy had no effect on the learning outcomes of bonus-recipient schools. This suggests that public information about school quality, even when attached to rewards and consequences, does not always prompt improvements in student achievement.

Monitoring Teacher Effort

Evidence on initiatives devised specifically to monitor teacher performance is much clearer than evidence on ranking schools. Duflo, Hanna, and Ryan (2010) evaluated an initiative in Rajasthan, India in 2003 that monitored teacher absenteeism on a daily basis using cameras, and that made teachers' salaries a function of their attendance by randomly assigning schools to either this treatment or no treatment. The intervention reduced teacher absenteeism from 42 to 23 percent. After a year, test scores in treatment schools were .17 of a standard deviation higher than in control schools, and grade completion increased. These findings indicate that initiatives to induce improvements in teacher effort that specifically target the behavior they try to change can achieve remarkable results.

Monitoring Teacher Performance

A recent study examined the extent to which teachers' effectiveness in one classroom (classroom A)—as measured by a number of indicators including a master's degree, experience, subject-specific pedagogical knowledge, NBPTS certification, principal ratings, student surveys, classroom observations, and value added—was predictive of their value added in another classroom (classroom B). In particular, the study found that teachers identified as effective through a composite of three measures (student surveys, classroom observations, and value added) in classroom A tended to also have high value-added scores in classroom B (Kane and Staiger 2010, 2012). In fact, Kane et al. (2013) found that a student-level standard deviation in the composite measure of teacher effectiveness on one year corresponded to roughly a student-level standard deviation in the value added for the following year, when comparing teachers teaching the same subject and grade in the same school. The studies used data for grades 4–8 in 2009–11 in

Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City; and Memphis, Tennessee.

Mihaly et al. (2013) compared four ways of weighting the three measures of effective teaching: (1) weighting them for maximum accuracy in predicting next-year gains on state tests (which resulted in an 81 percent weight on value added, 2 percent weight on classroom observations, and 17 percent weight on student surveys); (2) a 50 percent weight on value added and 25 percent each on both classroom observations and student surveys; (3) equal weights (33 percent) on all three components; and (4) a 50 percent weight on classroom observations. The first model maximized the predictive power of the composite, as measured by its correlation with next-year value added, while the third and fourth models maximized reliability (i.e., the proportion of variation in the scores on the composite reflecting consistent differences in practice between individual teachers). The authors did not recommend an "ideal" set of weights, but argued that the weights of the different measures should correspond to the main goal of the teacher evaluation system and its associated school and teacher accountability system.

Supporting Teachers to Improve Instruction

Building on the work of Greenwald (1986) and Acemoglu and Pischke (1998), it is noted here that when a firm invests in general skills training, it engages in "asymmetric learning"—that is, it learns more about the effect of training on the worker's productivity than its competitors learn. Thus, the firm's decision to invest in such training depends entirely on the effect that the training has on its productivity. These insights can provide a rationale for education systems to invest in training their teachers once they have entered the profession. In the case of education, the threat of turnover is mitigated by the fact that the education system as a whole bears the costs of training and that much of the competition for teachers is between schools, rather than between education and other fields.

Table 7 shows two types of interventions that have been rigorously evaluated: (1) providing or improving learning materials; and (2) providing in-service training.

Providing or Improving Learning Materials

Efforts to provide teachers with supplementary teaching materials have not been as successful as expected. Many initiatives have consisted of trying to incorporate computers in the classroom. Barrera-Osorio and Linden (2009) conducted an RCT to evaluate a program in Colombia in 2006 that sought to integrate computers, donated by the private sector, into the teaching of reading in public schools. The program had

virtually no effect on test scores or other outcomes, and the results were consistent across grade levels, subjects, and students' gender. Surveys administered to teachers suggested that the main reason behind the limited effect of the program was that few teachers actually incorporated the donated computers into their classroom practices. These findings indicate input-based policies are likely to have little impact on student learning if they do not change business-as-usual pedagogical processes.

Yet, not all of these interventions have been unsuccessful. Banerjee et al. (2007) used an RCT to evaluate a computer-assisted learning program in Vadodara, India, in 2002. Under the program, fourth graders shared a computer with a peer for two hours per week to play games solving math problems of age-appropriate difficulty. The program increased math scores by .35 of a standard deviation in the first year and by .47 of a standard deviation in the second year, and it was equally effective for all students. Interestingly, however, the program had no effect on students' reading scores, suggesting limited spillover effects. Further, effects seemed to fade shortly after the program ended. The difference between this program and the one in Colombia seems to be that it represented a meaningful change in classroom instruction.

Evidence from developed countries seems to support this conclusion. Rouse and Krueger (2004) used an RCT in 2001 to evaluate the impact of a reading instruction computer program on students with reading difficulties in an anonymous district in the United States. While the program improved some aspects of students' learning skills, these gains did not translate into a broader measure of language acquisition or into reading skills.

Together, the three evaluations reviewed here suggest that the impact of computer-assisted learning varies considerably, depending on the details of specific initiatives.

Other types of input-based policies have been even less successful in raising student learning. Borkum, He, and Linden (2009) conducted an RCT in Bangalore, India in 2007 of the impact of a program staffed by trained librarians in primary schools. The program was implemented using a "hub and spoke" system in which there were physical libraries in "hub" schools, while "spoke" schools were served by a mobile librarian. The program failed to increase language competency. In fact, estimates were sufficiently precise to rule out effects larger than .11 and .13 of a standard deviation based on the confidence intervals (Box 10). These findings indicate that improving school libraries is unlikely to have a positive impact on student achievement.

Box 10. What Is a Confidence Interval?

A confidence interval gives the range of possible values that an estimate could potentially take if the same evaluation were to be done over and over again. For example, suppose that a study estimates that a given intervention has an effect of .20 of a standard deviation. The confidence interval could say that, if the evaluation were to be repeated many times, the effect could be as small as .10 of a standard deviation or as large as .30. While many readers might not have had to interpret confidence intervals in the context of impact evaluations, they have likely heard of them in opinion polls. When poll results are reported, their "margins of error" are often mentioned. These margins are nothing more than the width of the confidence interval on each side of the estimated effect. For example, in the case mentioned above, the margin of error would be +/– .10 of a standard deviation.

In fact, at least in developing countries, additional books seem to have little effect. Glewwe, Kremer, and Moulin (2009) used an RCT to evaluate a program in Busia and Teso, Kenya that provided free textbooks to schools. The intervention had no impact on average test scores or student attendance. Interestingly, the textbooks increased the test scores of already high-achieving students by .06 of a standard deviation. But few children in lower grades (15–29 percent of median students) were actually able to read these books, since they were in English, which was the students' third language. This study is a reminder that input-based policies are likely to have a limited impact when there are good reasons to believe that the inputs provided will not be appropriately used.

A study in Busia and Teso, Kenya, conducted by Glewwe et al. (2004) found that the limited impact of input-based policies was true not only for students, but also for teachers. The authors used an RCT to evaluate the impact of free flipcharts on student learning. The flipcharts included science charts and a teacher's guide for science, charts for health, another set for math, and a wall map—mostly for grades 7 and 8. There was no evidence that the charts increased test scores. This finding was particularly interesting, since other less rigorous studies had found an effect on test scores of .20 of a standard deviation. This evaluation illustrates the importance of rigorous research to make critical policy decisions.

Providing In-service Training

One intervention geared toward supporting teachers' work that has been rigorously evaluated has been in-service teacher training. Jacob and Lefgren (2004b) evaluated a reform in Chicago in 1996 that placed 71 of the district's 489 primary school teachers on academic probation and provided them with funding for in-service training. The authors conducted an RDD, exploiting the fact that eligibility for probation was determined according to a cutoff in students' standardized reading scores. They found that marginal

increases in in-service training had no effect on either reading or math achievement. While more rigorous research is needed on the impact of in-service teacher training, these findings are a reminder that while school districts invest heavily in this type of training, there is little rigorous evidence to show it works.

Similarly, in Jerusalem, Israel, Angrist and Lavy (2001) conducted a DDA in 1995 of a handful of public schools that received a special infusion of funds primarily earmarked for in-service training. In-service teacher training in secular (i.e., nonreligious) schools improved students' test scores by .20–.40 of a standard deviation, but the effects in religious schools were less robust, meaning that they depended on the way in which the authors modeled the relationship between the treatment and student achievement. Angrist and Lavy speculated that this differential effect might be explained by the fact that the intervention started later and was implemented on a smaller scale in religious schools. The authors compared this intervention to alternative school improvement strategies and concluded that it was a relatively inexpensive way of improving achievement.

It is challenging to draw conclusions about rigorous studies on different forms of teacher support because the evidence is uneven. However, there is good reason to believe that input-based policies (mainly in developing countries) appear to have a limited impact on student achievement, especially when they do not influence classroom instruction, and that the impact of in-service teacher training might depend at least on the context where it is implemented and probably as well on its format.

Motivating Teachers to Perform

Economists typically think about incentives in the framework of principal-agent problems. The key insight in this framework, as presented in the work of Mirrlees (1976), Holmstrom (1979), and Shavell (1979), is that employers face a tradeoff when choosing how to remunerate their employees between offering them a base salary that ensures that they will receive a minimum level of compensation and making their pay conditional on their performance in order to induce them to work hard. An "efficient" contract is one that balances these two goals of "full insurance" and "first-best incentives."

Economists have noted that there might be several unintended consequences in incentive contracting. A key insight put forth by Holmstom and Milgrom (1991) and Baker (1992) is that a worker's measured performance might be quite different from the worker's total contribution to firm value, and that incentive schemes that reward the former might discourage employees' contributions to the work of their peers or the long-run effects of their actions (Gibbons and Waldman 1999). The other drawback of incentive contracting is the trade-off between intrinsic and extrinsic motivation—that is, mechanisms that

provide extrinsic rewards to employees for actions from which they derive intrinsic motivation might attenuate the latter (Bénabou and Tirole 2003).

Table 8 shows two types of interventions that have been rigorously evaluated: (1) hiring contract teachers; and (2) paying teachers for raising student achievement.

Hiring Contract Teachers

In many school systems, teachers are public employees, and once they obtain permanent employment status they can only be dismissed in extreme circumstances through a process that typically is time consuming. Therefore, instead of hiring new teachers as public employees, some countries have implemented reforms to hire teachers through annual, renewable contracts.

The most recent evaluation of this type of initiative was conducted by Muralidharan and Sundararaman (2010) in Andhra Pradesh, India starting in 2005. The program provided an extra contract teacher to 100 randomly-chosen government-run primary schools. At the end of two years, students in schools with an extra contract teacher outperformed those in comparison schools by .15 of a standard deviation in math and .13 in reading. While all students benefited from the program, those in their first year in school and in remote schools benefited the most. Contract teachers were absent less frequently than regular teachers (16 percent as opposed to 27 percent of the time) and were more likely to be found teaching (49 percent as opposed to 43 percent of the time). Finally, while the students of contract and regular teachers made similar achievement gains, contract teachers only cost one-fifth of what regular teachers cost.

In some contexts, contract teachers have been shown to positively affect achievement even when they lack formal training. In 2001, Banerjee et al. (2007) used an RCT to evaluate an initiative in Mumbai and Vadodara, India that provided government schools with an extra teacher to work with third and fourth graders who were struggling academically. These teachers, typically young women, were recruited from local communities and had only finished secondary school. Yet, they improved students' test scores by .14 of a standard deviation in the first year and .28 in the second year of the program. In fact, treated schools still scored .10 of a standard deviation higher than control schools one year after the program had ended. Perhaps most importantly, children at the bottom of the initial test score distribution and those receiving remedial teaching made the largest gains.

Paying Teachers to Improve Student Achievement

By far one of the most popular and rigorously evaluated ways in which education systems have recently tried to motivate teachers is by making part of their pay conditional on the achievement of their students.

In 2007, New York City adopted a teacher incentive program in more than 200 high-need schools. The program awarded a school up to $3,000 for every full-time unionized teacher if it met the annual performance target set by the U.S. Department of Education based on school report card scores. The target was based on student achievement and progress on the state exams for primary and middle schools, a high school leaving exam, student attendance, and a learning environment survey administered to teachers, parents, and students. The program also gave schools $1,500 per full-time unionized teacher if they met at least 75 percent of their mandated target. Schools that received the bonus could choose to distribute it at their discretion. Fryer (2011) took advantage of the fact that schools were randomly assigned to the program to evaluate its impact and found no impact on student achievement or evidence that the program affected teacher retention, absenteeism, or the learning environment. The author considered several reasons why the incentives program might have been ineffective, including the possibility that incentives were not large enough, the incentive scheme was too complex, group-based incentives may not be effective, or teachers may not know how they can improve student achievement.

Results from an experiment in Nashville, Tennessee were equally discouraging. Springer et al. (2010) evaluated the impact of the Project on Incentives in Teaching (POINT), an initiative that offered middle-school math teachers bonuses of up to $15,000 per year for getting their students to make unusually large gains on the state exam—the Tennessee Comprehensive Assessment Program (TCAP)—during the 2006–08 school years. The authors conducted an RCT in which half of the teachers who volunteered for the study were randomly assigned to a group that was eligible for the bonuses and half were assigned to a group that was not eligible. Springer and his collaborators found that students assigned to eligible teachers on average performed no differently than those assigned to noneligible teachers. While fifth graders of eligible teachers outperformed their peers in the second and third years of the study, this effect did not persist once these students moved on to the next grade.

Most other rigorous studies on performance pay plans for teachers were conducted in developing countries. Muralidharan and Sundararaman (2009) carried out an RCT to evaluate a teacher incentive program in the Indian state of Andhra Pradesh. The program offered bonuses of about 3 percent of teachers' annual pay, based on the average

improvement of their students' test scores in independently administered tests. After two years, the bonus program increased student achievement by .28 of a standard deviation in math and .16 in reading. Interestingly, incentive schools performed better on both mechanical and conceptual components of the tests, suggesting these improvements were not driven by "teaching to the test." In fact, students in treated schools also did better in nonincentivized subjects such as science and social studies. Finally, the results of individual- and group-based incentives were not different in the first two years of the program. This study suggests that well-designed merit pay plans can have important effects on student achievement in developing countries.

A study in Kenya, however, drew attention to the importance of what these programs measure and reward. Glewwe, Ilias, and Kremer (2010) conducted an RCT of a program in Busia and Teso, Kenya that rewarded primary school teachers with in-kind prizes based on students' scores on district exams. The program assigned low scores to students who did not take the exam so as to eliminate perverse incentives for schools to discourage low-scoring students from taking the exams. Students in treated schools scored higher than those in control schools in the first year. Yet, most of the gains were focused on the aspects measured by the reward formula, and the program had no effect on teacher behavior (including teacher attendance, the amount of homework that they assigned, and teacher pedagogy). The bonus increased test preparation by 4.2 percent in the first year and by 7.4 percent in the second year. These findings illustrate that, unless they are adequately designed, incentive programs can lead teachers to focus on test preparation but produce little change that is conducive to better learning.

In Chile, Rau and Contreras (2009) used DDA, RDD, and PSM to evaluate a program that offered bonuses to all teachers in public or publicly subsidized primary and middle schools based on students' performance and progress on the national student achievement test and other school factors. The bonus was small, but far from negligible (about 10-40 percent of a teacher's monthly wage). The program increased average learning outcomes by .07–.12 of a standard deviation. Interestingly, however, there was no evidence that winning the bonus led to any additional gains. These findings hint at the possibility that while the introduction of the program might have led schools to work harder to raise student learning through easy efficiency gains, the complexity (and therefore, predictability) of the bonus might have given them little clue about how to accomplish improvements beyond these initial gains.

Some of the evaluations of teacher incentives also shed light on aspects that can be useful in the design of these initiatives. For example, Lavy (2008, 2009) used an RDD to evaluate a program in Israel that awarded monetary bonuses to individual teachers based on students' average scores on matriculation exams and passing rates, relative to

predicted scores (adjusted for students' socioeconomic status). The bonus was large (70–300 percent of a teacher's monthly wage) and teachers were likely to receive it (about 48 percent of teachers got some award). The program resulted in students in treated schools having pass rates that were 14 percent higher in math and 5 percent higher in English. These students also had 10 percent higher scores in math and 4 percent higher scores in English. Yet, Lavy found that while the effects of the bonus did not differ by gender, female teachers were more pessimistic about their likelihood of getting it. This study suggests that even when bonuses are predictable, different groups of teachers might react to their predictability differently.

An evaluation of a teacher incentive program in Mexico shed light on the relative importance of the different aspects being rewarded and the persistence of rewards in a merit pay program. McEwan and Santibáñez (2005) used an RDD to evaluate the program, which awarded teachers and principals if they scored above a cutoff on an index that considered their education, experience, and students' test scores, among other factors. The awards were not only substantial, they also persisted through teachers' careers. However, there was no evidence that student test scores improved as a result of the reform. The authors speculated that this might have been because the bulk of the award (about 80 percent) was determined by the background of teachers and principals, rather than by their performance on the job. This suggests that, in designing incentive programs, education systems should ensure that bonuses are linked to the types of changes in employee behavior and productivity that they seek to promote.

Finally, a study in Kenya offers important insight on the potential of input-based rather than output-based teacher incentives. Kremer et al. (2001) used an RCT to evaluate an individual bonus program at the pre-school level that offered principals resources to award teachers bonuses for good attendance. The bonus was sizable, constituting up to 300 percent of teachers' monthly wages. Importantly, however, principals distributed the bonus among all teachers in their schools, regardless of their actual attendance. The bonus had no impact on teacher absenteeism (on average about 29 percent), student attendance, or achievement. In addition to lending further support to the perils of free-riding raised by the study on the New York City bonus, this study seems to suggest that there is little evidence that input-based bonuses have much promise to increase student learning.

In sum, the studies on interventions seeking to raise teacher effort and productivity yield at least two clear lessons. First, hiring contract teachers seems to be a promising way for developing countries to expand their teacher workforces. Second, while the effects of merit pay programs in developed countries seem mixed, the evidence in developing countries is much more encouraging. Yet, key design features (e.g., group versus

individual bonuses, coverage, performance measures, award processes, predictability, bonus size, etc.) are crucial in mediating the impact of these plans (Bruns and Santibáñez 2011).

Making Sense of the Theory and Evidence: Take-Aways and Caveats

This paper has tried to provide an analytical framework to make sense of the evidence on teacher policies in developed and developing countries and to engage readers in a conversation about what is known and how well it is known. Unlike many prior reviews, this paper has not sought to "distill" the key lessons from the available evidence for a nontechnical audience. Rather, the aim has been to explain the methods used in the studies of teacher policies in hopes of promoting an informed dialogue that includes all stakeholders, from citizens to teachers, government officials, experts, and others.

This concluding section offers some final observations about the state of the evidence on teacher policies and puts forth some directions for future research.

First, evidence on teacher policies is uneven and one should be cautious about how to interpret it. Much is known about some policies and very little or nothing about others. This is for several reasons. One is that there are certain areas in which there have not been opportunities to conduct rigorous research, mostly because rules and regulations make it difficult for researchers to implement some of the methodological strategies that have been reviewed here. Another reason is that some interventions are still quite new and it will take some time for researchers to evaluate their results.

In the absence of rigorous studies on a particular intervention, it can be tempting to look at the best available study, even if it is not able to distinguish the impact of an intervention independent from other factors that may affect the outcomes of interest. This perspective is understandable: as stated at the onset of this paper, policy decisions need to be made on a daily basis and governments cannot wait for the most rigorous studies to be published before they make a decision. However, it is also clear that this "next-best research" approach has not always served the education community well in the past. There have been several important teacher policies for which initial evidence looked promising until rigorous studies indicated otherwise. Rather than relying on nonrigorous studies, governments should design new interventions in ways that render them amenable to rigorous evaluation. In this way, a policy itself can inform future decisions.

Second, while the focus here has been on the impact of teacher policies on student achievement, important outcomes that governments need to take into account have been left out. As several studies on the economics of education have shown, there can often be

important "sleeper" (i.e., dormant) effects of policies that cannot be identified until children reach adolescence or even adulthood (e.g., health, crime, family planning, earnings, civic participation). Governments need to put in place robust data systems that allow researchers to assess these key long-term outcomes.

Third, this paper has identified a need to better understand not only the effects of each teacher policy in isolation, but also the interactions between teacher policies. Several studies have begun to do this by designing evaluations that assess the impact of different combinations of interventions (e.g., class reductions by themselves, class reductions with teacher monitoring, and class reductions with teacher monitoring and ability tracking). We applaud these efforts and encourage others to follow suit.

Finally, we believe there is a pressing need for more cost-benefit considerations in impact evaluations. It is often the case that several interventions can tackle the same problem effectively, but information is lacking about cost-benefits to make sound decisions about which intervention promises a bigger "bang for the buck." Some scholars have begun to pay attention to these issues in other areas of education (Dhaliwal et al. 2011), but more energy should be devoted to examining cost issues in the evaluation of teacher policies, in particular.

Table 1. Evidence on Setting Clear Expectations for Teachers

Scaffolding Teachers' Work				
Intervention	**Location**	**Study**	**Method**	**Results**
Scripted lessons (pre-school, grade 2) with a structured curriculum, reading activities, ability tracking, and compulsory reading at home.	United States	Borman et al. (2007)	Very Strong (RCT)	▪ Effect sizes ranged from .21 SD on the passage comprehension literacy measure to .33 SD on the word attack measure.
Pre-packaged curriculum and activities (pre-school and grade 1) for reading.	Mumbai, India	He, Linden, and MacLeod (2009)	Very Strong (RCT)	▪ Gains in students' reading scores of .26–.70 SD, proving most effective in pre-school classes and with low-performing students. ▪ The version of the program taught out of school time is more effective than the school-day version, yielding a .24 additional SD effect over the .26 SD effect.
Pre-packaged classroom activities (grades 1–5) either with a machine or flashcards.	Maharasthra, India	He, Linden, and MacLeod (2007)	Very strong (RCT)	▪ Gains in English achievement of .30 SD. ▪ Older and lower-performing students benefited the most from the program. ▪ The version of the program implemented by teachers improved not only students' English scores, but also their math scores, suggesting positive spillovers.
Reading marathon (grade 4) accompanied by a training program for teachers and provision of reading materials.	Tarlac, Philippines	Abeberese, Kumler, and Linden (2011)	Very strong (RCT)	▪ Increased the number of books read in the month after the program from 2.3 to 9.5. ▪ Students' reading scores increased by .13 SD. ▪ Three months after the reading marathon, treated students still read 3.1 books more than the comparison group and had reading scores that were .06 SD higher.

Manual of operations and school-level report cards (grade 3-4).	Madagascar	Lassibille et al. (2010); Glewwe and Maïga (2011)	Very strong (RCT)	▪ No statistically significant effects on test scores. ▪ The impact did not vary by teacher type (i.e., whether civil service, contract, or student teachers). ▪ The program raised attendance and reduced grade repetition only when coupled with interventions at the subdistrict and district levels.
Increasing Instructional Time				
Intervention	**Location**	**Study**	**Method**	**Results**
After-school program and summer school (grades 5–6), including academic and enrichment activities and mentoring after school and during the summer.	Washington, DC	Linden, Herrera, and Grossman (2011)	Very strong (RCT)	▪ Higher test scores by .09 SD (reading) and .12 SD (problem-solving) by the second year. ▪ Participating students were also more likely to visit a high school, get information, and talk to adults and peers about high schools to decide where to apply.
Summer school (grades 3 and 6), which is compulsory for all students who do not pass a competency exam to move on to the next grade.	Chicago	Jacob and Lefgren (2002)	Strong (RDD + IVE)	▪ .11–.13 SD higher test scores in reading and math in grade 3 but no higher test scores in grade 6.
Reducing Class Size				
Intervention	**Location**	**Study**	**Method**	**Results**
Class-size reductions (pre-school to grade 3) through small classes or regular-size classes with a teacher aide.	Tennessee	Krueger (1999)	Very strong (RCT)	▪ The average per-student effect was about .048 SD in reading and math. ▪ The effect was strongest for minority students and those with low socioeconomic status.
Class-size reductions (grade 1) by hiring a contract teacher, tracking students by initial ability, and/or empowering parents to monitor	Busia and Teso, Kenya	Duflo, Dupas, and Kremer (2007)	Very strong (RCT)	▪ Reducing class size (on average from 80 to 46 students), in the absence of any other reform, led to lower teacher effort and no discernible improvement in student achievement.

teacher performance.				▪ Combining class-size reductions with improved incentives led to significantly higher test scores—either by hiring contract teachers and increasing parental oversight (an improvement of .19 SD in literacy and numeracy) or hiring contract teachers and tracking students (an improvement of .25–.31 SD).
Class-size reductions (grades 1–6) emerging from idiosyncratic variations in student population or from maximum/minimum rules.	Connecticut	Hoxby (2000)	Strong (IVE + RDD)	▪ No effect on student achievement. ▪ It is possible to rule out even modest effects such as .02—.04 SD for a 10% reduction in class size.
Class-size reductions (grades 3–5) through a rule that capped the maximum number of students per class at 40.	Israel	Angrist and Lavy (1999)	Strong (RDD)	▪ The per-student effect was .017–.019 SD in fourth grade and .036–.071 SD in fifth grade, combining student achievement in both subjects.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: IVE = instrumental variables estimation; RCT = randomized control trial; RDD = regression discontinuity design; SD = standard deviation.

Table 2. Evidence on Attracting the Best People into Teaching

Setting Requirements for Entry into Teaching				
Intervention	**Location**	**Study**	**Method**	**Results**
Professional certification (grades 2–5) based on classroom videos and essays.	Los Angeles	Cantrell et al. (2008)	Very Strong (RCT)	▪ Students of teachers with professional certification scored .22 SD higher than those of teachers who applied to become certified but did not receive the certificate. However, students of certified teachers were not more effective than students of nonapplicant teachers. ▪ Students of successful applicants also achieved greater gains than those of unsuccessful applicants, but not greater than those of nonapplicants. ▪ Students of high-scoring successful applicants outperformed those of low-scoring successful applicants.
Content knowledge tests (grade 6) for teachers already in service.	Peru	Meltzer and Woessmann (2010)	Tentative (student, teacher, and subject fixed effects)	▪ 1 SD increase in teacher test scores in math and reading increases student test scores by .10 SD.
Traditional teaching credentials (grades 3–5), including teaching experience, test scores, and licensure.	North Carolina	Clotfelter, Ladd, and Vigdor (2007a)	Tentative (student, school, subject, and year fixed effects)	▪ A teacher's experience, test scores, and regular licensure have positive effects on student learning, with larger effects for math than for reading. ▪ The magnitudes are large when compared to the effects of changes in class size or to the socioeconomic characteristics of students.
Teacher certification (grades 4–8).	New York City	Kane, Rockoff, and	Tentative (school, grade, and	▪ The certification status of a teacher has a small effect on student performance. ▪ Among teachers with the same certification status,

		Staiger (2006)	year fixed effects)	there are large and persistent differences in teacher effectiveness. ▪ Even high turnover groups from alternative pathways into teaching would have to be only slightly more effective in their first year to offset the effects of their high exit rates.
Content knowledge tests (grades 4–8) for teachers already in service.	Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City; and Memphis, Tennessee	Cantrell and Kane (2013)	Tentative (correlations over two sections of students)	▪ Tests of teachers' subject-specific pedagogical knowledge in math and reading are correlated with teachers' performance on subject-specific observations, but not with teachers' value added. ▪ The tests did not correlate with teachers' value added at any point of the value-added distribution (i.e., it was not useful in identifying the bottom-performing, average-performing, or top-performing teachers).
Relaxing Entry Requirements for Outstanding Individuals				
Intervention	**Location**	**Study**	**Method**	**Results**
Alternative pathway into teaching (grades 1–5) that recruits outstanding college graduates to teach in high-need schools for two years.	United States	Decker, Mayer, and Glazerman (2004)	Very strong (RCT)	▪ The program attracted teachers who were more likely to attend a competitive college and to attain their education degree while teaching and less likely to have education-specific training and student-teaching experience than others in their schools. ▪ The students of teachers in the program performed .15 SD higher in math than students of other teachers (or 10% of a grade equivalent), but no differently in reading. ▪ Compared to students of other novice teachers,

				students of teachers in the program performed .26 SD higher in math, but no differently in reading.
Alternative pathway into teaching (pre-school to grade 5) that combines all nontraditionally trained teachers.	United States	Constantine et al. (2009)	Very strong (RCT)	▪ Traditionally and alternatively certified teachers (TC and AC, respectively) varied widely in their hours of pre-service training. ▪ TC and AC teachers did not differ in their scores on college entrance exams, the selectivity of their college, or their educational attainment. ▪ TC and AC teachers did not differ in their effectiveness in raising student learning. ▪ The amount or content of teacher training did not seem to impact the effectiveness of AC teachers.
An alternative pathway into teaching (grades 7–9) that recruits outstanding college graduates to teach in high-need schools for two years.	Chile	Alfonso, Santiago, and Bassi (2011)	Tentative (PSM)	▪ Increased scores by .22–.51 SD in reading and by .17–.43 SD in math in participating schools in its first year and by .75 SD in Spanish and .33 SD in math in its second year. ▪ Participating schools also had students with higher measures of self-esteem, self-efficacy, and intellectual and meta-cognitive skills.
Rewarding Advanced Educational Qualifications and Experience				
Intervention	**Location**	**Study**	**Method**	**Results**
Teaching experience (grades 4–8) from 3 to 12 years.	Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City;	Kane and Staiger (2012)	Tentative (school, grade, and subject fixed effects, two sections of students)	▪ Teachers with 12 or more years of experience have students with .01 SD higher scores in math and .02 SD higher scores in reading on state tests compared to students of teachers with fewer than three years of experience. These results were lower when math and reading were measured on supplemental tests. ▪ Teachers with a master's degree had students with .03 SD higher scores in math and.02 SD lower scores in reading on state tests compared to teachers without a master's. These results were higher when math and reading were measured on supplemental tests. ▪ Gains from experience were considerably smaller

	and Memphis, Tennessee			than gains on a combined measure of teacher effectiveness, which included teachers' value added, as well as their performance on a student survey and on classroom observations.
Advanced degrees and certification (grades 3–8).	Texas	Hanushek et al. (2005)	Tentative (school and student fixed effects)	▪ Teacher quality appears to be unrelated to advanced degrees or certification, but experience in the first year of teaching seems to matter. ▪ Good teachers tend to be effective with all student ability levels but there is a positive value of matching students and teachers by race.
Increasing Teacher Pay				
Intervention	**Location**	**Study**	**Method**	**Results**
Improvements in teacher pay (elementary and secondary school) in terms of levels and distribution.	United States	Hoxby and Leigh (2004)	Strong (IVE)	▪ Pay compression increased the share of lowest-aptitude female college graduates who became teachers by 9 percentage points and decreased the share of highest-aptitude female college graduates who become teachers by 12 percentage points. ▪ Improvements in pay parity reduced the share of women who taught by 3.2 percentage points for the highest aptitude group and 0 for the three lower aptitude groups.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: AC = alternatively certified teachers, IVE = instrumental variables estimation; PSM = propensity score matching; RCT = randomized control trial; SD = standard deviation; TC = traditionally certified teachers.

Table 3. Evidence on Preparing Teachers with Useful Training and Experience

Setting Requirements for Entry into Teaching				
Intervention	**Location**	**Study**	**Method**	**Results**
Compulsory mentoring (grades 4–8) in which new teachers with less than one year of experience met weekly with a mentor.	New York City	Rockoff (2008)	Strong (DDA + IVE)	▪ Teachers without prior teaching experience were 4.5% more likely to complete their first year if they had been assigned to a mentor. ▪ Retention in a school is higher when a mentor has previous experience at that school. ▪ For every 10 hours of mentoring, students of mentored teachers scored .05 SD higher in math and .04 SD higher in reading.
Comprehensive mentoring (grades 2–5) with intensive mentorship support for new teachers, orientations, professional development, classroom observations, and feedback.	United States	Glazerman et al. (2010)	Very strong (RCT)	▪ 90% of new teachers in districts assigned to a comprehensive mentoring program had a mentor assigned to them, compared to 72% of those in districts with a regular mentoring program. ▪ Teachers receiving comprehensive mentoring for one year were no more effective, as measured by observations and student achievement, than those with business-as-usual mentoring. ▪ Teachers with two years of comprehensive mentoring raised student achievement by .11 SD in reading and .20 SD in math in the third year of the study. ▪ Comprehensive mentoring had no effects on teacher retention or satisfaction.
Including a Practice Component in Teacher Preparation				
Intervention	**Location**	**Study**	**Method**	**Results**
Teacher residency (grades 4–8) with local recruitment, year-long residency with a mentor	Boston	Papay et al. (2011)	Tentative (school,	▪ Participating teachers are more racially diverse than other novice teachers in their district, more likely to teach math and science, and more likely to remain

four days a week, gradual classroom responsibilities, and a three-year commitment.			student, grade, and year fixed effects)	teaching in their district through their fifth year. ▪ Initially, these teachers are no more effective at raising student test scores than other novice teachers in language and less effective than them in math, but their effectiveness improves rapidly over time, such that by their fourth and fifth year they outperform veteran teachers.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; IVE = instrumental variables estimation; RCT = randomized control trial; SD = standard deviation.

Table 4. Evidence on Matching Teachers' Skills with Students' Needs

Offering Bonuses for Teachers to Work in High-need Schools				
Intervention	**Location**	**Study**	**Method**	**Results**
Bonuses to attract and retain talented novice teachers to low-performing schools (all grades) for at least four years, selected based on grade point averages, recommendation letters, CVs, essays, and interviews.	California	Steele, Murnane, and Willett (2010)	Strong (IVE)	▪ A $20,000 signing bonus increased by 28 percentage points the probability that its recipients taught in low-performing schools. ▪ 75% of both bonus and nonbonus recipients who began working in low-performing schools stayed in such schools for at least four years.
Offering Bonuses for Teachers to Teach Critical Shortage Areas				
Intervention	**Location**	**Study**	**Method**	**Results**
Bonuses to retain certified teachers in math, science, and special education (grades 10–12) in low-income or low-performing schools.	North Carolina	Clotfelter et al. (2008)	Strong (DDA)	▪ Increased year-on-year retention by 10–13%. ▪ Math teachers receiving the bonus were 18% less likely to depart the following year than those who did not get the bonus, but science and special education teachers were no less likely to leave. ▪ Middle-school teachers were 27% less likely to leave if they had received a bonus, but bonus-receiving high school teachers showed no statistically significant difference in attrition patterns.
Improving Working Conditions				
Intervention	**Location**	**Study**	**Method**	**Results**
Job matching (grades 3–5) as measured by teachers who switch schools.	North Carolina	Jackson (2013)	Tentative (teacher, school, and year fixed	▪ Teachers' value-added scores improved by .09 SD in math and .07 SD in reading when they moved to a school that is a better match for them. ▪ Match quality was also negatively correlated with

			effects)	school-switching, unrelated to exit from the profession, and increased with experience.
Effective peers (grades 3–5) as measured by the prior student achievement of students of peers of a teacher who switches schools.	North Carolina	Jackson and Bruegman (2009)	Tentative (student, teacher, school, and year fixed effects)	▪ Teachers who experience a 1 SD improvement in observable peer characteristics are associated with .10 SD in the test scores of their students in math and reading. ▪ A 1 SD increase in a teacher's peers' mean value added increases student test scores by .05 SD in math and reading.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; IVE = instrumental variables estimation; SD = standard deviation.

Table 5. Evidence on Leading Teachers with Strong Principals

Hiring More Effective Principals				
Intervention	**Location**	**Study**	**Method**	**Results**
Hiring more effective principals (grades 3–8) as measured by the average performance of students in their schools.	Texas	Branch, Hanushek, and Rivkin (2012)	Tentative (principal and school fixed effects)	▪ The annual impact of having an effective rather than an ineffective principal is .05–.21 SD, depending on the method used to identify this effect. ▪ Teachers who leave schools with the most successful principals are more likely to have been among the least effective teachers in that school than teachers leaving schools run by less successful principals. ▪ It is not always the case that more effective principals are more likely to remain in their position and less likely to leave public schools.
Hiring more effective principals (grades 3–10) as measured by the average performance of students in their schools.	Miami	Grissom, Kalogrides, and Loeb (2012)	Tentative (school, neighborhood, and peer fixed effects)	▪ The choice of model to estimate principal effects is substantively important for assessment. ▪ While some models identify principal effects as large as 0.15 SD in math and 0.11 SD in reading, others find effects as low as 0.02 SD in both subjects for the same principals.
Hiring more effective principals (grades 3–10) as measured by the average performance of students in their schools.	British Columbia, Canada	Coelli and Green (2012)	Tentative (school, neighborhood, and peer fixed effects)	▪ Getting a principal who is 1 SD better in the principal effects distribution implies that graduation rates will be .33 SD higher and English scores will be 1 SD higher in grade 12.

Setting Requirements for Becoming a Principal				
Intervention	**Location**	**Study**	**Method**	**Results**
Surveys of parents, teachers, and assistant principals on principal effectiveness.	Miami	Grissom and Loeb (2011)	Tentative (teacher and student fixed effects)	▪ A 1 SD increase in principals' self-rating on organizational skills is associated with a .10 SD increase in school accountability performance. ▪ Principals' self-rating on organizational skills is also related to parental ratings of their children's schools, but not consistently with teacher satisfaction. ▪ The ratings given by assistant principals of principals' organizational skills are also associated with math and reading gains.
Principal experience and education as measured by selectivity of undergraduate and graduate institution, prior work experience, and professional development.	New York City	Clark, Martorell, and Rockoff (2009)	Tentative (school effects)	▪ Little relationship between school performance and the selectivity of a principal's undergraduate or graduate institution. ▪ Little relationship between performance and a principal's prior work experience. ▪ Principals with three years of experience had students whose math scores were .04 SD higher than those of students with principals in their first year, and principals with five or more years of experience had students whose math scores were .06 SD higher. ▪ Mixed evidence on the relationship between principal training and professional development programs and school performance.
Granting Principals More Authority over Staffing Decisions				
Intervention	**Location**	**Study**	**Method**	**Results**
Providing teacher value-added scores (grades 4–8) in which principals received performance measures of their teachers based on student test score outcomes.	New York City	Rockoff et al. (2011)	Very strong (RCT)	▪ Increased the likelihood that teachers with low performance estimates exited their schools after the information was provided, causing a small improvement in teacher productivity at these schools. ▪ Receipt of "hard" performance data did not "crowd

				out" information that principals collect through classroom observation.
Discretion for principals to dismiss probationary teachers (grades 3, 5, and 8) for any reason in an expedited way.	Chicago	Jacob (2010a and b)	Strong (DDA + matching)	▪ Reduced absences among probationary teachers by roughly 10% and reduced the prevalence of teachers with 15 or more absences by 20%. ▪ Increased student achievement in elementary. Effects are only partly explained by changes in teacher absenteeism, suggesting that the policy affected teacher behavior. ▪ There is tentative evidence that principals consider teacher absences and value-added measures, along with several demographic characteristics, in choosing whom to dismiss.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; IVE = instrumental variables estimation; RCT = randomized control trial; SD = standard deviation.

Table 6. Evidence on Monitoring Teaching and Learning

Increasing Community and Parental Involvement in School Affairs				
Intervention	**Location**	**Study**	**Method**	**Results**
Increasing community participation in school management or support (grades 1–6) by explaining the role of village education committees to community members, training community volunteers to administer assessments, or training community volunteers to provide remedial instruction.	Uttar Pradesh, India	Banerjee et al. (2010)	Very strong (RCT)	▪ A program that explained the role of village education committees to their members and their communities increased committee members' knowledge of their role by .39 SD. ▪ A program that trained volunteers to administer and report on reading, writing, and math assessments increased committee members' knowledge of their role by .35 SD. ▪ A program that trained local volunteers to provide remedial reading classes after school increased committee members' knowledge of their role by .32 SD and increased students' reading test scores by about .02 SD.
Providing information to communities about their role in school management (grades 1–5) through state-mandated roles and responsibilities in school management	Karnataka, Madhya Pradesh, and Uttar Pradesh, India	Pandey, Goyal, and Sundararaman (2009)	Very strong (RCT)	▪ Increased number of meetings of village education committees and member participation in school inspections by 25% in Uttar Pradesh and increased the percentage of parents who talked to teachers about the quality of education in Madhya Pradesh. ▪ The program also raised teacher attendance by 11% in Uttar Pradesh and increased teacher classroom activity by 30% in Madhya Pradesh. ▪ The program increased reading achievement in grade 3 in both Uttar and Madhya Pradesh.
Organizing meetings with parents (grade 6) to discuss helping their children with homework, report cards, and other school issues.	France	Avvisati et al. (2008)	Very strong (RCT)	▪ Increased parental involvement at school and home. ▪ Children of treated families developed more positive behavior and attitudes in school and had fewer literacy problems.

				▪ There are large spillover effects on classmates of treated families.
Grading or Ranking Schools Based on Student Achievement				
Intervention	**Location**	**Study**	**Method**	**Results**
School- and student-level report cards (grade 3) with notes and rankings in English, math, and Urdu, with other schools and villages used as benchmarks.	Punjab, Pakistan	Andrabi, Das, and Khwaja (2009)	Very strong (RCT)	▪ Higher learning by .10 SD. ▪ Gains are heterogeneous: report cards produce a .34 SD increase in schools that initially scored below the median baseline, they have no effect on schools that initially scored above the median, and they cause a .10 SD increase in scores in government schools. ▪ Private school fees decrease by 18%.
Assigning grades to schools (grades 3–10) using an A-to-F scale based on proficiency rates in reading, writing, and math.	Florida	Feng, Figlio, and Sass (2010)	Strong (DDA)	▪ Teachers working in schools that received a lower grade than expected are 11% more likely to leave their school and those that received a higher grade than expected are 2.3% less likely to leave. ▪ Teachers at schools downgraded to a failing grade are 42% more likely to leave and 67% more likely to move to a school in the same district. ▪ The effectiveness of teachers staying in downgraded schools increased by .05 SD in student achievement, but since the teachers leaving these schools were of higher quality, the net effect on the teacher composition was nil.
Grading schools and linking grades to rewards and consequences (grades 4–8), including possible school closure.	New York City	Rockoff and Turner (2010)	Strong (RDD)	▪ Increased student achievement in math and English and improved parental evaluations of school quality.
Ranking schools by student achievement (grades 4, 8, 10) accounting for their socioeconomic status, and offering them a	Chile	Mizala and Urquiola (2007)	Strong (RDD)	▪ No consistent effect on the learning outcomes of top-ranked schools that receive the award.

monetary incentive for good performance.				
Monitoring Teacher Effort				
Intervention	**Location**	**Study**	**Method**	**Results**
Monitoring and rewarding teacher attendance (grades 1–6) using tamper-proof cameras and making salary a function of attendance.	Rajasthan, India	Duflo, Hanna, and Ryan (2010)	Very strong (RCT)	▪ Rural schools that monitored teacher performance with cameras and made individual teachers' salaries a direct function of their attendance reduced teacher absenteeism from 42% to 23%. ▪ After a year, test scores in treatment schools were .17 SD higher than in control schools and grade completion increased.
Monitoring Teacher Performance				
Intervention	**Location**	**Study**	**Method**	**Results**
Combining multiple measures of teacher effectiveness (grades 4–8), including classroom observations, student surveys, and principal surveys.	Charlotte-Mecklenburg, North Carolina; Dallas, Texas; Denver, Colorado; Hillsborough County, Florida; New York City; and Memphis, Tennessee	Kane et al. (2013)	Very strong (RCT)	▪ A composite measure of teacher effectiveness that combines student surveys, classroom observations, and a teacher's track record of student achievement gains on state tests identified teachers who produced higher average student achievement following random assignment of teachers to classrooms. ▪ The magnitude of the achievement gains produced by teachers identified as effective prior to random assignment was proportional to their performance on the composite measure. ▪ A subset of teachers identified as effective prior to randomization also had students who performed better on "audit" tests of math and reading.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; RCT = randomized control trial; RDD = regression discontinuity design; SD = standard deviation.

Table 7. Evidence on Supporting Teachers to Improve Instruction

Providing or Improving Learning Materials				
Intervention	**Location**	**Study**	**Method**	**Results**
Providing computers to public schools (grades 3–9) to be integrated into the teaching of reading.	Colombia	Barrera-Osorio and Linden (2009)	Very strong (RCT)	▪ Little effect on students' test scores and other outcomes. The results are consistent across grade levels, subjects, and students' gender. ▪ Surveys indicate few teachers incorporate the computers into their curriculum.
Providing computer-adaptive software (grade 4) in math during and after school (two hours per week) with local instructors.	Vadodara, India	Banerjee et al. (2007)	Very strong (RCT)	▪ Increase in math scores by .36 SD in the first year, but had no effect on reading, suggesting limited spillover effects.
Providing language/reading software (grades 3–6).	United States	Rouse and Krueger (2004)	Very strong (RCT)	▪ Improved some aspects of students' learning skills, but these gains did not translate into a broader measure of language acquisition or into actual reading skills.
Introducing libraries (grades 1–6) through a "hub-and-spoke" system and trained librarians who assist students.	Bangalore, India	Borkum, He, and Linden (2009)	Very strong (RCT)	▪ No increase in language competency scores.
Providing textbooks (grades 3–8) in English, math, and science.	Busia and Teso, Kenya	Glewwe, Kremer, and Moulin (2009)	Very strong (RCT)	▪ No effect on average test scores or student attendance. ▪ Test scores increased .06 SD for already high-achieving students, but few children in lower grades were able to read the (English language) texts (15–29% of median students).
Providing flipchart materials (grades 3–8), including science charts, teachers' manuals, health charts, math charts, and geography maps.	Busia and Teso, Kenya	Glewwe et al. (2004)	Very strong (RCT)	▪ No effect on student test scores.

Providing In-Service Teacher Training				
Intervention	**Location**	**Study**	**Method**	**Results**
Providing in-service training (grades 2–8) through funding for low-performing schools.	Chicago	Jacob and Lefgren (2004b)	Strong (RDD)	▪ No effect on either reading or math achievement.
Providing in-service training (grade 4) on a weekly basis during school by external trainers; focus on Hebrew, math, and English teaching.	Jerusalem, Israel	Angrist and Lavy (2001)	Strong (DDA + matching)	▪ Teacher training in secular schools leads to an improvement in student achievement of .20–.40 SD. ▪ Estimates for religious schools are not clear-cut, perhaps because training in religious schools started later and was implemented on a smaller scale.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; RCT = randomized control trial; SD = standard deviation.

Table 8. Evidence on Motivating Teachers to Perform

Hiring Contract Teachers				
Intervention	**Location**	**Study**	**Method**	**Results**
Hiring contract teachers (grades 1–5) chosen by school committees, on an annual basis, without labor protections and a lower salary.	Andhra Pradesh, India	Muralidharan and Sundararaman (2010)	Very strong (RCT)	▪ Test scores were .15 SD higher in math and .13 SD higher in reading. ▪ Contract teachers were absent 16% of the time while regular teachers were absent 27% of the time. ▪ Contract teachers were found teaching 49% of the time, regular teachers 43% of the time. ▪ There is tentative evidence that the students of contract and regular teachers make the same achievement gains, although contract teachers cost one-fifth of what regular teachers cost.
Hiring contract teachers (grades 3–4) who pull children out of their regular classrooms to give them remedial instruction.	Mumbai and Vadodara, India	Banerjee et al. (2007)	Very strong (RCT)	▪ Test scores were .14 SD higher in the first year and .28 SD higher in the second year than those of control schools. ▪ Treated schools still scored .10 SD higher one year after the program ended. ▪ Children at the bottom of the initial test score distribution and those receiving remedial teaching made the largest gains.
Paying Teachers to Raise Student Achievement				
Intervention	**Location**	**Study**	**Method**	**Results**
Bonuses for performance and inputs (grades 3–12) based on attendance, climate, level, and progress in student achievement. Schools decide	New York City	Fryer (2011)	Very strong (RCT)	▪ No increase in student achievement. ▪ No evidence that teacher incentives affect teacher

how to distribute the bonuses.				retention, absenteeism, or the learning environment.
Bonuses for performance (grades 5–8) for reaching a minimum performance level in math.	Nashville, Tennessee	Springer et al. (2010)	Very strong (RCT)	▪ Middle-school students assigned to teachers eligible to receive bonuses for improving student achievement on average performed no differently than those assigned to noneligible teachers. ▪ Fifth grade students of eligible teachers outperformed their peers in the second and third years of the study, but this effect did not persist once these students moved on to the next grade.
Bonuses for individual and group performance (grades 2–5) for reaching a minimum threshold and for additional increases.	Andhra Pradesh, India	Muralidharan and Sundararaman (2009)	Very strong (RCT)	▪ Improved student achievement by .28 SD (math) and .16 SD (reading) after two years. ▪ Incentive schools performed better both on mechanical and conceptual components of the test, suggesting little evidence of adverse consequences. ▪ Students in incentive schools also do better in nonincentivized subjects, such as science and social studies, suggesting positive spillover effects. ▪ The results of individual- and group-based incentives were not statistically significant in either the first or the second year.
Bonuses for group performance (grades 4–8) for ranking according to performance level and change.	Busia and Teso, Kenya	Glewwe, Ilias, and Kremer (2010)	Very strong (RCT)	▪ Schools where school committees gave bonuses to teachers whose students did well on exams scored .14 SD higher than control schools in the first year. ▪ The program had no effect on teacher attendance, the amount of homework that teachers assigned, or teacher pedagogy. ▪ The program increased test preparation by 4.2% in the first year and 7.4% in the second year.
Bonuses for group performance and inputs (grades 4, 8, and 10) based on student learning outcomes and teacher or school inputs.	Chile	Rau and Contreras (2009)	Strong (DDA + PSM + RDD)	▪ .07–.12 SD increase in average learning outcomes. ▪ No evidence that winning the bonus led to additional gains.

Bonuses for individual performance (grades 10–12) based on students' passing rates and average scores on their high school matriculation exams.	Israel	Lavy (2008, 2009)	Strong (RDD + matching)	▪ Students in treated schools had 14% higher pass rates and 10% higher scores in math, and 5% higher pass rates and 4% higher scores in English. ▪ The effects of the bonus did not differ by gender, although female teachers were more pessimistic about their possibility of receiving an award.
Bonuses for individual performance and inputs (grades 3–6) based on educational credentials, experience, professional development, peer evaluations, subject matter knowledge and student achievement.	Mexico	McEwan and Santibáñez (2005)	Strong (RDD)	▪ No evidence that student test scores improved as a result of the performance-based pay reform.
Bonuses for group performance and other factors (grades 9–11) based on school rankings on credits, graduation, dropout rates and student achievement.	Israel	Lavy (2002)	Strong (RDD)	▪ .13 SD improvement in student achievement and modest increases in credits earned and the percentage of students taking matriculation exams.
Bonuses for individual inputs (pre-school) including absenteeism. Principals decide how to allocate the bonuses.	Busia and Teso, Kenya	Kremer et al. (2001)	Very strong (RCT)	▪ No effect on teacher absenteeism, student attendance, or achievement.

Source: **[you need to add source, even if it says "Prepared by the author."]**
Note: DDA = differences-in-differences analysis; PSM = propensity score matching; RCT = randomized control trial; RDD = regression discontinuity design; SD = standard deviation.

References

Abdulkadiroglu, A., J.D. Angrist, S.M. Dynarski, T.J. Kane, and P.A. Pathak. 2011. Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots. *Quarterly Journal of Economics* 126: 699–748.

Abeberese, A.B., T.J. Kumler, and L.L. Linden. 2011. *Improving Reading Skills by Encouraging Children to Read: A Randomized Evaluation of the Sa Aklat Sisikat Reading Program in the Philippines.* Cambridge, MA: Innovations for Poverty Action (IPA).

Acemoglu, D., and J.-S. Pischke. 1998. Why Do Firms Train? Theory and Evidence. *Quarterly Journal of Economics* 113(1): 79–119.

Akerlof, G.A. 1970. The Market for "Lemons:" Quality Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84(3): 488–500.

Akerlof, G.A. 1982. Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics* 97(4): 543–69.

Alcázar, L., F.H. Rogers, N. Chaudhury, J. Hammer, M. Kremer, and K. Muralidharan. 2006. Why Are Teachers Absent? Probing Service Delivery in Peruvian Primary Schools. International Journal of Educational Research 45: 117–36.

Alfonso, M., A. Santiago, and M. Bassi. 2011. *Estimating the Impact of Placing Top University Graduates in Vulnerable Schools in Chile.* Washington, DC: Inter-American Development Bank.

Andrabi, T., J. Das, and A. Khwaja. 2009. *Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets.* Washington, DC: World Bank.

Angrist, J.D., and J. Guryan. 2008. Does Teacher Testing Raise Teacher Quality? Evidence from State Certification Requirements. *Economics of Education Review* 27(5): 483–503.

Angrist, J.D., and Lavy, V. 1999. Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics* 114(2): 533–75.

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

Arrow, K.J. 1963. Uncertainty and the Welfare Economics of Medical Care. *American Economic Review* 53(5): 941–73

Autor, D.H. 2003a. Lecture Note: Self-Selection – The Roy Model. MIT 14.661 (November 14). Massachusetts Institute of Technology.

Autor, D.H. 2003b. Lecture Note: Monitoring, Measurement and Risk. MIT 14.661. November 13. Massachusetts Institute of Technology.

Avvisati, F., M. Gurgand, N. Guyon, and E. Maurin. 2008. *Quels Effets Attendre d'une Politique d'Implication des Parents d'Élèves dans les Collèges? Les Enseignements d'une Expérimentation Contrôlée.* Paris: Ecole d'Economie de Paris.

Baker, G. 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy* 100: 598–614.

Banerjee, A.V., R. Banerji, E. Duflo, and M. Walton. 2011. *What Helps Children to Learn? Evaluation of Pratham's Read India Program in Bihar and Uttarakhand.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Banerjee, A.V., S. Cole, E. Duflo, and L. Linden. 2007. Remedying Education: Evidence from Two Randomized Experiments in India. *Quarterly Journal of Economics* 122(3): 1235–264.

Banerjee, A.V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani. 2010. Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal* 2(1): 1–30.

Barlevy, Gadi, and Derek Neal. 2011. Pay for Percentile. NBER Working Paper No. 17194. Cambridge, MA: National Bureau of Economic Research.

Barrera-Osorio, F., and L.L. Linden. 2009. *The Use and Misuse of Computers in Education: Evidence from a Randomized Controlled Trial of a Language Arts Program.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Becker, G.S. 1964. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* Chicago: University of Chicago Press.

Becker, G.S., and K. Murphy. 2000. *Social Economics.* Cambridge, MA: Harvard University Press.

Bénabou, R., and J. Tirole. 2003. Intrinsic and Extrinsic Motivation. *Review of Economic Studies* 70(3): 489–520.

Borjas, G. 1987. Self-Selection and the Earnings of Immigrants. *American Economic Review* 77: 531–53.

Borkum, E., F. He, and L. Linden. 2009. *School Libraries and Language Skills in Indian Primary Schools: A Randomized Evaluation of the Akshara Library Program.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Borman, G., R.E. Slavin, A. Cheung, A. Chamberlain, N.A. Madden, and B. Chambers. 2007. Final Reading Outcomes of the National Randomized Field Trial of Success for All. *American Educational Research Journal* 44(3): 701–31.

Boyd, D., H. Lankford, and S. Loeb. 2005. The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis and Management* 24(1): 113–32.

Boyd, D., H. Lankford, S. Loeb, and J. Wyckoff. 2011. Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness. *Education Finance and Policy* 6(3): 439–54.

Boyd, D., P.L. Grossman, H. Lankford, S. Loeb, and J. Wyckoff. 2009. Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis* 31(4): 416–40.

Boyd, D., P. Grossman, K. Hammerness, H. Lankford, S. Loeb, M. Ronfeldt, and J. Wyckoff. 2010a. Recruiting Effective Math Teachers: How Do Math Immersion Teachers Compare? Evidence from New York City. NBER Working Paper 16017. Cambridge, MA: National Bureau of Economic Research.

Boyd, D., H. Lankford, S. Loeb, M. Ronfeldt, and J. Wyckoff. 2010b. The Role of Teacher Quality in Retention and Hiring: Using Applications-to-Transfer to Uncover Preferences of Teachers and Schools. NBER Working Paper 15966. Cambridge, MA: National Bureau of Economic Research.

Boyd, D., P. Grossman, M. Ing, H. Lankford, S. Loeb, R. O'Brien, and J. Wyckoff. 2011. The Effectiveness and Retention of Teachers with Prior Career Experience. *Economics of Education Review* 30: 1229–241.

Bruns, B., D. Filmer, and H.A. Patrinos. 2011. *Making Schools Work: New Evidence on Accountability Reforms.* Washington, DC: World Bank.

Bruns, B., and L. Santibáñez. 2011. Making Teachers Accountable. In *Making Schools Work: New Evidence on Accountability Reforms,* ed. by B. Bruns, D. Filmer, and H.A. Patrinos. Washington, DC: World Bank.

Cantrell, S., and T.J. Kane. 2013. *Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study.* Seattle: Bill and Melinda Gates Foundation.

Cantrell, S., J. Fullerton, T.J. Kane, and D.O. Staiger. 2008. National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment. NBER Working Paper 14608. Cambridge, MA: National Bureau of Economic Research.

Chetty, R., J.N. Friedman, N. Hilger, E. Saez, D.W. Schanzenbach, and D. Yagan. 2011. How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star. *Quarterly Journal of Economics* 126(4): 1593–660.

Clark, D., P. Martorell, and J. Rockoff. 2009. School Principal and School Performance. CALDER Working Paper No. 38. Washington, DC: National Center for Analysis of Longitudinal Data in Educational Research (CALDER), the Urban Institute.

Clotfelter, C.T., H.F. Ladd, and J.L. Vigdor. 2007a. Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects. *Economics of Education* 26(6): 673–82.

Clotfelter, C.T., H.F. Ladd, and J.L. Vigdor. 2007b. Are Teacher Absences Worth Worrying About in the U.S.? NBER Working Paper 13648. Cambridge, MA: National Bureau of Economic Research.

Clotfelter, C., E. Glennie, H. Ladd, and J. Vigdor. 2008. Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina. *Journal of Public Economics* 92(5-6): 1352–370.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence.

Constantine, J., D. Player, T. Silva, K. Hallgreen, M. Grider, and J. Deke. 2009. *An Evaluation of Teachers through Different Routes to Certification: Final Report.* Washington, DC: Mathematica Policy Research/Institute of Education Sciences.

Das, J., S. Dercon, J. Habyarimana, and P. Krishnan. 2007. Teacher Shocks and Student Learning: Evidence from Zambia. *Journal of Human Resources* 42(4): 820–62.

Decker, P.T., D.P. Mayer, and S. Glazerman. 2004. *The Effects of Teach for America on Students: Findings from a National Evaluation.* Princeton, NJ: Mathematica Policy Research, Inc.

Dee, T.S. 2004. Teachers, Race and Student Achievement in a Randomized Experiment. *The Review of Economics and Statistics* 86(1): 195–210.

Dee, T.S. 2007. Teachers and the Gender Gaps in Student Achievement. *Journal of Human Resources* 42(3): 528–54.

Deming, D.J. 2011. Better Schools, Less Crime? *Quarterly Journal of Economics* 126(4): 2063–115.

Dhaliwal, I., E. Duflo, R. Glennerster, and C. Tulloch. 2011. *Comparative Cost-Effectiveness Analysis to Inform Policy in Developing Countries: A General Framework with Applications for Education.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Dobbie, W., and R.G. Fryer. 2009. Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem. NBER Working Paper 15473. Cambridge, MA: National Bureau of Economic Research.

Dobbie, W., and R.G. Fryer. 2011. The Impact of Youth Service on Future Outcomes: Evidence from Teach for America. NBER Working Paper 17402. Cambridge, MA: National Bureau of Economic Research.

Duflo, E., P. Dupas, and M. Kremer. 2007. *Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Duflo, E., P. Dupas, and M. Kremer. 2009. *Additional Resources versus Organizational Changes in Education: Experimental Evidence from Kenya.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Duflo, E., R. Hanna, and S. Ryan. 2010. *Incentives Work: Getting Teachers to Come to School.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Feng, L., D.N. Figlio, and T. Sass. 2010. School Accountability and Teacher Mobility. NBER Working Paper 16070. Cambridge, MA: National Bureau of Economic Research.

Ferraz, C., and B. Bruns. Forthcoming. Incentives to Teach: The Effects of Performance Pay in Brazilian Schools. Washington, DC: World Bank.

Fryer, R.G. 2011. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. NBER Working Paper 16850. Cambridge, MA: National Bureau of Economic Research.

Ganimian, A.J., and E. Vegas. 2011. What Are the Different Profiles of Successful Teacher Policy Systems? SABER-Teachers Background Paper No. 5. Washington, DC: World Bank.

Gibbons, R., and M. Waldman. 1999. Careers in Organizations: Theory and Evidence. In *Handbook of Labor Economics*, Vol. 3, ed. by O. Ashenfelter and D. Card. Amsterdam: Elsevier.

Glazerman, S., and A. Seifullah. 2012. *An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years: Final Report.* Washington, DC: Mathematica Policy Research.

Glazerman, S., E. Isenberg, S. Dolfin, M. Bleeker, A. Johnson, M. Grider, and M. Jacobus. 2010. *Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study.* Washington, DC: Mathematica Policy Research/Institute of Education Sciences.

Glewwe, P., and E. Maïga. 2011. *The Impacts of School Management Reforms in Madagascar: Do the Impacts Vary by Teacher Type?* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Glewwe, P., N. Ilias, and M. Kremer. 2010. Teacher Incentives. *American Economic Journal: Applied Economics* 2(3): 205–27.

Glewwe, P., M. Kremer, and S. Moulin. 2009. Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics* 1(1): 112–35.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz. 2004. Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics* 74: 251–68.

Goodman, S., and L. Turner. 2010. Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program. Paper prepared for the conference on "Merit Pay: Will It Work? Is It Politically Viable?" Program for Education Policy and Governance, Harvard Kennedy School, June 3-4.

Greenwald, B.C. 1986. Adverse Selection in the Labour Market. *Review of Economic Studies* 53: 325–47.

Grissom, J.A., and S. Loeb. 2011. Triangulating Principal Effectiveness: How Perspectives of Parents, Teachers and Assistant Principals Identify the Central Importance of Managerial Skills. *American Educational Research Journal* 48(5): 1–33.

Hanushek, E.A. 2003. The Failure of Input-Based Schooling Policies. *The Economic Journal* 113: F64–F98.

Hanushek, E.A., and L. Woessmann. 2007. *Education Quality and Economic Growth.* Washington, DC: World Bank.

Hanushek, E.A., J.F. Kain, D.M. O'Brien, and S.G. Rivkin. 2005. The Market for Teacher Quality. NBER Working Paper 11154. Cambridge, MA: National Bureau of Economic Research.

He, F., L.L. Linden, and M. MacLeod. 2007. *Helping Teach What Teachers Don't Know: An Assessment of the Pratham English Language Program.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

He, F., L.L. Linden, and M. MacLeod. 2009. *A Better Way to Teach Children to Read? Evidence from a Randomized Controlled Trial.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Hicks, J. 1939. The Foundations of Welfare Economics. *The Economic Journal* 49(196): 696–712.

Holmstom, B. 1979. Moral Hazard and Observability. *Bell Journal of Economics* 9: 74-91.

Holmstrom, B., and P. Milgrom. 1991. Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership and Job Design. *Journal of Law, Economics and Organization* 7: 24–52.

Horng, E.L., D. Klasik, and S. Loeb. 2010. Principal's Time Use and School Effectiveness. *American Journal of Education* 116: 491–523.

Hoxby, C.M. 2000. The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics* 115(4): 1239–285.

Hoxby, C.M., and A. Leigh. 2004. Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States. *American Economic Journal* 93(2): 236–40.

Jackson, C.K. 2013. Match Quality, Worker Productivity and Worker Mobility: Direct Evidence from Teachers. *Review of Economics and Statistics.*

Jackson, C.K., and E. Bruegmann. 2009. Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers. *American Economic Journal: Applied Economics* 1(4): 85–108.

Jacob, B.A. 2010a. The Effect of Employment Protection on Worker Effort: Evidence from Public Schooling. NBER Working Paper 15655. Cambridge, MA: National Bureau of Economic Research.

Jacob, B.A. 2010b. Do Principals Fire the Worst Teachers? NBER Working Paper No. 15715. Cambridge, MA: National Bureau of Economic Research.

Jacob, B.A., and L. Lefgren. 2004a. Remedial Education and Student Achievement: A Regression-Discontinuity Analysis. *Review of Economics and Statistics* 86(1): 226–44.

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

Jacob, B.A., and L. Lefgren. 2005. Principals as Agents: Subjective Performance Measurement in Education. NBER Working Paper No. 11463. Cambridge, MA: National Bureau of Economic Research.

Jacob, B.A., and L. Lefgren. 2007. What Do Parents Value in Education? An Empirical Investigation of Parents' Revealed Preferences for Teachers. *Quarterly Journal of Economics* 122(4): 1603–637.

Jacob, B.A., and S.D. Levitt. 2003. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. *Quarterly Journal of Economics* 118(3): 843–78.

Johnson, W. 1978. A Theory of Job Shopping. *Quarterly Journal of Economics* 92: 261–77.

Kaldor, N. 1939. Welfare Propositions in Economics and Interpersonal Comparisons of Utility. *The Economic Journal* 49(195): 549–52.

Kane, T.J., D.F. McCaffrey, T. Miller, and D.O. Staiger. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Seattle: Bill and Melinda Gates Foundation.

Kane, T.J., and D.O. Staiger. 2010. *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project.* Seattle: Bill and Melinda Gates Foundation.

Kane, T.J., and D.O. Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Seattle: Bill and Melinda Gates Foundation.

Kane, T.J., J.E. Rockoff, and D.O. Staiger. 2006. What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City. NBER Working Paper No. 12155. Cambridge, MA: National Bureau of Economic Research.

Kane, T.J., E.S. Taylor, J.H. Tyler, and A.L. Wooten. 2010. Identifying Effective Classroom Practices Using Student Achievement Data. NBER Working Paper No. 15803. Cambridge, MA: National Bureau of Economic Research.

Kemple, J. J., and C.J. Willner. 2008. *Career Academies: Long-Term Impacts on Labor Market Outcomes, Educational Attainment and Transitions to Adulthood.* New York: MDRC.

Kremer, M.E., D. Chen, P. Glewwe, and S. Moulin. 2001. *Interim Report on a Teacher Incentive Program in Kenya.* Cambridge, MA: Harvard University.

Krueger, A. 1999. Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics* 114: 497–532.

Lang, K., and W.T. Dickens. 1987. Neoclassical and Sociological Perspectives on Segmented Labor Markets. NBER Working Paper No. 2127. Cambridge, MA: National Bureau of Economic Research.

Lassibille, G., J.-P. Tan, C. Jesse, and T.V. Nguyen. 2010. Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions. *The World Bank Economic Review* 24(2): 303–29.

Lavy, V. 2002. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *The Journal of Political Economy* 110(6): 1286–317.

Lavy, V. 2008. Gender Differences in Market Competitiveness in a Real Workplace: Evidence from Performance-Based Pay Tournaments among Teachers. Paper prepared for the conference on "Economic Incentives: Do They Work in Education?" Program for Education Policy and Governance, Harvard Kennedy School, May 16-17.

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

Lavy, V. 2010. Do Differences in School's Instruction Time Explain International Achievement Gaps in Math, Science, and Reading? Evidence from Developed and Developing Countries. NBER Working Paper No. 16227. Cambridge, MA: National Bureau of Economic Research.

Lazear, E.P., and S. Rosen. 1981. Rank-Order Tournaments as Optimum Labor Contracts. *Journal of Political Economy* 89(5): 841–64.

Linden, L.L., C. Herrera, and J. Grossman. 2011. *Achieving Academic Success After School: A Randomized Evaluation of the Higher Achievement Program.* Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Lucas, R.E., and E.C. Prescott. 1974. Equilibrium Search and Unemployment. Journal of Economic Theory. *Journal of Economic Theory* 7(2): 188–209.

McEwan, P., and L. Santibáñez. 2005. Teacher and Principal Incentives in Mexico. In *Incentives to Improve Teaching: Lessons from Latin America*, ed. by E. Vegas. Washington, DC: World Bank.

Metzler, J., and L. Woessmann. 2010. The Impact of Teacher Subject Knowledge on Student Achievement: Evidence from Within-Teacher Within-Student Variation. IZA DP No. 4999. Bonn, Germany: Institute for the Study of Labor (IZA).

Mihaly, K., D.F. McCaffrey, D.O. Staiger, and J.R. Lockwood. 2013. *A Composite Estimator of Effective Teaching.* Seattle: Bill and Melinda Gates Foundation.

Miller, R.T., R.J. Murnane, and J.B. Willett. 2007. Do Teacher Absences Impact Student Achievement? Longitudinal Evidence from One Urban School District. NBER Working Paper 13356. Cambridge, MA: National Bureau of Economic Research.

Mincer, J. 1962. On-the-Job Training: Costs, Returns and Some Implications. *Journal of Political Economy* 70: 50–79.

Mirrlees, J. 1976. The Optimal Structure of Incentives and Authority within an Organization. *Bell Journal of Economics* 7: 105–31.

Mizala, A., and M. Urquiola. 2007. School Markets: The Impact of Information Approximating Schools' Effectiveness. NBER Working Paper 13676. Cambridge, MA: National Bureau of Economic Research.

Montgomery, J. 1991. Equilibrium Wage Dispersion and Interindustry Wage Differentials. *Quarterly Journal of Economics* 106: 163–79.

Muralidharan, K., and V. Sundararaman. 2009. Teacher Performance Pay: Experimental Evidence from India. NBER Working Paper 15323. Cambridge, MA: National Bureau of Economic Research.

Muralidharan, K., and V. Sundararaman. 2010. *Contract Teachers: Experimental Evidence from India.* Washington, DC: World Bank.

Murnane, R.J. 2010. Progress and Puzzles in Educational Policy Research. *Harvard Education Letter* 26(2).

Murnane, R.J., and J.B. Willett. 2011. *Methods Matter: Improving Causal Inference in Educational and Social Science Research.* New York: Oxford University Press.

Myung, J., S. Loeb, and E.L. Horng. 2011. Tapping the Principal Pipeline: Identifying Talent for Future School Leadership in the Absence of Formal Succession Management Programs. *Educational Administration Quarterly* 47(5): 695–727.

Nelson, R.R., and E.S. Phelps. 1966. Investment in Humans, Technological Diffusion and Economic Growth. *The American Economic Review* 56(1/2): 69–75.

Osterman, P. 1984. *Internal Labor Markets.* Cambridge, MA: MIT Press.

Pandey, P., S. Goyal, and V. Sundararaman. 2009. Community Participation in Public Schools: Impact of Information Campaigns in Three Indian States. *Education Economics* 17(3): 355-375.

Papay, J.P., M.R. West, J.B. Fullerton, and T.J. Kane. 2011. Does Practice-Based Teacher Preparation Increase Student Achievement? Early Evidence from the Boston Teacher Residency. NBER Working Paper 17646. Cambridge, MA: National Bureau of Economic Research.

Pissarides, C.A. 1979. Job Matching with State Employment Agencies and Random Search. *The Economic Journal* 89(356): 818–33.

Prendergast, C. 1993. The Role of Promotion in Inducing Specific Human Capital Acquisition. *Quarterly Journal of Economics* 108(2): 523–34

Prendergast, C. 2002. The Tenuous Tradeoff between Risk and Incentives. Journal of Political Economy, 110(5): 1071–102.

Rau, T., and D. Contreras. 2009. Tournaments, Gift Exchanges, and the Effect of Monetary Incentives for Teachers: The Case of Chile. Department of Economics Working Paper 305. Santiago: University of Chile.

Reardon, S.F. 2011. The Widening Academic Gap between the Rich and the Poor: New Evidence and Possible Explanations. In *Whither Opportunity? Rising Inequality, Schools and Children's Life Chances,* ed. by G. Duncan, G. and R.J. Murnane. New York: Russell Sage Foundation and Spencer Foundation.

Reich, M., D.M. Gordon, and R.C. Edwards. 1973. A Theory of Labor Market Segmentation. *American Economic Review* 639(2): 359–65

Rockoff, J.E. 2008. Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City. NBER Working Paper 13868. Cambridge, MA: National Bureau of Economic Research.

Rockoff, J.E., and L.J. Turner. 2010. Short Run Impacts of Accountability on School Quality. *American Economic Journal: Economic Policy* 2(4): 119–47.

Rockoff, J.E., B.A. Jacob, T.J. Kane, and D.O. Staiger. 2008. Can You Recognize an Effective Teacher When You Recruit One? NBER Working Paper 14485. Cambridge, MA: National Bureau of Economic Research.

Rockoff, J.E., D.O. Staiger, T.J. Kane, and E.S. Taylor. 2011. *Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools.* New York: Columbia Business School.

Rosen, S. 1978. Substitution and Division of Labor. *Económica* 45: 235–50.

Ross, S. 1973. The Economic Theory of Agency: The Principal's Problem. *American Economic Review* 63(2): 134–39.

Rouse, C.E., and A.B. Krueger. 2004. Putting Computerized Instruction to the Test: A Randomized Evaluation of a "Scientifically Based" Reading Program. *Economics of Education Review* 23(4): 323–38.

Roy, A. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers*: 235–46.

Sattinger, M. 1975. Comparative Advantage and the Distributions of Earnings and Abilities. *Econometrica* 43: 455–68.

Schultz, T.W. 1963. *The Economic Value of Education.* New York, NY: Columbia University Press.

Shapiro, C., and J. Stiglitz. 1984. Equilibrium Unemployment as a Worker Discipline Device. *American Economic Review* 74: 433–44.

Shavell, S. 1979. Risk Sharing and Incentives in the Principal and Agent Relationship, *Bell Journal of Economics* 10: 55–73.

Spence, A.M. 1974. *Market Signaling: Informational Transfer in Hiring and Related Screening Processes.* Cambridge: Harvard University Press.

Springer, M.G., D. Ballou, L. Hamilton, V.-N. Le, J.R. Lockwood, D. McCaffrey, M. Pepper, and B.M. Stecher. 2010. *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching.* Pittsburgh: The RAND Corporation.

Steele, J.L., R.J. Murnane, and J.B. Willett. 2010. Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence from California. *Journal of Policy Analysis and Management* 29(3): 451–58.

Stiglitz, J. 1975. The Theory of Screening, Education and the Distribution of Income. *American Economic Review* 65: 283–300.

Suryadarma, D., A. Suryahadi, S. Sumarto, and F.H. Rogers. 2006. Improving Student Performance in Public Primary Schools in Developing Countries: Evidence from Indonesia. *Education Economics* 14(4): 401-429.

Taylor, E.S., and J.H. Tyler. 2011. The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-Career Teachers. NBER Working Paper 16877. Cambridge, MA: National Bureau of Economic Research.

Terviö, M. 2003. *Mediocrity in Talent Markets.* Berkeley, CA: Haas School of Business, University of California, Berkeley.

Tyler, J.H. 2011. If You Build It, Will They Come? Teacher Use of Student Performance Data on a Web-Based Tool. NBER Working Paper 17486. Cambridge, MA: National Bureau of Economic Research.

Urquiola, M., and E. Verhoogen. 2009. Class-Size Caps, Sorting and the Regression Discontinuity Design. *American Economic Review* 99(1): 179–215.

Vegas, E., A.J. Ganimian, N. Goldstein, A. Paglayan, S. Loeb, and P. Romaguera. 2011. What Are Teacher Policy Goals, How Can Education Systems Reach Them and How Will We Know When They Do? SABER-Teachers Background Paper No. 2. Washington, DC: World Bank.

Williamson, O. 1975. Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange. *Bell Journal of Economics* 16: 250–80.