Teacher Evaluation and Training[†]

Eric S. Taylor
Harvard University and NBER

October 2022

Evaluation and training are important features of the employment relationship between teachers and the schools they work for. The first feature, evaluation, involves performance measures and often performance incentives linked to those measures, like bonuses or the threat of dismissal. This chapter reviews research on whether and how evaluation and incentives change teaching, including unintended effects. Potential mechanisms include changes in a teacher's effort or skills, or changes in the composition of the teacher workforce through selection. Many (quasi-)experiments document increases in the measures used to determine rewards or consequences for teachers, but it is less clear whether those increases represent improvements in student learning or welfare. Research on the second feature, training, typically focuses on formal training programs, where evidence of benefits is inconsistent at best. This chapter reviews evidence on both formal training, as well as informal ways in which teachers appear to learn new skills at work.

JEL No. I28, J24, J33, J45, M52, M53

Keywords: Performance evaluation, Performance incentives, Training, Skill development

[†] Taylor: eric_taylor@harvard.edu, Gutman Library 469, 6 Appian Way, Cambridge, MA 02138. This chapter was prepared for the *Handbook of the Economics of Education*. The editors, in particular Eric Hanushek, provided helpful feedback on earlier drafts. Fei Yuan provided excellent research assistance.

Teachers work for schools. That employment relationship is central to the economics of schooling. Teacher labor costs dominate school budgets, and teachers' contributions to student achievement dominate the variation in achievement created by schools. As employers, schools make a variety of personnel decisions—hiring and firing, performance evaluation, incentives, training, job design, etc.—intending to improve teachers' contributions. But schools' personnel choices are constrained by how teachers' own choices will respond.[1] These employer-employee interactions can affect the success of schools through both changes in individual teachers' skills or effort, and changes in the composition of the teacher workforce. A growing empirical literature documents the effects of school personnel strategies and teacher responses. This chapter reviews the literature on teacher evaluation and training.[2]

Evaluation measures and performance incentives are familiar features in many occupations and sectors. Teachers and schools are no different. In the education sector, the category "teacher evaluation" has come to encompass a variety of policies, programs, and proposals. All involve measuring teacher job performance, and section 1 provides a primer on common performance measures. Most also involve performance incentives linked to those measures. This chapter includes many examples, from pay-for-performance bonuses to the threat of dismissal for low performance.

Underlying the practical features are a variety of rationales for how evaluation might improve (average) teacher performance. These rationales involve three broad categories of mechanisms: changes in teachers' effort,

[1] For simplicity I use "schools" or "school systems" to refer to the employer which is making personnel decisions. In practice, because most schools are publicly funded, teacher personnel decisions are made by a variety of actors: school principals or head teachers, superintendents, school boards, legislatures, ministers of education, etc.

[2] For other reviews see Neal (2011) in this Handbook, and Jackson, Rockoff, and Staiger (2014). For general reviews of performance incentives see Oyer and Schaefer (2011), Gibbons and Roberts (2013), Lazear and Oyer (2013).

changes in teachers' skills, and changes in the composition of the teacher workforce through (self-)selection. The most common rational, especially in the economics literature, is an agency-theory rationale which emphasizes changes in effort. Briefly, performance incentives make a teacher's job utility a (stronger) function of her school's success, and thus induce more effort at work or a better allocation of effort across tasks.

As this chapter demonstrates, it is no longer difficult to find examples—convincing (quasi-)experimental examples—where teacher performance incentives cause improvements in achievement test scores or other outcomes. Performance incentives can change teacher effort. Much less clear is whether teacher performance incentives improve student welfare. For example, a teacher may keep total effort fixed but reallocate effort: decreasing effort on job tasks which are unmeasured or non-incentivized, and increasing effort on tasks which increase the performance measures used to determine rewards or punishments for the teacher. This is the multitask problem originally described by Holmstrom and Milgrom (1991). Section 2 focuses on studies which test for this kind of reallocation and other potential distortions in teacher effort. Section 2 also covers other features which may weaken or strengthen teachers' effort response, including noisy performance measures, team incentives, the availability of inputs which complement teacher effort, and others.

Teacher selection is a second common rationale for evaluation. But there are two distinct mechanisms: selection of teachers by schools based on measured performance, and teacher self-selection into or out of teaching in response to evaluation incentives. These topics are the focus of section 3. One strand of literature considers optimal dismissal rules for schools. That analysis suggests dismissals are limited more by costs than by imperfect performance measures. The costs include both higher salaries to compensate for greater employment uncertainty, and short-run losses in student achievement because a novice

replaces the dismissed teacher. However, in practice, school systems still rarely dismiss teachers based on performance, which limits empirical study of the effects of selection by schools.

Evaluation may cause individuals to self-select into or out of teaching. Self-selection assumes (prospective) teachers have private information, specifically about their potential scores on the incentivized performance measures. Empirical evidence remains relatively scarce. However, recent contributions suggest one potentially important pattern: Prospective teachers may have little relevant information, private or otherwise, before they begin working, and thus little scope for self-selection effects at entry. Teachers with some work experience, by contrast, appear more likely to have private information on which to self-select.

A third rationale is that evaluation can improve job performance through causal effects on teachers' job skills. Potential mechanisms arise through the costs and returns to teachers' investments in skills. First, performance measurement can reduce the costs of skill investments. For example, classroom observation evaluations typically measure performance in many separate teaching tasks. Figure 1 shows observation rubric examples for asking questions and responding to student misbehavior. By comparing her own ratings and feedback across different tasks, a teacher receives direction on where to focus her effort in skill improvement. Second, performance incentives can increase the returns to skill investments. If the teacher expects repeated evaluation and performance incentives over time, then the potential stream of future rewards creates an incentive for the teacher to improve her skills. Section 4 discusses evidence that may differentiate between skill investment mechanisms and the effort mechanisms typical of agency-theory rationales. For example, if skill investment, then evaluation's effects on performance can persist after the evaluation measures or incentives end because skills persist.

Formal teacher training programs provide an important comparison to teacher evaluation. Both are personnel strategies intended to improve teacher job performance, and, for a school system, one may be the opportunity cost of the other. Section 5 reviews empirical evidence on formal teacher training programs, both pre-service programs which prepare prospective teachers and training programs for working teachers. There remains little causal evidence that formal training programs improve teacher performance. But formal training is not the only opportunity for teachers to learn new skills. Section 6 focuses on informal training, including learning through experience and learning from other teachers. For example, one quite consistent result in the literature is that performance improves steeply over the first few years of teaching.

This chapter discusses empirical examples from the United States and other high-income countries, alongside examples from low- and middle-income countries. The economic themes are common across these settings, and the chapter is organized by those themes. However, there are empirical patterns that differ across settings. For example, several experimental studies have documented positive effects of pay for performance, but those studies occurred outside the United States. Evidence from the U.S. is much less convincing of pay for performance benefits. This pattern has persisted as the literature has grown (Neal 2011, Jackson, Rockoff, and Staiger 2014), suggesting some caution in applying lessons from other countries to the U.S., or vice versa. These differences in effect estimates could be due to differences in incentive design features, for example, whether structured as a tournament or not, the size of the bonus, the type of performance measure, etc. Alternatively, the differences could be due to more fundamental differences across settings, for example, the baseline level of teacher effort and thus where the average teacher is on her cost of effort curve.

Teachers' contributions to their students' outcomes vary widely, and those differences between teachers are the broader context for this chapter's personnel-

focused topics. In the case of teachers' contributions to students' academic achievement, one standard deviation in the teacher contribution distribution is typically 10-20% of the total variability in student test scores. While many estimates of teacher "value-added" to achievement scores come from the United States, others have found similarly-sized teacher effects in a variety of lower- and higher-income countries (see Bacher-Hicks and Koedel's review of these estimates in volume 6 of this Handbook). Moreover, a teacher's contributions to her young students' academic achievement can further affect those students' success as adults, as measured by things like college going and labor market earnings (Chetty, Friedman, and Rockoff 2014b). Beyond academic achievement, teachers also likely differ in their contributions to students' social and emotional development, and those contributions can also have lasting effects in adulthood (Jackson 2018). In short, there are large potential gains from improving teacher performance through personnel strategies like evaluation and training.

1. PERFORMANCE MEASURES

Schools use several performance measures to evaluate teachers. Some measure output, including scores based on student test scores or other student outcomes. Others measure inputs, like ratings from classroom observations of teaching practices. The same types of performance measures also appear in empirical research, even when not part of the evaluation or incentive scheme under study. This section is a brief introduction to the most common performance measures for teachers.

1.1. Value-Added Measures

Perhaps the most intuitive component of teacher job performance is the teacher's contribution to what her students learn. Measures of an individual teacher's contribution to academic achievement are commonly called "teacher value-added" scores. Economists began working in the 1970s on estimating a

teacher's contribution to student test scores (Hanushek 1971, Murnane 1975, Summers and Wolfe 1977; see also Seyfert and Tyndal 1934). That work has acerated rapidly in recent decades, mainly using data from schools in the United States but not exclusively. In this Handbook, Bacher-Hicks and Koedel (in-press) review the literature on estimating and interpreting teacher value-added measures, including a summary of estimates from a variety of lower- and higher-income countries.[3]

The use of value-added scores by schools for teacher evaluation has also expanded in recent decades. The Tennessee Department of Education was the first to distribute reports with value-added scores for individual teachers, beginning in the early 1990s (Sanders and Horn 1998). Widespread use of teacher value-added scores by schools in the United States began later in the 2010s. Alongside a growing research base, the Obama administration's education policy efforts emphasized teacher evaluation, including individual value-added measures. The Obama administration explicitly incentivized states and school districts to adopt new teacher evaluation measures, through the Race to the Top grant competition and NCLB flexibility. By 2015, most all school systems in the U.S. were using value-added scores for teacher evaluation (Steinberg and Donaldson 2016, Ross and Walch 2019).[4] However, the Obama-era policies were supplanted in 2015 by the reauthorization of the Elementary and Secondary Education Act (also known as the Every Student Succeeds Act). A number of states subsequently revised their teacher evaluation rules, and, as of 2019, only two-thirds of U.S. states still required the use of value-added. While widely used, value-added scores are

[3] Estimates from outside of the United States come from Ecuador, England, Ethiopia, India, Japan, Pakistan, Uganda, and Vietnam; citations are provided in Bacher-Hicks and Koedel (in-press). Other reviews include Hanushek and Rivkin (2010), Jackson, Rockoff, and Staiger (2014), and Koedel, Mihaly, and Rockoff (2015).

[4] Forty-two states and D.C. required that teacher evaluation include value-added scores. California did not have a state requirement, but many California districts did including its largest district Los Angeles Unified. For those interested in more details, the National Council on Teacher Quality (NCTQ) maintains a longitudinal data base of state-level teacher policies in the U.S.

typically given much less weight than classroom observation scores in a teacher's overall evaluation (Steinberg and Donaldson 2016).

In the typical estimates, one standard deviation in the teacher performance distribution is 0.10-0.20 student test-score standard deviations (σ = student standard deviations). An average class of students with a 75th-percentile teacher will score 0.13-0.27σ higher than the same class would with a 25th-percentile teacher. Teacher value-added scores are typically scaled so that the mean is zero. A negative value-added score does not necessarily imply a loss of student knowledge, but rather slower achievement growth than in the class of the average teacher.

Value-added scores are causal claims. The causal relationship of interest is the effect of a given teacher on her students' achievement test scores. The main identification challenge is the assignment of students to teachers. The typical identifying assumption is that student-teacher assignment is ignorable conditional on observables, where the key conditioning variable is the student's prior test score.[5] Several (quasi-)experimental results support the plausibility of this particular conditional independence assumption (Kane and Staiger 2008, Kane et al. 2013, Chetty, Friedman, and Rockoff 2014a, Bacher-Hicks, Kane, and Staiger 2014, Bacher-Hicks et al. 2019), though Rothstein (2010, 2017) provides a skeptical perspective. On balance, value-added measures are today generally taken as unbiased estimates of teachers' causal effects on student achievement scores.

[5] For example, in Chetty, Friedman, and Rockoff (2014a) estimation begins with a regression where the dependent variable is a student's test score at the end of school year t, and the key independent variable is the student's score from year $t-1$. Residuals from that regression are then used to form value-added estimates. Nearly all other value-added score approaches are based on a similar lagged-score regression specification (Bacher-Hicks and Koedel i-press). Moreover, this approach—lagged-score specification and conditional independence assumption—is commonly used in (quasi-)experiments on teachers where the outcome is student test scores, even when the researchers do not explicitly estimate value-added scores. This chapter includes many examples.

Unbiasedness is a useful property when value-added scores are used in personnel management strategies, like probationary screening or performance incentives. One cost is that, because of the lagged-test-score data requirement, school systems often have value-added scores for fewer than half of teachers.

Even if unbiased, individual teacher value-added estimates can be quite noisy. A statistic often cited is the correlation between a teacher's value-added scores in two consecutive years; estimates from several studies range from 0.20 to 0.65 (Bacher-Hicks and Koedel in-press). However, a correlation less than 1 is partly noise but also partly changes in true performance.[6] Motivated by the measurement error component, researchers often "shrink" value-added estimates. Each teacher's initial score is multiplied by the estimated proportion of signal variance in that score, thus making the final score closer to the mean score of zero. Noisier scores have less signal and thus get shrunk more. This shrinking is well-motivated when the value-added score will subsequently be used as an explanatory variable itself, but the tradeoffs between shrunk and unshrunk are not always as clear in other uses.

Measurement error is an important property to consider when value-added scores are used by schools in evaluation and reward schemes. Measurement error represents uncontrollable risk to the teacher and thus weakens incentives. A common practice by schools is to average a teacher's scores across multiple school years, which reduces noise, but weakens incentives in other ways. School systems also often use "shrunken" value-added scores, though without a clear statistical justification. Moreover, schools generally do not report information about a score's error in a teacher's evaluation (for an exception see Rockoff et al. 2012).

[6] If teacher performance is fixed over time, at least over two-year periods, then 0.20-0.65 is the reliability of value-added scores.

Setting aside the practical constraints, schools might prefer to evaluate a teacher based on her contributions to the long-run success of her students, like success in college or the workforce. But test-based value-added scores are a much more feasible option for personnel management. Encouraging evidence shows teachers with higher value-added scores also positively affect long-run outcomes like college going and earnings as adults (Chamberlain 2013, Chetty, Friedman, and Rockoff 2014b, Jackson 2018). Those long-run effects occur even though teachers' measured effects on test scores can fade out over time (Jacob, Lefgren, and Sims 2010). In short, test-based value-added scores are informative proxies for teacher performance more generally. Though Jackson's (2018) results suggest teacher contributions not captured by achievement tests, like student behavioral improvement, may be more strongly related to long-run outcomes.

Jackson (2018) and others have begun building evidence on how teachers contribute to other near-term outcomes, especially students' social and behavioral skills (see also Kraft and Blazar 2017, Kraft 2019, Liu and Loeb 2021, Mulhern and Opper 2021, Petek and Pope 2021). These new measures often adopt the estimation machinery developed for test-based value-added scores, though it remains unclear whether the unbiasedness or other measurement properties will be similar for non-test outcomes.

1.2. Student Proficiency Rates

A second measure—widely used by schools for evaluating teachers—is the percent of students who pass the year-end test. In schools this measure is commonly known as the "percent proficient" or "proficiency rate."[7] The percent proficient measure is most often used to evaluate teams of teachers—a

[7] Test scores are typically divided into more categories than just "pass" and "fail." A common categorization is to divide the test score distribution into "achievement levels" like advanced, proficient, basic. The consequential number for teachers is nearly always the percent of student who score proficient or higher, thus "percent proficient" and "percent passing" are used interchangeably.

department, grade level, or school—not individual teachers per se. For many years public schools in the United States were required to use the percent proficient measure, under the federal regulations known as "No Child Left Behind" (NCLB) adopted in the early 2000s (Dee and Jacob 2011). Under NCLB schools were expected to increase their percent proficient measures year after year, with negative consequences for repeatedly missing targets. While made famous by NCLB, percent passing measures had been in use since at least since the late 1980s (Jacob 2005, Deming et al. 2016, and others study example cases). In 2015, NCLB was replaced by the "Every Student Succeeds Act" (ESSA). Percent proficient remains a prominent performance measure in the ESSA era, though somewhat less central than during the NCLB years.

A basic feature of any percent passing measure is the cutoff score which separates passing from failing. Koretz (2008) provides an overview of the different methods used for setting that cutoff. In short, the cutoff is set by the judgment of a panel, not by a statistical or algorithmic process. The panel review the test item by item, informed partly by psychometric data like item difficulty, but mainly applying their own judgment to identify items which differentiate proficient from not proficient. Panel members do not necessarily have experience with teaching or testing. Empirically, what a student must know to be considered "proficient" can vary greatly across tests and jurisdictions (Koretz 2008, Reardon, Kalogrides, and Ho 2021), though the cutoffs are typically stable across years for a given test in a given jurisdiction.

A "percent proficient" is just a simple mean for a sample of students, not an estimate of their teacher's (or school's) contribution to who passed the test. That simplicity makes percent proficient quite unlike value-added estimates. Quite like value-added scores, however, measurement error in "percent proficient" is also a concern when using the score for evaluation (Kane and Staiger 2002).

1.3. Classroom Observation Ratings

Schools have long evaluated job performance by observing teachers at work in the classroom (see for example Barr 1928). Today, in many schools, detailed scoring rubrics guide the observations and structure the resulting ratings. Observation ratings measure inputs—the teacher's skills and effort in certain job tasks—not outputs like student learning. For example, many rubrics ask the observer to score the nature and frequency of questions the teacher asks her students, but observers are not asked to assess whether the questions generated student learning.[8]

Contemporary observation rubrics produce dozens of separate item-level scores for each teacher, covering a range of job tasks. Those many item scores are then (typically) reduced to a single scalar rating. Figure 1a shows two of the 32 items on the Cincinnati Public Schools rubric, which was central to the quasi-experiment in Taylor and Tyler (2012). The evaluator scores each item separately, after each observation visit, by matching the behavior she observed from the teacher (and students) to the text descriptions under each of the four possible ratings. That process leaves room for some degree of subjectivity. The ratings are usually given normative labels, as in this example's "distinguished," "proficient," "basic," and "unsatisfactory." Cincinnati's rubric is a version of the Framework for Teaching (FFT; Danielson 1996) which has been widely used and adapted by school systems for teacher evaluation. A second common rubric is the Classroom Assessment Scoring System (CLASS; Pianta, LaParo, and Hamre 2008), for example in Araujo and coauthors' work in Ecuador (2016). Figure 1b shows one of the ten items scored with CLASS. Comparing Figures 1a and 1b illustrates the

[8] Observing teachers is often a component of school inspection programs, for example, in England (Allen and Burgess 2012, Hussain 2015) and the Netherlands (Luginbuhl, Webbink, and De Wolf 2009). In many ways, school inspections are a form of team evaluation for teachers. Observations of teachers are typically designed to serve that team evaluation purpose rather than to make inferences about individuals.

main ways rubrics vary in general: the number of scored items, how tasks are grouped together or not, detail in the text descriptions, different labels for ratings, etc. For a detailed, empirical comparison of rubrics read work generated from the Measures of Effective Teaching project (Kane and Staiger 2012, see also Gill et al. 2016). Similar to value-added scores, this style of rubric first appeared in the 1980s and 90s, but wide-spread use by U.S. schools started in the Obama-era (Kennedy 2010, Donaldson and Papay 2015). In the typical U.S. state, classroom observation scores are given 50 percent or more weight in determining the teacher's overall evaluation (Steinberg and Donaldson 2016). In about half of states teachers are observed annually, and in some states multiple times per year. The use of classroom observation measures in the U.S. has declined somewhat since ESSA in 2015, but less so than the declines in the use of value-added scores (Ross and Walsh 2019).

When used in a school system's evaluation program, the observer is often the teacher's supervisor—the school principal or some other leader at the school—though some school systems employ specialized observers (for examples of the latter see Taylor and Tyler 2012, Dee and Wyckoff 2015). Each scored observation can be as brief as 10-15 minutes in the classroom, and the number of observations typically ranges between one and six in a school year.

One common criticism of classroom observations is that few teachers receive low ratings. Leniency bias is relevant to many uses of teacher ratings, though the criticisms are often overstated, and leniency bias is common in many occupations beyond teaching (Prendergast 1999). Low scores are uncommon. On the typical rating scale of 1-4 or 1-5, often fewer than 5-10% of teachers score 1-2 for a given task (synonymously, rubric item) in a given observer visit (Ho and Kane 2013, Kraft, Papay, and Chi 2020, Taylor 2022). However, each teacher is rated on many items in multiple visits, and averaging across items reveals greater variation in composite performance. These overall average scores, for a typical

sample of teachers, often have a mean around 3.5 on a 1-4 scale, and a standard deviation around 0.5 or half a rubric scale point, or a coefficient of variation of about 0.15 (Kane et al. 2011, Papay et al. 2020, Bell et al. 2022). Yet, in practice a teacher's final performance measure is typically an integer rating which discards much of the available variation. Evidence often cited in leniency bias criticisms is based on these final integer ratings (e.g., Weisberg et al. 2009, New York Times 2013, Kraft and Gilmour 2017). Observation ratings are somewhat less skewed when trained external evaluators score observations, but very low ratings are still uncommon (Kane and Staiger 2012).

Other properties of observation scores are also relevant to their use in performance evaluation. The first is measurement error and other unintended variation in observation scores. Consider the average score, over many item ratings, but all from a single observation visit. Across a variety of rubrics, Kane and Staiger (2012) report that 14-37% of the variation in a single observation score is signal reflecting persistent differences between teachers. Reliability of 0.14-0.37 is lower than the 0.20-0.65 estimate for value-added scores. Kane and Staiger estimate that to reach a reliability of 0.65 would require four observation visits scored by four different evaluators. Though adding evaluators reduces noise faster than additional observation visits by the same evaluator (Ho and Kane 2013). Beyond the signal, variation in observation scores can also reflect the students in the class (Steinberg and Garrett 2016, Campbell and Ronfeldt 2018), as well as the relationship between teacher and evaluator (Ho and Kane 2013, Chi in-press, Grissom and Bartanen 2022). In contrast to value-added scores, the literature includes little about the (un)biasedness of observation scores (Bacher-Hicks et al. 2019). Observation scores are typically not adjusted for measurement error, student characteristics, or evaluator.

A teacher's scores across rubric items (teaching tasks) tend to be strongly correlated. Pairwise correlations across items tend to be between 0.50-0.80, with

the first principal component or factor of item scores explaining 50-90% of the total variation (Kane and Staiger 2012, Kane et al. 2011, Aucejo et al. 2022, Bell et al. 2022, Burgess et al. 2022). Teacher average observation scores are typically correlated between 0.10-0.30 with value-added scores (Kane and Staiger 2012). Despite the detailed rubrics, that correlation is not meaningfully different from the correlation between value-added scores and more subjective ratings from supervisors (Jacob and Lefgren 2008, Rockoff and Speroni 2010, Rockoff et al. 2012).

1.4. Other Performance Measures

Teacher performance measures based on student test scores and classroom observations are widely used in academic research and personnel management. Several other performance measures appear in the literature less frequently. First, beyond value added and percent proficient, student assessments also appear in more ad-hoc ways. One example is a test-based goal: a goal set by the teacher herself, under some constraints, and defined in terms of student performance on a future test or other assessment (Donaldson and Papay 2015). Teachers in the Denver Public Schools received a one percent bonus for meeting their test-based goal (Goldhaber and Walch 2012). Second, estimates of teachers' "value-added" style contributions to student behavioral outcomes, as mentioned already (Jackson 2018, Liu and Loeb 2021, Mulhern and Opper 2021). Third, some school systems now ask students to rate their teachers, using surveys like the one used in the Methods of Effective Teaching project (MET; Kane and Cantrell 2010). Students are asked about items like "In this class, the teacher accepts nothing less than our full effort," one of three dozen such items. In the MET project fewer than half of secondary school students agreed with this statement in bottom-quartile classes, and more than 80% agreed in top-quartile classes. In the U.S., only a handful of states require the use of student surveys; most leave the decision up to individual school districts (Ross and Walsh 2019). A fourth performance measure, relevant

in some settings, is how frequently the teacher is absent from work (Banerjee and Duflo 2006, Chaudhury et al. 2006, Duflo, Hanna, and Ryan 2012).

2. PERFORMANCE INCENTIVES AND EFFORT

A variety of rationales for "teacher evaluation" each emphasize different mechanisms by which performance measures and incentives might improve teacher performance. These rationales and mechanisms are applications of well-known ideas in labor and personnel economics. Yet, while the empirical literature on these topics continues to grow, that growth has been uneven across the different rationales.

The most common rational for evaluation and incentives is an agency-theory rational: Performance incentives tie an employee's (teacher's) utility to their employer's (school system's) success, and thus induce more effort at work or a better allocation of effort across tasks. The mechanism for improvement is a change in teacher effort, where "effort" is shorthand for things like the teacher's attention, time, and decisions. Let $h(e)$ be the teacher's contribution to her school's success; a function which is increasing in teacher effort, e. To simplify our discussion, assume $h(e)$ is the teacher's contribution to her students' achievement growth, though the framework can be used with a broad range of contributions. The teacher makes decisions at work, e, to maximize her job-related utility:

$$\max_{e} \quad u = w + v[h(e)] - c(e). \tag{1}$$

The disutility cost of effort, $c(e)$, is increasing and convex in e. The teacher receives a salary, w, which is unaffected by the effort decision. A conventional starting point, for the agency-theory rationale, is to assume $v[h(e)] = 0$ or at least that $\partial u / \partial h = 0$ beyond some particular level of effort. That is, the teacher's utility does not change if her contribution to student achievement rises or falls at

the margin. The basic prediction, then, is that a school system could increase $h(e)$ by introducing some performance incentive, such that $\partial u/\partial h > 0$. For example, a cash bonus based on teacher value-added scores, or performance-based requirements for earning employment protections like tenure.

This section reviews empirical tests of that basic prediction. That prediction, however, is (potentially) attenuated by three features of the teacher case. These features also complicate claims about the efficiency of performance incentives in education, and thus often become arguments against teacher evaluation in policy debates. First, teachers are "motivated agents," at least to some degree (Dixit 2002). That is, the teacher's utility already depends on $h(e)$, even absent any performance incentives offered by the school. For motivated agents, in teaching or any occupation, the potential response to a new incentive may be smaller but not zero. The reason is that even teachers dislike their jobs at the margin. The disutility of effort, $c(e)$, is convex and thus at some point the teacher will choose time for her family or leisure over more time for her students. Adding some new reward for $h(e)$ will shift that switching point out further in favor of students and raise achievement. How much it shifts out will depend on where the teacher's pre-incentive optimal effort is on the cost curve.

Second, performance measures for teachers are often quite noisy, which weakens the effect of performance incentive schemes. Across occupations, employers generally do not observe $h(e)$ but rather only some measure of it, $h^* = h(e) + \eta$, where η is an unobserved random shock. In practice, performance incentives are linked to h^*. From the employee's perspective η is uncontrollable risk, reducing the expected reward for giving greater effort. This gives rise to the well-known tradeoff between risk and incentives. The most salient example of η in the teacher literature is measurement error in value-added scores, h^*, which proxy for contributions to achievement, $h(e)$. As discussed in section 1,

measurement error can be a substantial contributor to variation in teacher value-added scores. Moreover, the risk-incentive tradeoff may be steeper in the teacher case; the typical teacher contract, with relatively-low pay but relatively-high job security, likely selects for individuals with greater risk aversion (see Dohmen and Falk 2010).[9]

Third, performance incentives may change teacher effort in unintended ways. Economists have described a variety of empirical examples, which I review in this section. The first type is conceptually straightforward: effort which improves the teacher's performance measure, h^*, but has little or no effect on the teacher's contribution to student achievement, h. The simplest example is cheating on student tests. The second type arises from the multitask nature of the teaching job. Imagine $h(e) = h_1(e_1) + h_2(e_2)$. In their seminal paper on the multitask problem, Holmstrom and Milgrom (1991) use the example of teachers, with h_1 testable basic knowledge and h_2 untestable higher-ordered thinking. Introducing incentives based on student test scores raises h_1, likely at the expense of reducing h_2. However, "motivated agents" would mute this kind of multitask distortion, assuming teachers intrinsically care about both h_1 and h_2. Neal (2011) provides a particularly useful description of the multitask problem in the case of teachers.

Do performance incentives change teacher effort and performance? It is not difficult to find examples—well-identified (quasi-)experimental examples—documenting increases in student achievement test scores, and other improvements, caused by teacher performance incentives. The discussion in this

[9] The typical and reasonable assumption is that effort, e, is not observable and thus cannot be the basis of any contract. Occasionally some have argued that classroom observation scores measure teacher effort; not just energy and time generally, but also skilled allocation across teaching tasks. Nevertheless, as discussed in section 1, observation scores themselves have measurement error over even the constructs they attempt to measure; and the effort and actions classroom observations attempt to measure only partially explain between-teacher differences in student achievement outcomes.

chapter includes many examples from a variety of settings, as listed in table 1. Though, even in these examples, there are often reasons to be cautious about interpreting the positive effect estimates as good for students, and there are also examples which find no effects.

The effects of teacher incentives on student outcomes, now documented in many settings, imply that performance incentives change teacher effort. The nature of that change is much less clear; direct measures of effort are scarce. Moreover, the nature of change in teacher effort has implications for whether we should interpret positive treatment effect estimates as evidence of welfare improvements for students. The remainder of this section reviews empirical studies of teacher performance incentives, beginning with studies which compare incentivized and non-incentivized student outcomes.

2.1. Incentivized and Non-Incentivized Tasks

Claims about the benefits of teacher performance incentives are complicated by the multitask nature of teaching jobs. Performance incentives often favor a subset of the teacher's job responsibilities—for example, students' math and reading skills—over other responsibilities—science skills or social-emotional development. In response to those incentives, the teacher may keep her total effort fixed, but reallocate effort away from non-incentivized tasks to incentivized tasks. That reallocation may well improve the performance measures which determine rewards or sanctions, but at the cost of poorer performance and outcomes in other job responsibilities. A growing empirical literature tests for these potential multitask distortions in teacher effort.

2.1.1. Different Academic Subjects

Consider a straightforward example to begin: reallocating effort from one academic subject area to another. In practice, it is common to have teacher rewards (or consequences) determined by student test scores in math and language, but not subjects like science and social studies, even when one teacher

(or school) is responsible for all subjects. Forgone achievement in non-incentivized subjects is an important potential cost.

Muralidharan and Sundararaman (2011) study an experiment in Andhra Pradesh, India where teachers earned cash bonuses for student scores in math and language. Test scores were scaled in percentage points, and bonuses were based on the average change in scores across students, $\overline{\Delta}$. If $\overline{\Delta} > 5$, then teachers received a bonus of Rs. $500 * (\overline{\Delta} - 5)$. The average bonus was about 3% of average annual salary. Over a two-year period, the incentive treatment increased student scores by 0.28σ in math and 0.17σ in language (σ = student standard deviations). The first of many examples of the improvement in teacher performance measures that can be generated by new performance incentives. But test scores also increased by 0.11σ in science and 0.18σ in social studies—meaningful gains even though neither science nor social studies tests affected teacher bonuses.

That pattern of results in Andhra Pradesh does not necessarily rule out some reallocation of effort away from science and social studies and to math and language. Reallocation may occur but not be sufficient to create losses in non-incentivized subjects. When asked in surveys, teachers in other studies have self-reported that they do reallocate class time to incentivized subjects (for example, Koretz and Barron 1998). Nevertheless, even if there was reallocation in Andhra Pradesh, students were better off in all four subjects when their teachers were incentivized in two subjects. One potential mechanism is that improved math and language skills helped students learn more science and social studies. Such spillovers could be evidence that math and reading test score gains reflect real gains in math and reading skills relevant to students' mastery of other subjects. Or the spillovers could indicate that students learned skills for taking tests, a topic discussed in section 2.2.

Several other studies have also tested for effects on non-incentivized subjects. None of the empirical examples show clear losses in these other subjects, and sometimes there are gains in the non-incentivized subjects as there were in Andhra Pradesh. Still, research testing for cross-subject distortions remains relatively scarce.

Mbiti and coauthors (2019) conducted a field experiment in Tanzania, where teachers in randomly-assigned schools were eligible for cash bonuses. Teachers received a bonus for each student in their class who passed government exams in math, Kiswahili, and English. For the average teacher the maximum potential award was 125% of one month's salary. At the end of two years, Mbiti and coauthors estimated the effects on student achievement using a low-stakes exam separate from the government exams which determined bonuses. The new teacher incentives raised math achievement somewhat, 0.07σ (standard error 0.04), but had little effect on Kiswahili, 0.01σ, or English, 0.00σ.[10] Nor was there a change in a non-incentivized subject: science, -0.01σ. However, the Tanzania experiment included two additional treatment conditions. A second in which randomly-assigned schools received an unconditional cash grant, which could be spent on anything except teacher compensation. And a third that combined both the bonuses and the grant. While teacher incentives alone had little effect on student achievement, the combination of incentives and grants created meaningful gains. Student scores on the low-stakes exams increased by 0.18-0.21σ in math, Kiswahili, and English—the incentivized subjects—but also increased by 0.09σ in science. As in Andhra Pradesh, students in Tanzania were better off in all subjects. These results suggest some complementarities between incentives and other school inputs which we return to in section 2.7. Nevertheless, whether

[10] There is evidence that teachers responded to the new incentives. Treatment significantly raised scores on the government exams which determined teacher bonuses: between 0.12-0.17σ across the three subjects. We return to the contrast between low- and high-stakes tests of the same subjects in section 2.2, which discusses a different dimension of potential distortions in effort.

incentives alone or incentives with grants, the results from Tanzania are not consistent with strong distortions across subject areas.

The teacher incentives literature includes examples of both rewards and sanctions. The Andhra Pradesh and Tanzania examples involve cash bonuses for good performance. Many schools and teachers are threatened with sanctions for poor performance. The most well-known examples of the latter are "school accountability" programs in the United States, especially under federal No Child Left Behind (NCLB) rules in the 2000s. But individual states and districts began similar programs in the 1990s. These programs are team evaluations, with performance measures at the school level, and sanctions that range from writing a plan for improvement to dismissal of the schools' teachers. The next examples are cases of these school accountability incentives.

The Chicago Public Schools began test-based accountability for schools in the late 1990s. Teachers and school administrators were threatened with dismissal or reassignment if too many of the school's students scored below national norms in math and reading on the Iowa Test of Basic Skills (ITBS). Jacob (2005) shows that math and reading ITBS scores improved as a result of these incentives. Chicago's average ITBS scores also improved in science and social studies, although the math and reading gains were 2-4 times as larger. However, Jacob (2005) does find some evidence of reallocation for lower-achieving students—from science and social studies to math and reading. The Chicago performance measure, based on whether students score above or below a cutoff, may have strengthened the incentive to reallocate for students at risk of missing the cutoff.

In the early 2000s, public schools in Florida were graded A-F based on student test performance in math, reading, and writing. If a school was graded F twice in four years, the students were offered vouchers to transfer to another public or private school. Winters, Trivitt, and Greene (2010) use a regression discontinuity design to compare F and D graded schools (see also Rouse et al.

2007, 2013). Schools graded F, but which just missed D, subsequently improved in math, 0.20σ, and reading, 0.09σ. But the same schools also improved in science, 0.10σ.

In an important contribution to the school accountability evidence, Dee and Jacob (2011) estimate effects of the incentives created by No Child Left Behind (NCLB). Using a difference-in-differences style design, the authors compare states which already had consequential school accountability programs before NCLB to states where the NCLB requirements created new accountability programs. Notably, Dee and Jacob measure student achievement using National Assessment of Educational Progress (NAEP) scores, not the high-stakes state tests which determined consequences for schools. Dee and Jacob find evidence of gains in math as a result of NCLB incentives, but no improvement in reading or science. Of the three subjects, only math and reading directly determined NCLB performance measures.

Altogether there is little evidence, currently, that performance incentives consequentially distort teachers' decisions in allocating effort across subject areas. At least no evidence that students are made worse off in non-incentivized subjects. Still, the examples so far focus on closely related outcomes: academic achievement measured by standardized tests, even if in different subjects.[11]

2.1.2. Unmeasured and Long-Run Student Outcomes

In response to incentives for academic test scores, teachers may reallocate work effort in ways that negatively affect students' social and emotional development or longer-term success. Academic skills—in subjects like math, reading, and science—are not the only outcomes teachers contribute to in

[11] Holmstrom and Milgrom (1991) suggest specialization—for example, teachers specializing in one subject area—as way to reduce distortion by matching evaluation and job design features. Fryer (2018) conducts a field-experiment in which elementary school teachers specialized in math or reading; the results show important potential costs of specialization to weigh against any benefits from improved incentive design.

educating their students. And recent work demonstrates teachers' contributions to students' social and emotional skills (see Jackson 2018, Kraft 2019), and contributions to their students' success in adulthood, including labor market earnings (Chetty, Friedman, and Rockoff 2014b). A clear opportunity for future contributions in this literature would be to consider a broader range of teacher job responsibilities and student outcomes.

One such contribution comes from Andrabi and Brown (2022). The setting is a field experiment in Pakistan where, in treatment schools, teachers competed for bonuses of between 0-10% of annual salary based on their performance rank. The design also involved experimental variation in the type of performance measure used to rank teachers. In half of treatment schools, performance rank was based on student test score improvement, and, in the other half, based on the school principal's subjective rating. In the subjective condition, principals had discretion over what evaluation criteria they used but were required to inform teachers of the (potential) criteria at baseline. Principals often included both teacher contributions to student test scores as one criterion, alongside other measures of pedagogy.[12]

Both objective and subjective incentives produced nearly identical gains in student achievement tests: roughly 0.10σ in math, science, Urdu, and English. Yet, the similarity of those reduced-form achievement effects masks notable differences in how teachers responded to objective and subjective performance measures. Briefly, data from classroom observations show that teachers in the objective condition spent more class time on practice tests and test-taking skills and were more likely to discipline students. Results consistent with concerns about "teaching to the test" strategies discussed later in section 2.2. By contrast, teachers in the subjective condition improved their basic teaching skills, as

[12] A companion paper, Brown and Andrabi (2021), discusses additional features of the experiment, including testing for the influence of self-selection in incentive effects. See section 3.

measured by the CLASS rubric. In short, both conditions improved test scores but likely through different changes in teacher effort. Those different teacher responses may also have different consequences for student social-emotional outcomes.

Further results from Andrabi and Brown (2022) do suggest teacher incentives can affect students social-emotional outcomes. Neither incentive type—student test scores (objective) or principal ratings (subjective)—had a statistically significant effect on students' socio-emotional outcomes compared to the control condition. However, comparing between objective and subjective conditions, students' social-emotional outcomes were somewhat better in the subjective condition. For example, when teacher bonuses were based on test scores alone, students were less likely to say they enjoyed learning, and more likely to want to switch schools. There was no difference in measures of students' ethical behavior, citizenship, or inquisitiveness.[13]

Longer-term student outcomes might also reveal distortions in teacher effort. Assume that, in the absence of job performance measures or incentives, the teacher allocates time and effort to maximize her students' future success in school and adulthood. A new performance incentive, linked only to student test scores, might well cause her to increase class time and effort for teaching tested knowledge and skills, at the expense of other skills relevant to future success but unmeasured(able). We would predict negative effects of test-based incentives on students' future outcomes.

Two empirical examples find the opposite result: positive effects of teacher incentives on students' future success. Lavy (2009, 2020) studies

[13] Andrabi and Brown (2022) make a notable contribution to evidence on subjective performance measures in teaching. However, the achievement gains caused by the subjective incentives may or may not generalize to other settings. Using quasi-experimental designs, Atkinson and coauthors (2009) and Martins (2009) examine the effects of subjective evaluations in England and Portugal, respectively, used to determine salary increases. The estimates in Atkinson et al. (2009) are mixed between null and positive effects. Martins (2009) finds negative effects on student achievement.

performance incentives for high school teachers in Israel. Teachers were ranked based on their students' test outcomes, both pass rates and average score, and top ranked teachers received large cash bonuses. Teachers competed against others who were teaching the same subject in the same school. The introduction of the bonus tournament improved incentivized outcomes: student average test scores and pass rates increased (Lavy 2009). Importantly, in a second paper, Lavy (2020) then follows students between high school graduation and age 30, well beyond the time when student outcomes determine teacher bonuses. Benefits continued for the students of incentive-eligible teachers. They were more likely to enroll in college and complete more years of college. They were also more likely to be employed and were earning 8-9% more around age 30. These longer-term benefits are consistent with real improvements in students' achievement during high school, and encouraging evidence that student welfare improved when their high-school teachers competed for performance incentives.

Deming and coauthors (2016) report on the longer-term effects of Texas' 1990s-era school accountability program. Texas rated high schools with four categories—low performing, acceptable, recognized, and exemplary—based on TAAS scores for 10th grade students, attendance, and dropout rates. Deming and coauthors' identification strategy compares different cohorts of 10th grade students: cohorts attending the same school but facing different pressure to score well. The between-cohort within-school variation in incentive strength comes from policy rules which increased the minimum threshold for an "acceptable" rating each year. In 1995 "acceptable" required that 25% of students to pass, but that percent rose by 5 points each year. The requirements for "recognized" also increased over time. Schools rated "low performing" (below "acceptable") faced sanctions that could include layoffs or closing the school.

The threat of sanctions did improve 10th grade test scores, but also improved longer-term student outcomes well beyond high school. In years when a

Texas high school was at risk of being rated "low performing," scores on the 10th grade exams improved. That improvement on the incentivized measure is perhaps not surprising. However, those same cohorts were also more likely to graduate on time and complete more math courses during high school. They were more likely to attend and graduate from college, and at age 25 they had higher earnings in the labor market. Moreover, these gains mainly accrue to previously lower-achieving students.

While only two studies, the results from Lavy (2020) and Deming and coauthors (2016) are inconsistent with concerns about costly distortions. Indeed, the individual bonuses and school accountability consequences led to better futures for students, suggesting the incentive programs improved the work of teachers and schools.

2.1.3. Students

Evaluation can affect how teachers allocate effort among their many students. Some performance measures do not give equal weight to all students' test scores, creating an incentive for teachers to treat students differently. A well-known example is the differential weights that occur when the performance measure is the percent of students who pass the exam (the "percent proficient" approach). Students who are on the margin of passing or failing—colloquially "bubble kids"—have an outsized influence a teacher's (school's) score, and likely a larger return on teacher effort. By contrast, there is likely little return on teacher effort given to students who are far above or far below the passing threshold.

Using data from the Chicago Public Schools, Neal and Schanzenbach (2010) demonstrate that teachers do respond to the bubble-kids incentives. After the start of Chicago's new school accountability program in 1996, student test scores improved, as Jacob (2005) also showed. Neal and Schanzenbach further show that students in the middle of the achievement distribution improved more than their lower- or higher-achieving peers. The same pattern occurred again in

Chicago at the start of NCLB in 2002. Other researchers have documented this pattern elsewhere in the United States (Reback 2008, Springer 2008, Macartney, McMillan, and Petronijevic 2018, 2021) and in England (Burgess et al. 2005). Still, it may be that this shift of teacher attention toward the bubble kids was intended by the policy designers, and the welfare implications are unclear.

Aucejo, Romano, and Taylor (2022) study an evaluation rule that created much sharper differences in incentives across students, and clearer evidence of allocative inefficiency. Under a short-lived policy in North Carolina, students who failed the end-of-year state tests were re-tested 2-3 weeks later, but then only the higher of their two scores was used to calculate the teacher's final evaluation score. Thus, effort given to failing students over the 2-3 weeks could only increase the teacher's final score, but effort given to students who had already passed would have no effect on the teacher's score. Aucejo and coauthors use a regression discontinuity design to estimate effects on student math achievement one year later. Students who barely failed the initial test in year t—and thus likely received more teacher effort for 2-3 weeks—scored 0.03σ higher in at the end of year $t+1$ than did their peers who had barely passed in year t. At the end of year $t+2$ the difference was still 0.02σ.

In this North Carolina case, teacher incentives changed discontinuously at the cutoff between failing and passing the initial test, encouraging teachers to give much more effort to the failing students for the 2-3 weeks before the retest. Yet, while teachers' incentives differed sharply between students, the barely-failing and barely-passing students were otherwise identical, and would have benefited equally from the extra teacher effort. The extra benefits to barely-failing students, 0.02-0.03σ higher math achievement in the longer-run, are evidence that otherwise-identical students were treated differently by their teachers because of the different evaluation incentives attached to different students.

Another example of teachers reallocating effort across students in response to incentives comes from a field experiment in China's Shaanxi and Gansu provinces. Loyalka and coauthors (2019b) randomly vary the performance measures used to determine teacher bonuses. In all three treatment conditions, teachers were ranked based on their students' math test scores. Top ranked teachers received a cash bonus equal to one month's salary, then bonus amounts scaled down linearly with rank, and bottom ranked teachers received 10% of the top bonus amount. What differed across the three conditions was how teacher performance was measured. First, in a "levels" condition, each teacher was ranked using the average test score, year t, among students in her class, without any adjustment for prior achievement. Second, in a "gains" condition, teacher rankings used the class-average change in test scores, from year $t-1$ to t. The third condition was a "pay-for-percentile" condition, with a measure following Barlevy and Neal (2012). Each student was ranked based on his year t score, but only among a comparison group of students defined by year $t-1$ score, thus yielding a conditional percentile for each student. The teacher's percentile rank was the average of her students' conditional percentiles.

Teachers in the Shaanxi-Gansu experiment responded differently to the three performance measures, in ways which suggest they understood that their incentives differed between the measures, and further that they should treat students differently. The effects of teacher incentives were largest in the pay-for-percentile condition, where students scored 0.15σ higher than the no-incentive control. There was little if any effect in the gains condition, 0.01σ, and the levels condition was intermediate, 0.08σ. The contrast between "pay-for-percentile" and "gains" is notable because both adjust for students' prior achievement, which should reduce (teachers') concerns about differential assignment of students to classes. However, the pay-for-percentile measure distributes incentives more uniformly across students and classes, which might contribute to the difference in

treatment effects. The gains measure incentivizes teachers to focus more effort on students who can make larger score gains, between $t-1$ and t, and the size of potential gains likely varies across the distribution of baseline, $t-1$, scores. By contrast, a student's potential conditional percentile does not vary with baseline score.

Teachers' expectations about students play an important role, as Loyalka and coauthors (2019b) demonstrate empirically. At baseline the researchers measured teachers' expectations about the effect of their own effort on their students' achievement growth. Each teacher was asked for two predictions about each student in her class: how much the student's test score would improve over the school year with and without one extra hour per week of individualized instruction. In both levels and gains conditions, treatment effects were larger for students whom the teacher expected would benefit more from her effort. Heterogeneity consistent with the incentives of those two conditions. However, in the pay-for-percentile condition effects were more uniformly distributed across students. Moreover, while treatment effects varied with teachers' baseline expectations, effects varied much less with baseline test scores.

A final example of distortion comes from Macartney (2016)—distortion within students over time. Many performance measures are based on growth in student test scores over the course of one school year. This creates a ratchet effect feature: larger performance improvements today, in year t, will make future rewards, in year $t+1$, more difficult to achieve. This creates an incentive for teachers to reallocate effort intertemporally. To test for response to these incentives, Macartney compares North Carolina schools which have different grade level configurations—K-5, K-6, and K-8—before and after the introduction of test-based accountability. Different grade configurations create different incentives. For schools serving grades K-5, the ratchet effect feature does not apply to 5th grade test scores; any gains in 5th grade affect some other school's

performance measure. And indeed, under test-based accountability for schools, 5th graders score 0.04-0.06σ higher in K-5 schools than do 5th graders in K-6 or K-8 schools.

2.2. Teaching to the Test

Distortions in teacher effort can also occur within subject. Consider a teacher who is only responsible for teaching one subject, say, math. In practice, any one math test can only cover a relatively small sample of math topics or skills. Attaching rewards or sanctions to a specific test, then creates an incentive to focus effort on the sample of topics or skills that test covers. If the teacher can anticipate which topics are more likely to be on the test, then she might give more class time to those topics. This is sometimes described as "narrowing instruction" or "teaching to the test."

The welfare consequences of narrowing instruction are not obvious. Students' mastery of specific skills—actual mastery not just test scores—may well improve, though likely at the cost of less mastery of other skills not emphasized by the incentivized test. That potential tradeoff is the source of ongoing policy debates. Lazear (2006) provides a framework for the problem, and describes conditions under which narrowing instruction would benefit student welfare. Koretz (2002, 2008) discusses this topic from the perspective of psychometricians.

The colloquial "teaching to the test" or "test prep" is also sometimes used to describe other actions taken by teachers: explicitly teaching students the skills of taking a test, giving class time to intensive review of material just prior to the test, or having students take practice tests. These actions also often exploit specific knowledge of a particular test. For example, even if two different math tests cover the same topics and skills, test makers may have idiosyncratic differences in how they design test questions. Additionally, test makers often include some of the same exact items from one year to the next to allow for

equating scores and scales over time. A teacher might exploit these features by having students take practice tests by the same test maker or coaching students on specific items.

2.2.1. Same Subject, Different Tests

One way to approach this problem empirically is to use two tests of the same subject. Imagine a setting with two separate student exams, both designed to measure the same underlying skills, but only one of the two determines teacher performance incentives. If student scores improve on the incentivized exam, but not on the second low-stakes exam, we should be skeptical about whether the change in teacher effort produced real improvements in students' mastery of the subject. The literature includes several examples of this kind of generalizability test. The results across studies are mixed, and sometimes mixed within one incentive program.

One example comes from Mbiti and coauthors' (2019) experiment in Tanzania, described earlier. Recall that teachers received bonuses based on students passing a year-end government exam. Pass rates increased by 37% in math, 17% in Kiswahili, and 70% in English; across the three subjects, scores increased by 0.21σ (standard error 0.07). Yet, scores increased only 0.03σ (standard error 0.04) on separate low-stakes exams taken about three weeks before the incentivized exams. An example where apparent gains did not generalize.

In contrast, Sojourner, Mykerezi, and West (2014) document a case where incentives did improve scores on both low- and high-stakes student tests. Sojourner and coauthors study Minnesota's Q-Comp program, which included, among other features, small cash bonuses for teachers based on student test scores. In some school districts, students took two different exams: the state-required Minnesota Comprehensive Assessment (MCA) plus a test provided by the Northwest Evaluation Association (NWEA). In those districts, bonuses were

determined by only one of the two exams, but some districts chose to incentivize MCA while others chose NWEA. These details allow Sojourner and coauthors to compare incentivized versus non-incentivized outcomes, while also controlling for any general test differences between MCA and NWEA. In contrast to the Tanzania case, Q-Comp had the same effect on both incentivized and low-stakes exams. No matter which test the district incentivized, the estimated Q-Comp effect is 0.03σ on three out of four exams (MCA-reading, NWEA-reading, and NWEA-math) and null on the fourth (MCA-math).

Some of the variation in these results may come from the nature of the stakes in the nominally low-stakes exams. Perhaps Minnesota teachers had unobserved incentives, beyond the Q-Comp bonuses, and those incentives applied to both MCA and NWEA exams. Both exams were used by the school districts in their course of business. In Tanzania the low-stakes exam was administered by the researchers for the experiment's purposes only. One low-stakes exam used in this literature is the National Assessment of Educational Progress (NAEP), administered by the U.S. Department of Education. The sampling design of the NAEP makes it unlikely that any individual teacher or school would have an incentive to teach to the NAEP.

Vigdor (2009) studies performance incentives in North Carolina's ABCs program, comparing trends in the incentivized state exams to trends in NAEP scores. In the ABCs program, each teacher received a bonus, up to $1,500, if her school met a predetermined target for student test score growth on the state exams. Vigdor finds mixed results. For math, both state and NAEP scores showed a similar pattern of improvement. But apparent gains on the state reading exams were not reflected in the NAEP scores.

Again, this literature extends to the threat of sanctions in school accountability programs. Recall the examples of test-based accountability for schools in places like Chicago and Texas. In the Chicago Public Schools setting,

Jacob (2005) has data from the Iowa Test of Basic Skills (ITBS), used to determine school sanctions, and also data from the lower-stakes Illinois Goals Assessment Program (IGAP). The results are mixed. Lower-stakes IGAP scores did not show the same improvement as ITBS for younger students, but both ITBS and IGAP scores improved for older students. In Texas, famously, student scores on the Texas Assessment of Academic Skills (TASS) improved dramatically after the state's new school accountability program began. However, Klein and coauthors (2000) show that the TASS improvements were not consistently matched by NAEP trends. For example, Texas' 4th grade NAEP scores improved in math but not reading. Studying Kentucky during the same period, Koretz and Barron (1998) find no evidence that NAEP scores improved, despite large gains on the Kentucky state tests. The analyses of Chicago, Texas, and Kentucky are important evidence; still, these school systems were not the only places to adopt school accountability programs in the 1990s.

Hanushek and Raymond (2005) and Dee and Jacob (2011) contribute broader assessments of school accountability incentives. Both compare student achievement outcomes across U.S. states using low-stakes NAEP test scores. In the 1990s, before NCLB, Hanushek and Raymond (2005) find NAEP scores improved when a state's program included meaningful sanctions for schools, but not when the program simply reported test scores publicly. Dee and Jacob (2011) further show that, under NCLB, NAEP math scores improved in states that had not already adopted meaningful sanctions. Fourth-grade reading scores, however, did not improve. These between-state comparisons suggest optimism about the benefits of school accountability incentives. The tradeoff, compared to the studies of Chicago or Texas, is that the between-state designs in Hanushek and Raymond (2005) and Dee and Jacob (2011) cannot tell us about effects on the high-stakes exams which differed from state to state. We would like to know whether high-stakes state-test results are reasonable predictors of low-stakes test results.

In summary, performance incentives for teachers regularly generate increases in student scores on the tests which determine rewards or sanctions. Teacher effort does respond to incentives. However, even for a given subject, scores on low-stakes non-incentivized tests do not always show the same improvements. Thus, improvement in incentivized outcomes is insufficient evidence to claim that student mastery of the subject improved.

2.2.2. Same Test, Different Questions

An alternative empirical approach is to compare different questions (synonymously, items) from the same test. One kind of "teaching to the test" requires that the teacher can predict which topics or question types are more likely to be on one test. Cohodes (2016) and Jacob (2005) use variation between test items in the predictability of the item appearing on the test.

Cohodes (2016) uses item-level data from the Massachusetts Comprehensive Assessment (MCAS) exams, and the contrast between charter schools and traditional public schools. Both charter and traditional schools were subject to potential NCLB sanctions, but charter reauthorization rules created additional incentives to raise MCAS scores. Cohodes first details patterns of MCAS exam topics over time. In middle school math, for example, number sense and operations are tested much more frequently than data analysis, statistics, and probability. If charter schools responded to incentives by teaching to the test, then we would expect larger gains on expected test topics, like number sense and operations, and smaller or no gains on less-often tested topics, like data analysis. Cohodes finds no evidence to support that prediction; charter school students score consistently higher on all topics.

Jacob (2005) finds somewhat different results in Chicago. Test-based incentives for teachers caused students to score better both on basic math skills—computation and number concepts—and on more complex skills—estimation and word problems. Still, gains on the basic skills were twice as large. Questions on

basic math skills were more common on the ITBS and thus more predictable, but, as Jacob says, those basic skills are also potentially easier to teach. Even if different topics were equally likely to appear on the test, teachers may still have an incentive to reallocate effort, when the expected return on effort differs across topics. Moreover, Dee and Jacob (2011) find changes in NAEP sub-scores that are similar to Jacob's (2005) results. Of course, teachers were not "teaching to the NAEP," but the NAEP sub-score differences may be spillovers from distortions caused by the high-stakes exams.

2.2.3. Test-Taking Skills and Test Preparation

A different kind of "teaching to the test" involves teaching students the skills of test taking. Class time on test-taking skills is also likely an unintended distortion in effort. One example of test-taking skills is this: On a multiple-choice question, if you are going to guess, first eliminate answers you know are incorrect before guessing. That example is quite general, but test-taking skills can get very specific. An example from Koretz (2017): Math scores will improve, in expectation, if students are taught to choose the answer involving 3-4-5 or 5-12-13 every time they see a right triangle. Test scores improve even if the students know nothing about the Pythagorean theorem. Thus, test-taking skills may create differences in test scores which do not reflect differences in students' actual mastery of the subject.

Empirical evidence on teacher incentives and test-taking skills is quite scarce. One exception comes from Glewwe, Ilias, and Kremer's (2010) analysis of a field experiment in Kenya. Treatment teachers were evaluated as teams, all grade 4-8 teachers in the school, and rewarded with in-kind prizes. At the end of the experiment, student scores had improved on the incentivized government exam but not on a separate low-stakes exam (0.14σ versus -0.02σ and far from statistically significant). One key difference between the exams was item format. The low-stakes exam included many fill-in-the-blank questions and few multiple-

choice, while the incentivized exam was entirely multiple-choice. When Glewwe and coauthors looked just at the low-stakes exams, treatment effects were larger for multiple-choice items than for fill-in-the-blank items. Additionally, in the first year of the experiment, the two exams had a similar multiple-choice format, and there was no difference in treatment effects across the exams (0.048σ and 0.046σ, though both estimates are quite noisy). This pattern of results would be consistent with teachers coaching students on multiple choice test-taking skills.

Teachers may also reallocate class time to other forms of preparation for a specific test. Test-prep activities—like intensive review of material or taking practice tests—are sometimes counted among "teaching to the test." Teachers eligible for bonuses in Andhra Pradesh self-reported giving practice tests to students twice as much as control teachers, 30% vs 14%. Though treatment teachers also held extra classes beyond regular school hours, so the test-prep may not have crowded out other uses of class time (Muralidharan and Sundararaman 2011). Teacher survey data also show more test prep in the Kenya, Tennessee, and Mexico field-experiments (Glewwe, Ilias, and Kremer 2010, Springer et al. 2010, Behrman et al. 2015). As part of their field experiment in Pakistan, Andrabi and Brown (2022) measure test-prep activities in classroom observations, and find treatment teachers devote more class time to practice tests and test-taking skills.

What is less clear, in the empirical literature, is how test-prep affects test score measures. One prediction is that any gains from test-prep—intensive practice or review just preceding a test—should decay more quickly than achievement gains from other forms of learning and teaching. Among the teacher incentive (quasi-)experiments, there are some empirical examples which could be seen as consistent with that prediction. In the Kenya experiment (Glewwe, Ilias, and Kremer 2010), students scored 0.14σ higher (standard error 0.071) in year two, the last year their teachers were eligible for bonuses. But the same students only scored 0.08σ higher (standard error 0.071 as well) when tested again in year

three. That suggests decay of about half, if we assume treatment teachers reverted to behavior matching the control teachers in year three. A second example comes from the school accountability F grades in Florida. Both Rouse and coauthors (2007, 2013) and Chiang (2009) find statistically significant effects two years after treatment, and implied persistence rates greater than half in some cases.

Fade-out by 50% in one year may seem like rapid decay, but Jacob, Lefgren, and Sims (2010) estimate that teachers' normal contributions to student test scores, without performance incentives, also fade-out by about half after one year (see also Kane and Staiger 2008, Rothstein 2010, Chetty, Friedman and Rockoff 2014b). Indeed, the test-score gains from other educational interventions also decay at similar rates: reduced class size (Krueger and Whitmore 2001), increased class time (Taylor 2014), private schools (Andrabi et al. 2011), and others. In short, fade-out is not specific to incentive-induced gains, and may well be caused by factors other than test-prep. However, in one direct comparison, Macartney, McMillan, and Petronijevic (2018) find evidence that test-score gains caused by teacher performance incentives fade-out more quickly than gains arising from teachers' non-incentivized contributions to test scores.

In two other examples, the timing of key events provides somewhat sharper evidence on test-prep. The first is the North Carolina retest incentives studied by Aucejo, Romano, and Taylor (2022). Recall that teachers had just 2-3 weeks between the initial test and the retest. During that short period, teachers were strongly incentivized to focus their effort on students who had initially failed and who would be retested. Aucejo and coauthors' regression discontinuity estimates compare students who were treated identically before the initial test, but (potentially) treated differently for just a few weeks after. The initially-failing students scored higher one year after the 2-3 intensive weeks, 0.03σ, and two-years after, 0.02σ. Those results suggest meaningful persistence over time of the effects from just 2-3 weeks of extra teaching attention.

In the Tanzania experiment, Mbiti and coauthors (2019), students took the high-stakes incentivized exam after they had taken the low-stakes exam. Recall that there was a large treatment effect on the incentivized exam, 0.21σ, but no average effect on the low-stakes exam. Assume teachers and students were engaged in test-prep in the days or weeks leading up to the high-stakes exam, and that the benefits of test-prep would spill over to the low-stakes exam. Then the longer the period between the low- and high-stakes exams, the smaller should be any spillover effects. However, Mbiti and coauthors (2019) find that treatment effects on the low-stakes exam do not vary with the number of days between the two tests.

2.2.4. Pay for Percentile

Most of these "teaching to the test" strategies require that tests have predictable features, and that teachers can exploit those features. Barlevy and Neal (2012) propose a system of teacher performance measures and incentives—called "pay for percentile"—which removes the predictable features of tests. First the performance measure: At the beginning of the school year (baseline) students are grouped into comparison sets, based on their prior achievement scores and perhaps other considerations. Then each student's end-of-year test score is ranked within his comparison set, yielding a conditional percentile. The teacher's performance measure is a function of those conditional percentile ranks. This kind of conditional percentile approach has been proposed by others (for example, Betebenner 2009).

Barlevy and Neal (2012) add a key feature in in the pay-for-percentile design: remove the predictable features from the tests. Evaluating teachers with a conditional-percentile measure does not require equating test forms or scales, and thus eliminates the need for repeating specific test items or question formats from one year to the next. Barlevy and Neal describe other beneficial properties of this incentive strategy. One cost is that, in practice, it would likely require two

separate testing regimes: one high-stakes for teachers, and a second low-stakes for teachers used to assess student achievement growth. Though the two separate regimes would also improve inferences about student achievement from the second.

A number of recent field experiments have used a pay-for-percentile style performance measure (Fyer et al. 2012, Loalka et al. 2019b, Leaver et al. 2021, Brown and Andrabi 2021). As described earlier, Loalka and coauthors (2019b) experimentally compare a pay-for-percentile style measure with other common test-based measures and find quite different results. However, these examples do not fully test the potential benefits of Barlevy and Neal's (2012) proposal. First, some studies use conditional-percentile measures, but do not eliminate the predictable features of the tests. Second, teacher knowledge of a specific test's features develops over repeated exposure to the same test, but many experiments use a novel test or last just one or two years.

2.3. Other Responses to Incentives

Teachers' responses to performance incentives extend beyond changes in how they go about teaching their students day to day. The simplest example is cheating on student tests. Cheating increases the teacher's performance measure, capturing the linked reward, without any change in student mastery of the material. Reports of teacher cheating appear in the popular press, including a famous case in Atlanta (New York Times 2013, Washington Post 2022). But cheating and other subtler forms of manipulation have also been studied by economists.

Using data from the Chicago Public Schools, Jacob and Levitt (2003) estimate that serious cheating by teachers (or administrators) occurs in perhaps one in twenty elementary school classrooms each year. Moreover, cheating increased when Chicago introduced new incentives—the threat of sanctions for low scores—in the 1990s. This may partly explain why Jacob (2005) finds that

improvements on the tests Chicago incentivized (ITBS) are only partly reflected on a separate low-stakes test (IGAP). Aware of the threat of cheating, Mbiti and coauthors (2019) take steps to prevent cheating using proctors and many different test versions within any class. Behrman and coauthors (2015) detect cheating, apparently by the students themselves, and address it in their analysis of an incentive experiment in Mexico discussed in section 2.4.

Teachers may also manipulate performance measures in more subtle ways. For example, teachers could increase average test scores by excluding lower-achieving students from the test takers. Researchers have documented a variety of empirical examples: suspending students from school during the test period (Figlio 2006), encouraging students to be absent (Cullen and Reback 2006), designating students as special education (Jacob 2005, Cullen and Reback 2006, Figlio and Getzler 2006, Deming et al. 2016), among other possibilities. To address this threat, often evaluation programs now include test participation requirements.

Figlio and Winicki (2005) document a non-instructional response involving school lunch menus. The study makes use of detailed data on the nutritional content of school meals, for a random sample of Virginia schools, collected by the U.S. Department of Agriculture. Accountability rules in Virginia, as in other states, imposed sanctions on schools with sufficiently-low test score performance. Figlio and Winicki demonstrate that, in response to the threat of sanctions, schools increased the caloric content of school lunches on testing days. There is some evidence that the strategy succeeded in raising student test scores. The mechanism here is about student effort and cognition on the day of the test specifically, not learning or nutrition in the preceding weeks and months.

Student effort during a test, induced by teachers through extra calories or other means, is one potential explanation for the effects of teacher incentives on student achievement scores. Students with identical mastery of the material on a

test, can score differently simply because of effort on the test day. Gneezy and coauthors (2019) demonstrate the influence of student test effort in an experiment. At the beginning of the test, with no advance notice, students were offered cash rewards linked to their test score. Thus, the incentives could not affect student preparation, only effort during the test. In the United States sample, students attempted more questions on the test, and answered more questions correctly. These results show scope for teacher strategies involving student effort, though there were no teacher incentives in the experiment. However, in Chicago, Jacob (2005) finds similar evidence of changes in student test effort when the ITBS exam became high-stakes for teachers and schools. Students attempted more questions and answered more correctly. The effects of teacher incentives were particularly notable at the end of the reading exam. After controlling for question difficulty, student scores on the last 20% of questions, when effort might typically wane, improved nearly twice as much as on the first 20% of questions. Still, inducing more student effort during the test is not the same as cheating. Cheating clearly degrades inferences about student learning and future success. The consequences of test effort for similar inferences are less obvious. With greater student effort, test scores may better reflect student knowledge at the time, but the resulting scores may or may not be better predictors of success in the future (Balart, Oosterveen, and Webbink 2017, Segal 2012).

2.4. Team Incentives

Team incentives are ubiquitous in schools. Many of the empirical examples in this literature involve team incentives. That includes studies of "school accountability" incentives, where all teachers working in a school are evaluated as a team and low performing schools face sanctions (e.g., Jacob 2005, Rouse et al. 2007, 2013, Dee and Jacob 2011, Deming et al. 2016, Macartney 2016). Others involve bonuses for individual teachers determined by team performance measures (e.g., Ladd 1999, Lavy 2002, Vigdor 2009, Glewwe, Ilias,

and Kremer 2010, and others discussed in the next few paragraphs). Yet, direct comparisons of team and individual incentives are scarce, as is evidence on the mechanisms that might differ.

One experimental comparison of individual versus team incentives comes from the Andhra Pradesh study, described by Muralidharan and Sundararaman (2011). Recall that teachers earned bonuses scaled to the average change in student test scores. In half of the treatment schools, each teacher's bonus was based only the scores of students she taught. In the other half, all teachers in the school received the same bonus, based on the average score across all students in the school. The bonus formula was otherwise the same. Both incentive systems produced increases in student achievement scores. After one year of teacher incentives, student scores improved meaningfully, though the improvements were similar in individual-incentive and team-incentive conditions (0.16σ and 0.14σ, respectively, compared to the control condition). But after two years, achievement gains were nearly twice as large when teachers earned bonuses individually, 0.28σ compared to 0.15σ. In a follow-up, Muralidharan (2012) finds that individual incentives continue to produce larger gains after three, four, and five years.

A second experimental comparison comes from an experiment conducted by Fryer and coauthors (2022) in Chicago Heights, Illinois. In one treatment condition, individual teachers earned cash bonuses based on the test scores of their own students, using a pay-for-percentile measure and an otherwise low-stakes test. In another, teachers were paired in teams of two, and earned bonuses based on all students taught by either teacher. Each pair worked at the same school, and taught the same grade, subject, and similar students. Bonuses were scaled from a maximum of $8,000 down to zero linearly based on the average conditional percentile rank of the teacher's (or team's) students. Student scores on the incentivized ThinkLink test improved substantially; however, Fryer and

coauthors find no meaningful differences in those effects between individual and team incentives.[14]

Why might incentive effects be smaller with team incentives? Team evaluation creates the opportunity for free riding, weakening the incentives for any one individual teacher to give more effort (Holmstrom 1982). If free riding weakens incentives, then the effect of team bonuses should depend on the number of team members. Muralidharan and Sundararaman (2011) and Muralidharan (2012) test for heterogeneity of effects by team size, and find no differences, but the Andhra Pradesh schools were small with 2-5 teachers each. In Fryer and coauthors' (2022) experiment each team was just two teachers, by design.

Goodman and Turner (2013) do find some differences in effects by team size, studying randomized trial in New York City. Fryer (2013) describes the experiment and several results. Treatment schools could earn a bonus of $3,000 per teacher (5% of average salary) if the school met a performance target set by the school district, and $1,500 for meeting 75% of the target.[15] Targets were based on variety of measures: student test scores, student attendance, graduation rates for high schools, and others. Fryer (2013) finds no evidence that the incentives improved student achievement, attendance, or graduation, on average, and no change in measures of teacher behavior. The average school in the

[14] The design also included experimental variation in the framing of the bonuses. In a "loss aversion" condition, teachers were given $4,000 (the expected value of the bonus) before the school year began, and were required to pay back any difference from the earned bonus after the test results were known. In a conventional condition, teachers received bonuses after the test results were known. The loss-aversion incentives increased ThinkLink test scores dramatically, 0.20-0.40σ, while the conventional incentives had smaller and usually statistically insignificant effects. Though the pattern of loss-aversion versus conventional differences was the same for individual and team conditions. There is some suggestive evidence that, in the loss-aversion framing, team incentives had a larger effect on state ISAT tests, which did not affect bonuses, but were used for school accountability.

[15] The bonus was awarded to the school. A school committee then determined how the bonus pool was distributed to individuals, with the option of differential bonuses. In most schools, however, bonuses were based solely on position held (Fryer 2013). Marsh et al. (2011) also study the same experiment.

experiment employed nearly 60 teachers (standard deviation 22). However, the smallest (bottom quartile) elementary schools had fewer than ten teachers, and the smallest middle schools had fewer than five math teachers and six reading teachers. Goodman and Turner (2013) find that math scores improved roughly 0.08σ in that smallest one-quarter of schools, and scores declined in the larger three-quarters of schools. The pattern was similar for reading scores, but about half as large and not statistically significant.

Imberman and Lovenheim (2015) provide further evidence that team design influences the effects of team incentives. Under Houston's ASPIRE program, teams of high school teachers competed in rank tournaments for cash bonuses, based on the team's value-added score. Each teacher could earn a yearly bonus of up to $5,000-7,000. Teams were defined by the intersection of subject area, grade level, and school. For example, 10th grade science teachers at Westside High School competed as a team against 10th grade science teachers at other high schools. Imberman and Lovenheim point out that, in settings like Houston, incentive strength can vary between teachers within a team. The team's rank depends on the scores of many students, but some teachers teach a greater proportion of that group of students and other teachers a smaller proportion. And, empirically, student scores improved more in Houston when their teacher taught a larger share of the team's students. The effect of incentives increased by 0.05-0.09σ with a 10-point increase in share, among teachers teaching a relatively small share of students. Once teachers were responsible for more than one-quarter or one-third of students, further increases in share made little difference to incentive effects. As comparisons, the average Andhra Pradesh school had three teachers, implying shares of roughly one-third each in the team incentive condition. Shares in New York were quite small on average, but reached perhaps one-tenth to one-fifth in the schools where Goodman and Turner (2013) find positive effects. Imberman and Lovenheim also find that, conditional on the share

measure, the number of teachers on a team does not predict differences in treatment effects.

Free riding and monitoring are potential costs of team evaluation, likely reducing incentive effects. The potential benefits are gains through complementarities in production—gains which can be captured with greater coordination among team members. The net effect of team incentives, in any specific case, will thus depend on the scope for free riding, as Goodman and Turner (2013) and Imberman and Lovenheim (2015) show, but also depend on the size of any as-yet-unrealized gains from coordination.

Behrman and coauthors (2015) document a case where team incentives generate gains in student achievement, but individual teacher incentives do not. High schools in Mexico were randomly assigned to one of three incentive designs or a control. In one treatment condition, only teachers were incentivized. Each math teacher could earn a cash bonus based on the test scores of the students in her class. Those individual teacher incentives generated little to no effect on student test scores (point estimates were 0.01-0.04σ and not statistically significant). In sharp contrast, test scores rose dramatically in a team incentive condition: a treatment effect of 0.30σ after the first year, accumulating to 0.60σ after three years. In this team treatment, math teachers received the same bonus as in the individual-incentive condition, plus an additional bonus based on the math scores of other math teachers' classes in the school. But the "team" extended beyond math teachers. Non-math teachers and school principals also received smaller bonuses based on math test scores, and students received bonuses based on their own math test scores and the scores of their math classmates.

The gains in Mexico standout in this literature on incentives. Disentangling the mechanisms is complicated by the many different team members. The gains of 0.30-0.60σ could partly reflect complementarities between math teachers, who had more to gain from coordination than non-math teachers

and administrators. The gains could also partly reflect complementarities between math teachers and their students, or peer-effect complementarities among students. Nevertheless, the size of the gains suggests sorting out those mechanisms is worthy of additional experimentation.[16] The idea of student-teacher teams, and complementarities between teacher effort and student effort, are largely missing from the current literature. Jackson (2010, 2014) studies one other example where teachers and students were simultaneously incentivized. However, some of the 0.30-0.60σ gains in Mexico seem to come from individual student incentives. In a third experimental condition, only students were incentivized, and received bonuses based only on their own math scores. Individual student incentives increased math scores substantially, 0.20-0.30σ, but not enough to fully explain the team incentive effects.

2.5. Noisy Performance Measures

Noise in performance measures weakens incentives and should weaken any effort response. The risk-incentive tradeoff is a basic prediction of the agency theory models used as a rationale for evaluation, and teacher performance measures can be quite noisy in practice. Yet, there is little evidence of the empirical consequences of noise in teacher performance measures. One exception is Brehm, Imberman, and Lovenheim (2017).

Recall that, in practice, schools cannot observe true performance, $h(e)$, but only some measure of performance, h^*, and bonuses are determined by h^*. In the simplest case, $h^* = h(e) + \eta$, where η is an unobserved mean-zero random shock. This simple case is a reasonable approximation when h^* is a teacher value-added score based on student tests, and $h(e)$ is the teacher's true contribution. Let

[16] Barrera-Osorio and Raju (2017) contribute related evidence from a field experiment in Pakistan. The treatment conditions contrasted incentives for only the school principal and incentives for both principal and teachers. Barrera-Osorio and Raju find no evidence of effects on student test scores of any of the incentive designs.

λ be the proportion of signal variance in h^*: $\lambda = \frac{\sigma_h^2}{\sigma_{h^*}^2} = \frac{\sigma_h^2}{\sigma_h^2+\sigma_\eta^2}$. As the signal weakens—$\lambda$ shrinks towards zero—any incentive attached to h^* also weakens. Put differently, as λ shrinks towards zero, h^* becomes less and less responsive to changes in effort, e. In the extreme, h^* becomes a random number and any bonus program becomes a lottery.

The insight from Brehm, Imberman, and Lovenheim (2017) comes, in part, by focusing on the case of a rank tournament with a strong contrast between winners and losers. The empirical data come from Houston's ASPIRE program. In elementary and middle schools, individual teachers competed for cash bonuses based on their individual value-added scores. Teachers ranked above the 50th percentile earned $2,500, and those above the 75th earned $5,000.[17] While teachers ranked below the 50th percentile received no bonus. With the sharp award cutoff at the 50th (75th) percentile, the incentive to give greater effort should be strongest for teachers near the cutoff.

Brehm and coauthors (2017) provide a formal framework for what "near the cutoff" means when the performance measure is noisy. In school year t the teacher chooses effort e_t. At the end of the year, she learns her value-added score, $h_t^*(e_t)$, and also learns the award cutoff, c_t. Then, for school year $t+1$, the teacher must choose effort e_{t+1}. If there is little to no noise in the performance measure, $\lambda \to 1$, then the teacher's incentive to raise her effort level, e_{t+1}, will be increasing in the inverse of her prior distance from the award cutoff, $\frac{1}{|h_t^*(e_t)-c_t|}$. Imagine a Gaussian curve, centered at c_t, which measures the predicted change in effort, $(e_{t+1} - e_t)$, due to the performance incentive. This is the "near the cutoff" intuition, but it assumes little to no noise. As the performance measure becomes

[17] Bonus amounts increased to $3,500 and $7,000 in the later years of ASPIRE. Imberman and Lovenheim (2015), discussed already, also study Houston's ASPIRE program, but in high schools where teachers competed in school-grade-subject teams.

noisier, the Gaussian curve flattens out. For very noisy measures, $\lambda \to 0$, there is no relationship between $h_t^*(e_t)$ and e_{t+1}. Additionally, the peak of the incentive-response curve shifts further and further to the right of c_t as the performance measure becomes noisier. This occurs because the teacher, whose goal is to score above the award cutoff, must insure herself against the risk of a negative draw of η.[18]

Using data from Houston, Brehm and coauthors (2017) test whether there were larger student achievement gains, h_{t+1}^*, from teachers whose prior performance, h_t^*, was nearer the award cutoff. The authors find no evidence that being nearer the cutoff changed teacher performance. Briefly, the test involves predicting h_{t+1}^* as a flexible but parametric function of h_t^*, using teachers whose baseline performance, h_t^*, is well above or below the award cutoff. Extrapolating that prediction to values of h_t^* near the cutoff provides the counterfactual estimate. The lack of effects in Houston is consistent with the incentives being weakened by a quite noisy performance measure. Brehm and coauthors estimate that, for the value-added scores in their setting, the proportion signal was 0.35-0.45, which is similar to estimates from other settings. The authors discuss alternative explanations for the lack of effects. One simple explanation is that none of Houston's teachers responded to the ASPIRE incentives at all, but that explanation is difficult to reconcile with the results in Imberman and Lovenheim (2015).

The Brehm, Imberman, and Lovenheim (2017) framework suggests new ways of thinking about the well-known POINT field experiment in Nashville, Tennessee. Springer and coauthors (2010) describe the POINT experiment and its general lack of effects on teacher performance. The Nashville teachers randomly

[18] This summary abstracts from some features of the problem discussed in Brehm, Imberman, and Lovenheim (2017). For example, the teacher's choice depends on her own marginal return to effort, and the teacher also faces some uncertainty about the new award cutoff c_{t+1}.

assigned to treatment competed in a rank ordered tournament based on their value-added scores, much like the teachers in Houston. The POINT experiment focused on math teachers in middle schools. In Nashville the awards were larger but the award cutoffs much higher. Teachers ranked above the 95th percentile earned a $15,000 bonus, and those above the 90th and 80th percentiles earned $10,000 and $5,000 respectively. Over three years of the experiment there were effectively no differences in student test scores between treatment and control teachers.

One explanation for the lack of effects in Nashville, despite the large bonuses, is that the award cutoffs were set too high. Neal (2011) points out that for about half of Nashville teachers the chance of winning a bonus was less than one in five, creating little incentive to change. Springer and coauthors (2010) find reasons to be more optimistic about the incentives for some teachers, but do not test for related effect heterogeneity. The Brehm, Imberman, and Lovenheim (2017) framework suggests that noise in the value-added scores pushes the effective award cutoffs even higher. A Nashville teacher who wanted to earn the $15,000 bonus would have work for a performance target well above the 95th percentile to insure herself against the risk of a negative draw of noise. This only reinforces the "cutoffs were set too high" explanation. Moreover, it is possible that Nashville teachers would not have responded to the POINT incentives even if the award cutoffs were lower, because noise should weaken incentives at any award cutoff level, as Brehm and coauthors point out.

Implicit in the Brehm, Imberman, and Lovenheim (2017) framework is that teachers themselves had some understanding of the noisiness of value-added scores. What teachers understand and how their understanding affects their response to incentives are clear opportunities for future contributions to the literature. Houston had apparently not reported individual teacher value-added scores prior to the ASPIRE program, and teachers may have had poor priors about

noise in value-added scores. However, Brehm and coauthors do not find differences between the early and later years of ASPIRE over which teachers might learn more about the instability of value-added scores. In Tennessee, by contrast, the state had provided teacher value-added score reports for many years before the POINT experiment. Thus, a teacher in Nashville likely better understood how value-added scores can change from year to year without any change in her own behavior. Rockoff and coauthors (2012), as part of an experiment in New York City described in section 3, provide some related evidence on how school principals respond to information about the noise in teacher value-added scores.

A widely-cited example of teacher incentives from Washington, DC provides a useful contrast to Nashville and Houston. Dee and Wyckoff (2015) describe DC's IMPACT system and study its effects on teacher performance and turnover. While not a rank tournament, bonus and salary incentives in DC were based on sharp cutoffs. For the average teacher, the IMPACT performance measure was mainly based on classroom observation ratings.[19] That continuous measure was then discretized into four ratings categories: ineffective, minimally effective, effective, or highly effective. Ineffective and minimally effective ratings carried the threat of dismissal, as discussed later in section 3. Teachers rated "highly effective," roughly the top scoring 15% of teachers, received cash bonuses of between $5,000-25,000. What's more, teachers rated "highly effective" for two consecutive years received a permanent increase in salary, worth $6,000-27,000 per year and as much as $200,000 in present-discounted career earnings. Dee and Wyckoff use regression discontinuity methods to

[19] For 83% of teachers in the Dee and Wyckoff (2015) analysis, the IMPACT performance measure is a weighted average of classroom observation ratings (75% weight), school value-added score (5%), and two relatively-subjective ratings from the principal (20%). The other 17% are teachers for whom the district calculated individual teacher value-added scores. For these teachers the weights were: teacher value-added score (50%), observation ratings (35%), school-value added (5%), and principal ratings (10%).

compare teachers near the cutoff between "effective" and "highly effective." Those who scored above that cutoff in year t, had a strong incentive to score "highly effective' again in year $t+1$, and thus capture the permanent salary increase. Teachers responded accordingly and performance in year $t+1$ increased by 0.24 standard deviations in the teacher IMPACT score distribution.

A notable difference between DC and the Houston and Nashville cases is the performance measure. Classroom observation ratings, like those used in DC, are not free of noise. Still, the observation ratings in DC were likely more reliable than the typical teacher value-added score (Kane and Staiger 2012, Ho and Kane 2013); each teacher was rated by two different observers in five total observations across the school year. When they do have teacher value-added scores, Dee and Wyckoff (2015) find smaller effects of the "highly effective" incentives on value added. The larger incentive effects in DC may partly be a result of a less-noisy performance measure. However, classroom observation ratings also create new opportunities for teachers to manipulate their scores. Additional analysis of the DC case suggests teachers give more effort when they are (or expect to be) visited by the observer (Phipps and Wiseman 2021), and that teachers focus on parts of the observation rubric which are easier to demonstrate successfully (Adnot 2016).

An additional contribution on this topic comes from the Pakistan experiment studied by Andrabi and Brown (2022). Recall that teachers were randomly assigned to bonuses based on student test scores (objective) or based on a rating by their school's principal (subjective). When asked in a survey, teachers facing subjective ratings were more likely, than objectively scored teachers, to agree with the statements "those who work harder, earn more," "I feel motivated," and "the raise is [in] my control." Whatever the actual noise in the two performance measures, or other properties, teachers seemed to feel stronger incentives when bonuses were based on subjective ratings. However, those differences did not lead to larger treatment effects on student test scores, at least

not on average. Andrabi and Brown do show that, within the subjective rating condition, the gains in student achievement were larger when the teacher agreed more with these three survey items.

One final note on the topic. School systems, aware that noisy value-added scores weaken incentives, often adopt one or both of the following strategies: First, school systems often use value-added scores which average over two, three, or more school years to measure the performance of a given teacher. This may reduce noise but weakens incentives in a different way. Consider a teacher who increases her effort in school year t, motivated by a potential bonus or tenure. The effect of that real effort will be muted if the performance measure averages over year t, $t-1$, $t-2$, etc. Second, school systems often use "shrunken" value-added scores, as described in section 1, which adjust for noise by moving a teacher's score closer to the mean when she teaches a relatively small class. The consequences of both these strategies have not been studied.[20]

2.6. Incentives for Inputs

Teacher performance incentives linked to output measures, especially student achievement test scores, dominate the empirical literature. Yet, in practice, teachers and schools are more often incentivized on measures of inputs. The typical classroom observation rubric is designed to assess inputs not outputs. For example, observers score the nature and frequency of questions the teacher asks her students, but observers do not assess whether these questions generated student learning. Two examples of teacher incentives based on observation ratings come from Cincinnati, Ohio (Taylor and Tyler 2012) and France (Briole and Maurin in-press). In both cases incentives based on observation ratings generated improvements in student achievement test scores which were not incentivized; both cases are discussed in detail in section 4. Hussain (2015) similarly finds

[20] In hospitals, Schwartz (2021) finds that shrinking weakens incentives for small hospitals.

improvements in student test scores caused by school inspections, which focus on teaching observations and other inputs. Other examples described in this chapter involve observation ratings as a substantial component of teacher performance measures (Dee and Wyckoff 2015, Ng 2021, Taylor 2022).

In some schools, the typical rates of teacher absence from work suggest incentivizing attendance. Chaudhury and coauthors (2006) measured teacher absence in unannounced visits to a sample of schools in Bangladesh, Ecuador, India, Indonesia, Peru and Uganda. One in five teachers were absent, on average. In India, one in four were entirely absent and only half were actively teaching students. Duflo, Hanna, and Ryan (2012) describe a field experiment in Rajasthan, India where teachers' salaries depended on their attendance at work. In control schools, teachers received a fixed salary of Rs. 1,000 per month. In treatment schools, teachers received a base salary of Rs. 500, plus an additional Rs. 50 for each day they worked beyond a 10-day minimum. Teacher absence rates fell by half, from 42% in control schools to 21% in treatment schools.

Attendance is a straightforward performance measure, and Rajasthani teachers clearly responded to the new incentives. However, teachers showing up to work does not necessarily improve student welfare. Teachers could work more days each month but give less effort on any one day. Still, in this case, students were better off. Student achievement test scores improved by 0.17σ. When at work, treatment teachers were just as likely to be in their classroom, using blackboards, and interacting with students. Cilliers and coauthors (2018) also report improvements in both teacher attendance and student outcomes in Ugandan schools which incentivized attendance. Moreover, simply monitoring and reporting attendance, without linked bonuses, had no effect in Uganda. Banerjee and Duflo (2006) compare evidence on a variety of strategies for reducing teacher absence, including reward and punishment incentives.

2.7. Complementary Inputs

Teacher effort is far from the only input in schooling, and increasing teacher effort may be an insufficient or inefficient strategy for raising student achievement. Mbiti and coauthors (2019) study the complementarity of teacher incentives and other school inputs in Tanzanian schools. In one randomly-assigned treatment condition, described in section 2.1, teachers received a cash bonus for each student who passed a year-end government exam. The Tanzania experiment also included two other treatment conditions. In a grant condition, schools received unconditional cash grants. In grant schools, spending increased by 60% on expenditures other than teacher salaries, and most of the funds were used for textbooks and other classroom materials. A third treatment condition combined both the grants and teacher incentives.

The experimental results show strong complementarities between the grants and teacher incentives. Neither the grants nor the teacher incentives alone had much effect. Student test scores increased 0.03σ in the incentive-only condition and 0.01σ in the grant-only condition, but neither estimate is statistically significantly different from zero. However, the combination of grants and teacher incentives improved test scores by 0.23σ—an effect nearly five times as large as the sum of the grant- and incentive-only estimates. This striking pattern of results is, first, inconsistent with teachers as motivated agents. Cash grants alone had no effect student achievement; a result at odds with the prediction that teachers are constrained by resources and not their own effort costs. Yet, the results also show that, at least in this case, teacher incentives alone were insufficient to improve student achievement. Teacher effort did respond under the incentive-only condition, recall that scores on the incentivized exam rose 0.21σ, but that effect did not generalize to the low-stakes exams. The notable gains in Tanzania, 0.23σ on low-stakes exams, required both resources and incentives.

Muralidharan and Sundararaman (2010) describe a different pattern in results from Andhra Pradesh. Alongside the pay-for-performance experiment (described in Muralidharan and Sundararaman 2011), some Andhra Pradesh schools were randomly assigned to receive additional resources. Specifically, this "feedback" treatment included (a) diagnostic test and feedback to teachers on individual students at the beginning of the school year, and (b) low-stakes classroom observations throughout the year. Similar to the Tanzania results, these additional resources alone had no effect, 0.002σ (standard error 0.045), on student test scores in Andhra Pradesh. Again, a result inconsistent with strongly-motivated agents, though the feedback resources may be less helpful than unconditional grants. Unlike Tanzania, in Andhra Pradesh teacher incentives alone did generate improvements in student achievement (Muralidharan and Sundararaman 2011), as did the combination of incentives and feedback (Muralidharan and Sundararaman 2010).[21]

This topic—the complementarities, if any, between teacher incentives and other educational inputs—is a clear opportunity for contributions to the economics of education literature. The Tanzania and Andhra Pradesh results are two examples suggestive of the potential. Also, from this perspective, the results in Behrman et al. (2015) and Jackson (2010, 2014) suggest the possibility of complementarities between teacher incentives and student incentives.

3. EVALUATION AND SELECTION

Teacher selection is a second common rationale for evaluation. Or rather two related rationales: selection by schools based on measured performance, and self-selection into or out of teaching in response to evaluation incentives.

[21] The Andhra Pradesh experiment lacks a low-stakes exam like the Tanzania experiment, so perhaps the incentive effects in Andhra Pradesh would not generalize to a low-stakes exam. However, in Andhra Pradesh, the teacher incentives did improve student achievement in non-incentivized subjects, science and social studies.

3.1. Selection by Schools

Retention or dismissal based on job performance is, at least conceptually, nothing new to teachers and schools. Schools have long subjected new teachers to probationary screening, often ending with successful teachers granted the strong employment protections of "tenure." Earning tenure nominally requires performance above some threshold. However, in practice, tenure decisions are usually not selective. In the New York City public schools just 2% of teachers were denied tenure before reforms in 2010 (Loeb, Miller, and Wyckoff 2015, Dinerstein and Opper 2022), and denials were even rarer in Chicago before reforms in 2004 (Jacob 2011). Teachers may self-select in response to tenure expectations, but involuntary dismissal by schools is rare.

Interest in selecting teachers based on performance has been reenergized in recent years for two or three reasons. First, over the last decade or so, teacher value-added scores have become a plausible alternative to traditional classroom observations, in the role of selection criterion. A simple but critical advantage is that value-added scores reveal meaningful between-teacher differences in performance—differences to select on—while observation ratings typically do not differentiate as much among teachers. Second, over the same decade plus, economists and other researchers have tried but failed to identify predictors of teaching performance which could be used to screen prospective new hires (Jackson, Rockoff, and Staiger 2014 provide a review).[22] By contrast, observed performance on the job, even in just a teacher's first year, is a useful predictor of career performance. A third reason for interest in selection is that some read the

[22] Interest in teacher hiring decisions has also been reenergized recently, including perhaps some new optimism about pre-hire screening (for example, Jacob et al. 2018, Estrada 2019, Araujo et al. 2020). Another related problem is how schools select teachers for layoffs caused by unanticipated budget cuts, and the potential role of performance measures (for example, Boyd et al. 2011, Goldhaber and Theobald 2013, Kraft 2015).

evidence in section 2 as lacking clear empirical support for performance incentive strategies.

The school's selection decision, in its most basic terms, is between two options: (a) retain a low-performing teacher, or (b) replace the low-performer with a new hire drawn from the distribution of potential hires. Consider the potential benefits of choosing (b) over (a).

Hanushek (2011) estimates the gains from replacing the bottom 5-10% lowest-performing teachers with average performing teachers. Student achievement scores would improve by nearly half a standard deviation, accumulated over 12-13 years of schooling. Based on those achievement gains we would predict large future earnings gains for students; lifetime gains in present value of $10,000-20,000 per classroom per year. Chetty, Friedman, and Rockoff (2014b) reach a quite similar conclusion: replacing the bottom 5% of teachers is worth $9,000 per classroom per year in lifetime earnings. The important contribution of this estimate is that Chetty and coauthors can directly estimate the causal effect of higher-value-added teachers on students' future earnings. These estimates reveal large potential gains, in either achievement or earnings terms, but they make some simplifying assumptions.[23] One assumption is that the replacements will be average teachers—specifically, the new hire in option (b) will have a value-added score equal to the mean of the current teacher distribution. The following paragraphs discuss these assumptions and other features which have been the focus of analysis by Staiger and Rockoff (2010), Neal (2011), and Rothstein (2015).

Choosing the new hire, option (b) over (a), comes at a cost in the short run. The cost arises because the new hire, option (b), will be a novice teacher.

[23] These achievement and earnings estimates require assumptions about the true variance of teacher effects on achievement, how achievement scores relate to future earnings, average class size, and others. Hanushek (2011) provides estimates under various assumptions, and the gains remain quite large.

One individual school might replace a low-performer by convincing a successful veteran to switch schools, but that leaves the second school needing to hire, and in the end the school system will need to hire a novice. We can thus restate the options as: (a) retain a low-performing teacher, who has at least some experience; or (b) replace the low-performer with a draw from the novice teacher distribution. Choosing (b) over (a) will likely make the school worse off in the short run. Typical value-added scores for first-year teachers are roughly one standard deviation below average, and half a standard deviation below second-year teachers (Staiger and Rockoff 2010, Papay and Kraft 2015). Thus, during the new hire's first year on the job, she may very well be a less effective teacher than the teacher she replaced would have been if he had been retained.[24] These achievement losses predict future earnings losses for the students affected. However, these costs are short run. We would expect the novice new hire to improve year over year, and eventually outperform the teacher she replaced, or else be subject to replacement herself.

The problem is intertemporal. Think of options (a) and (b) as two streams of job performance over two teaching careers, and thus best compared in present value terms. Moreover, the (a) versus (b) choice will reoccur any time a new hire turns out to be lower performing than predicted. Thus, we might think of the school's choice as between two personnel strategies: (i) do not dismiss teachers based on observed performance, or (ii) choose (b) over (a) whenever the choice improves expected performance. Each of these strategies has a present value, each summing over several individual teachers' careers.

Staiger and Rockoff (2010) and Neal (2011) both derive optimal dismissal rules given these intertemporal features. Both find quite high dismissal rates:

[24] This cost is less likely when the replaced teachers are only the lowest-performing 5-10%, whose value added would be more than one standard deviation below average, but the cost becomes more salient for the higher dismissal rates we will soon consider.

Dismiss the 50-67% lowest-performers after their first year teaching. Then small fractions in subsequent years, for a total dismissal rate of 60-85% of a given hire cohort. Staiger and Rockoff estimate that average student achievement would be about 0.10σ higher in the new steady state (σ = student standard deviations), though that might take a decade or more to reach. Even a 50% dismissal rate is much higher than the current roughly 10% turnover after one year teaching (NCES 2016), and that 10% total turnover includes some high-performing teachers who leave for other reasons. Neither Staiger and Rockoff nor Neal suggest such dismissal rules should be policy proposals, and neither consider how teachers' actions or choices might change in response to such dismissal programs. Rather, these optimal-rule exercises sharpen our understanding of the problem's relevant features. Staiger and Rockoff, notably, cover a broad range of considerations and their effect on the optimal rules.

Here I highlight three features. First, the Staiger and Rockoff (2011) analysis includes "tenure," while Neal (2011) does not. Staiger and Rockoff derive optimal dismissal rules under the constraint that retained teachers are granted employment protections (tenure) after t years. This partly explains why the Staiger and Rockoff rates are higher than Neal, 67-85% versus 50-60% respectively. Tenure protections raise the value of selectivity before tenure. Second, the Neal approach maximizes present value, while Staiger and Rockoff maximize steady state outcomes. By explicitly discounting, Neal emphasizes the up-front cost of choosing (b) over (a)—the losses in student achievement because of the (typically) low-performance of novice teachers. This further contributes to Neal's lower dismissal rates. Staiger and Rockoff also discuss this issue, and report that the optimal dismissal rate in their analysis remains well above 50% for typical discount rates.

Third, Staiger and Rockoff (2011) give particular attention to the (un)reliability of value-added scores. A key limitation when using value-added

scores to make inferences about individual teachers, as discussed earlier, is the non-trivial measurement error. Critics of selection rules based on value-added scores emphasize the "false negative" problem: some teachers with true value added above the threshold will nevertheless be dismissed because of a bad draw of measurement error. Such mistakes are an important cost to account for. The papers cited in this section (nearly) always account for the typical reliability of value-added scores. A contribution form Staiger and Rockoff is to simulate how the optimal dismissal rules, and benefits of selection, change with higher or lower reliability (synonymously, proportion signal variance). Perhaps not surprisingly, the benefits increase almost linearly with reliability. However, optimal dismissal rates are much less sensitive to reliability, only falling sharply when reliability is below 0.05. Even when reliability is 0.30, the optimal rule is to dismiss 80% at the end of the first year. In short, acting on weak signals is still valuable, because tenure protections can lock in career-long streams of low performance, though stronger signals would provide even larger gains from selection.

In recent work, Dinerstein and Opper (2022) study selection by schools, but with a focus on the potential multitask problem. For simplicity, assume teachers' job responsibilities can be divided into two types: contributions to student skills measured by standardized tests (and reflected in teacher value-added scores), and contributions not measured by the school system. Evidence from Jackson (2018), among others, suggests teacher performance on these two dimensions is imperfectly correlated. This raises the concern that dismissal rules based on value-added scores alone risk dismissing teachers who make other positive but unobserved contributions. However, this concern will not necessarily occur, as Dinerstein and Opper show both theoretically and empirically.

Dinerstein and Opper (2022) develop a model of probationary screening with multitasking. Schools announce a dismissal rule based on value-added scores; teachers respond by reducing effort on unmeasured contributions and

increasing effort on their value-added score contributions. Dinerstein and Opper then describe conditions under which the value-added dismissal rule will select teachers with higher unmeasured contributions. To see the basic idea, imagine two teachers with the same value-added score absent the dismissal program. However, the first has a comparative advantage in value-added score contributions, and the second a comparative advantage in unmeasured contributions. Absent the dismissal rule, both teachers allocate their effort following their advantage, with the goal of maximizing their own total contribution. Thus, the second teacher is initially making larger unobserved contributions than the first, even though their observed value-added scores are the same. When the dismissal rule is announced, the second teacher shifts effort toward increasing her value-added score, and thus reduces her chances of being dismissed. The first teacher has much less scope to increase his value-added score. In the end, the second teacher is more likely to be retained. Dinerstein and Opper find empirical support for these predictions using data from the New York City schools, which adopted a new tenure policy in 2009 that emphasized teacher value-added scores based on student tests.

3.2. Teacher Self-Selection

Evaluation and incentives may well influence (prospective) teachers' own choice between teaching and an alternative occupation. Mechanisms which predict positive self-selection are a simple rational for teacher evaluation. But self-selection can also complicate other rationales.

The (prospective) teacher's basic decision is between (i) beginning or continuing a career in teaching, or (ii) working in the best alternative occupation.[25] Assume the teacher chooses between (i) and (ii) to maximize

[25] As with occupational choice in general, (prospective) teachers' choice between teaching and alternative work is influenced by expected compensation, suggesting meaningful scope for

expected compensation (or utility) over her remaining career. When the compensation in teaching depends on a job performance measure—through earning tenure protections or performance bonuses—the choice between (i) and (ii) will depend on the teacher's own beliefs about her likely measured performance. This choice can be formalized using a Roy model of self-selection; Brown and Andrabi (2021) is one example. Brown and Andrabi add a useful distinction between self-selection on ability and on anticipated response to incentives, both of which contribute to the teacher's own beliefs about her likely measured performance.

Rothstein (2015) extends the optimal dismissal rule analysis to include teachers' self-selection response. As the proportion dismissed before tenure increases, a novice teacher's expected compensation from a teaching career falls; and fewer individuals will choose teaching over the alternative occupation, unless schools simultaneously raise teacher salaries. Expected compensation falls even for a relatively high-performing novice, who must consider the possibility that a bad draw of measurement error in her performance measure will place her below the threshold. In Rothstein's simulation, the optimal dismissal rate is about 40% after 2-3 years, but that rate requires a 25% increase in the average teacher salary. The simulation assumes schools pay for higher salaries by employing fewer teachers and increasing class size. The gain in student achievement is about one-quarter of the gain under the Staiger and Rockoff (2010) optimal rules, mainly because larger classes reduce achievement. In the end, the anticipated gain in students' future earnings is smaller but still a substantial improvement over the typical dismissal rate of nearly zero.

One insight from Rothstein (2015) focuses on the potential role for differential self-selection—self-selection correlated with true performance or

---

improving schooling through teacher self-selection (for example, Hoxby and Leigh 2004, Nagler, Piopiunik, and West 2020).

ability. In Rothstein's results, performance-based dismissal policies do not induce more high-ability teachers (or fewer low-ability teachers) to begin a career in teaching. As the prospective teacher anticipates, dismissal rules are based on measured performance not true performance, and so noisy performance measures weaken any incentive to self-select based on her private information. The incentive is further weakened by the prospective teacher's own uncertainty about her true ability. In short, there is little differential self-selection at initial entry into teaching. After one year on the job, teachers can update their beliefs based on their first actual performance measure, and some choose to an alternative occupation at that point. Still, in the end, most performance-based selection in Rothstein's analysis is selection by schools not self-selection by teachers.

3.3. (Quasi-)Experimental Evidence

Several recent studies include (quasi-)experimental results relevant to selection rationales for evaluation, but the evidence base remains small. First consider settings where the school threatens to dismiss teachers based on performance measures.

A widely-cited example of turnover under a clear dismissal rule comes from Washington, DC's teacher evaluation program, studied by Dee and Wyckoff (2015). Recall from section 2 that DC teachers are annually rated one of ineffective, minimally effective, effective, or highly effective. By rule, teachers rated "ineffective" (roughly the bottom 1%) are dismissed immediately, and teachers rated "minimally effective" (below roughly the 10th percentile) in two consecutive years are also dismissed. Following these rules, DC fired 4% of its teachers in the first two years of the new rules. Another 4% of teachers voluntarily quit after they received their first minimally-effective rating.[26]

[26] Dee and Wyckoff (2015) study the first 2-3 years of the new evaluation program, but Dee, James, and Wyckoff (2021) show that these patterns generally continued beyond those early years.

Without a counterfactual, however, it is unclear how many of those 8% would have left or been fired without the new evaluation program.

DC's dismissal rules did change teachers' self-selection decisions. Dee and Wyckoff (2015) compare teachers at the margin between "minimally effective" and "effective" ratings, estimating regression discontinuity treatment effects. Teachers rated minimally effective were more likely to voluntarily quit the district, increasing the probability of turnover from roughly 20-30%. Among those who did continue working in DC the next school year, $t + 1$, teachers rated minimally effective in year t performed better, on average, than those rated effective in t. The RD estimate is one-fifth to one-half of a standard deviation increase in the teacher value-added score distribution. That increase in performance could be (partly) caused by self-selection—quit versus remain—based on teachers' own private information about their likely future performance. If there had been no effect on turnover, we would likely conclude the increase in performance was caused by a change in teacher effort, and effort could also partly explain the increase. Additionally, even with the increase in turnover, DC's subsequent new hires were higher performing than the teachers they replaced (Adnot, Dee, Katz, and Wyckoff 2017).

Dismissal rules in other school systems, beyond the DC example, have had little effect on turnover. Many school systems do have seemingly similar rules (Winters and Cowan 2013). For example, in New Jersey teachers can be dismissed if they are rated "partially effective" in two consecutive years. In contrast to DC but using the same RD approach as Dee and Wyckoff (2015), Ng (2021) finds no difference in turnover between teachers rated partially effective versus those rated effective. At the "partially effective"-"effective" margin 90% of teachers were retained regardless of rating. And the dismissal threat had no effect on teacher value-added scores. Brunner and coauthors (2019) also find little effect on turnover under similar rules in Michigan.

One potential explanation is that dismissal threat in DC was more credible or salient to teachers than it was in New Jersey. In New Jersey just 1-2% of teachers scored below the "partially effective" cutoff, compared to 10% in DC below the equivalent cutoff. While DC followed-through and fired teachers after two consecutive poor ratings, in New Jersey two poor ratings only increased the probably of exit somewhat, from 10% to 20% at the cutoff. Moreover, even in DC the effects on exit were much larger in the new evaluation program's second year compared to the first year. That pattern, Dee and Wyckoff (2015) argue, is consistent with teachers being initially skeptical of the district's new dismissal threat, but much less skeptical once teachers were fired at the end of year one.

Several U.S. states and schools districts also now have score-based rules for granting tenure. For example, Tennessee teachers must be rated "above expectations" or higher (above roughly the 33rd percentile) in two consecutive years to earn tenure. However, quite unlike the simulated policies discussed earlier (Staiger and Rockoff 2010, Neal 2011, Rothstein 2015), teachers who fail to earn tenure continue to work until they meet the performance score requirement. In New Jersey teachers must be rated "effective" or higher (above the 2nd percentile roughly) in two out of three years. New Jersey teachers who fail to earn tenure leave their current district, but can go to work in other districts where the tenure process resets. Taylor (2022) and Ng (2021), studying Tennessee and New Jersey respectively, both find that pre-tenure teachers' value-added performance improves in response to the new tenure requirements, but differ on what happens to after tenure. Both also find little evidence of selection into teaching.

The selection rational for evaluation does not require formal rules. New selection patterns may arise simply from new (or newly symmetric) information about teacher performance. Rockoff and coauthors (2012) describe a field experiment in New York City where school principals, in randomly assigned

treatment schools, were provided reports on value-added scores for their teachers.[27] In response to the new information, principals updated their subjective opinions of teachers to align more closely with the value-added scores. Principals updated more when their priors about a specific teacher were less precise, and when the value-added scores were more precise. In treatment schools, lower-value-added teachers were more likely to leave the school, even though principals were not required to use (or even asked to use) the new reports in their personnel decisions. In control schools, turnover remained uncorrelated with value-added scores. The following school year, student achievement scores increased in treatment schools, though turnover cannot fully explain the improvement.

Similar evidence comes from Chicago (Sartain and Steinberg 2015) and Houston (Cullen, Koedel, and Parsons 2021). Both cases have (quasi-)experimental introduction of new performance measures. In Chicago there were no formal consequences attached. In Houston there were no formal dismissal rules, but the school system created incentives to encourage dismissal and retention based on teachers' performance ratings. In both cases, as in New York, treatment induced a negative correlation between performance and exit. However, in Houston the change in turnover was too small to affect student achievement.

While exits are easy to observe, the deliberations leading up to an exit are not, making it difficult to separate self-selection from selection by schools. The administrative data collected by most school systems, do not differentiate between teachers who quit truly of their own volition, and teachers who quit under "counseling out" pressure from their school administrators. Both mechanisms involve some degree of self-selection. But given the data limitations, clear evidence of selection by schools is rare in the empirical literature. Two exceptions are Bates (2020) and Jacob (2011, 2013).

[27] This experiment preceded the tenure reforms studied in Dinerstein and Opper (2022).

Bates (2020) studies a quasi-experiment in North Carolina similar to the New York experiment from Rockoff and coauthors (2012). One school district, Gilford County, began providing value-added score reports to teachers and principals in 2000, several years before the rest of the state. What's more, principals could view reports for teachers they were considering hiring away from another school in the district. Subsequently, teachers with higher value-added scores were more likely to move to a new job in a more-desirable school in Gilford County. By contrast, teachers with lower value-added scores moved to schools in other school districts, schools where the principals did not know the new hire's value-added score from Gilford. The contrast demonstrates the influence of new performance information on principals' hiring decisions, but may also reflect teachers' decisions about where to apply given what principals know about them.

Jacob (2011, 2013) focuses on school principals' dismissal decisions. In 2004, principals in the Chicago Public Schools were given new authority to fire any probationary teacher for any reason, without the typical costs of paperwork and hearings. The available performance information did not change, but principals could now more easily act on any latent preferences they had for selecting teachers. Turnover among novice teachers increased from roughly 9-18% under the new rules. Principals dismissed more teachers with low value-added performance, low prior evaluation ratings, and those more often absent from work. Teacher absences fell by 10% on average, mostly through turnover and less through changes in individual teachers' behavior. And there is some evidence student achievement improved at previously low-performing schools.

A final topic is teacher self-selection in response to the offer of performance incentive contracts. Selection could well have been part of the rational for some of the evaluation systems described in section 2, but the (quasi-)experimental studies in section 2 were mainly designed to detect incentive

effects. Two recent experiments were designed to separate selection from incentive effects.

Studying performance incentives in Rwandan schools, Leaver and coauthors (2021) separate selection effects from effort effects by using a novel two-stage field experiment. During the first stage, advertisements for teaching positions included one of two contract offers: pay for performance (P4P) or fixed wage (FW). Offer type was randomly assigned at the level of labor market, defined by geographic district and subject taught. In the second stage, after new hires were in their jobs, the teacher's actual contract, P4P or FW, was assigned by a second randomization at the school level. Teachers and schools were not aware of the second randomization until it occurred. In P4P schools, the top 20% of teachers received an annual bonus equal to about 15% of their base salary. To determine P4P bonuses, the performance measure was a weighted average of contribution to student achievement scores, teacher attendance, and a classroom observation rating.[28]

The offer of pay-for-performance incentives had little effect on selection into teaching jobs. Over the two year experiment, new hires recruited with a P4P offer contributed no more or less to student achievement scores than did FW-offer hires (point estimate 0.01σ, randomization inference p-value = 0.75). By contrast, new hires actually working under a P4P contract increased student scores by 0.12σ (p-value = 0.01), and that actual-contract effect did not differ by advertised offer type (p-value = 0.51).

While that pattern is consistent with little or no selection effect, Leaver and coauthors (2021) add further evidence. At baseline the experimenters also measured teaching skills and intrinsic motivation toward students. New hires

[28] All teachers were given a one-time signing bonus that ensured no one was worse off as a result of the re-randomization. In FW schools, all teachers received an unconditional annual bonus of about 3 percent, making the payouts equal in expectation.

recruited with a P4P offer were no more or less skilled than those recruited with a FW offer, though they were somewhat less intrinsically motivated.[29] About one-fifth of teachers did quit or otherwise leave their new jobs over the experiment's two years. However, new hires working under P4P and FW contracts were equally likely to stay or leave, and contract type had no effect on the correlation between turnover and skills or intrinsic motivation. Additionally, Leaver and coauthors find no difference in schools' hiring decisions between P4P and FW markets.

A second novel field experiment—with some contrasting results—was conducted by Brown and Andrabi (2021) with private schools in Pakistan.[30] Similar to Leaver and coauthors (2021), schools were randomly assigned to either a pay for performance (P4P) or fixed wage (FW) contract. In P4P schools, teachers received an annual bonus of between 0-10% of salary depending on their performance rank. However, Brown and Andrabi study selection patterns among incumbent teachers, not just new hires. The key design feature is that, prior to randomization, each teacher chose her own contract type, P4P or FW, knowing she would get that chosen contract with probability one-third. The experimenters observe a second contract choice when some teachers move from a FW to P4P school, or vice versa.[31]

Across several results, Brown and Andrabi (2021) document sorting to P4P or FW based on teachers' (private) expectations about their performance and compensation. First, teachers who chose P4P at baseline were higher performing at baseline. The average difference in value-added scores was 0.05σ or about one-

[29] Teaching skills were measured by asking teachers to grade a student exam and evaluating the accuracy of their grading. Intrinsic motivation was measured by having teachers play a lab-in-the-field version of a framed dictator game.

[30] This is the same field experiment as Andrabi and Brown (2022) which emphasized the experimental variation in objective and subjective performance measures.

[31] The remaining 2/3 of schools were the schools randomly assigned to P4P or FW for the entire school. In the FW contract, teachers received an unconditional annual bonus of 5 percent. The experiment involved other details and results beyond the scope of this summary.

third of a standard deviation in teacher performance. That 0.05σ difference is effectively unchanged after controlling for the principal's baseline rating of the teacher, suggesting teachers do have private information about their performance. Moreover, the correlation between value-added score and choosing P4P was stronger when the teacher had better information about her own performance—better because she had more teaching experience, or because she had been explicitly told her value-added percentile before her contract choice (as a randomized treatment). Second, after the first school year of the experiment, higher-value-added teachers were more likely to move from FW to P4P schools, and lower-value-added teachers the reverse. These moves were the teacher's choice alone, effectively, and were more likely when a school with the teacher's preferred contract type was located nearer to her current school.

Last, teachers also self-select based on their own effort response to performance incentives. Among teachers who chose P4P, value added increased by 0.09σ if their school was randomly assigned to P4P. That improvement is nearly two-thirds of a standard deviation in teacher job performance. In contrast, the P4P effect was just 0.01σ among teachers who chose FW. This sorting on effort response is distinct from sorting on prior performance. The 0.09σ versus 0.01σ difference does not change after allowing for effect heterogeneity by baseline value-added score. Brown and Andrabi (2021) estimate that the total sorting effect is about two-thirds sorting on prior performance and one-third sorting on effort response.[32]

What might reconcile the sorting effects in Brown and Andrabi (2021) with the relative lack of sorting effects found by Leaver and coauthors (2021)? First, Leaver and coauthors test for self-selection among novice new hires who,

[32] In a related quasi-experiment, Goldhaber and Walch (2012) study the introduction of new performance measures and incentives in Denver, CO, where teachers could choose to participate in the incentives or not. Higher performing teachers were more likely to participate, but performance improved among both participants and non-participants.

compared to experienced teachers, likely have little information about their own potential performance scores, and thus little information to sort on. The lack of selection among novice hires matches Rothstein's (2015) prediction. Brown and Andrabi also find that among inexperienced teachers (less than three years) there is little to no correlation between value-added score and choosing P4P. Second, the selection effects in Brown and Andrabi are smaller when the switching costs are higher. The contract choice design has zero switching costs, since teachers do not change jobs, and the resulting sorting effects are about equal to the effort effects. The estimated selection effects are much smaller when teachers change schools, and smaller still when the switching costs are higher.

Biasi (2021) studies a related quasi-experiment in Wisconsin. In 2011 the state legislature ended required collective bargaining over teacher compensation. About half of local school districts nevertheless continued to use traditional seniority-based salary schedules. The other half adopted new more-flexible compensation strategies, including differentiating pay based on performance. Subsequently, as Biasi documents, higher value-added teachers were more likely to leave seniority-pay districts and take new jobs in flexible-pay districts. In their new jobs they earned more and increased their effort at work. By contrast, lower value-added teachers moved to seniority-pay districts or left teaching. While the average Wisconsin student was better off, sorting under flexible pay could widen achievement gaps among students (Biasi, Fu, and Stromme 2021) and gender pay gaps among teachers (Biasi and Sarsons 2022).[33]

Other research on teacher compensation is also consistent with the selection rational for performance incentives. Briefly, first, even when pay is unrelated to performance in teaching jobs, pay outside teaching can affect

[33] Similarly, Burgess, Greaves, and Murphy (2022) estimate the effects of teacher pay deregulation in England in 2010. In more-competitive labor markets, the option to differentiate pay based on performance generated improvements in student achievement.

selection into or out of teaching. For example, Nagler, Piopiunik, and West (2020) show that teachers hired during a recession, when outside options are poor, have higher value-added performance. Chingos and West (2012) and Leigh (2012) are other examples. Second, Johnston (2021) measures teacher preferences over different forms of compensation in a discrete-choice experiment. On most dimensions—base salary, retirement benefits, health insurance—higher and lower performing teachers have similar preferences. The exception is performance bonuses, where higher performing teachers do value the potential bonus much more.

Finally, predictions about the benefits of selection effects are usually predictions about some long-run equilibrium. Well-designed experiments can provide important insights but typically cannot test long-run predictions. To study the long-run, Woessmann (2011) compares 28 OECD countries on the PISA exam. In about half of the 28, teacher pay is based partly on performance, erring on the side of inclusion. Students score a quarter of a standard deviation higher in the performance pay countries, Woessmann estimates, likely as a result of some combination of selection effects and effort effects.

4. EVALUATION AND SKILL DEVELOPMENT

A third rationale focuses on skill development mechanisms: that evaluation can improve job performance through causal effects on teachers' job skills. An important prediction here is that evaluation's effects on performance can persist after the evaluation measures or incentives end because skills persist (even if with some depreciation).

Return to the teacher's problem in (1) but make teacher skill, s, an explicit input to the teacher's contribution to student achievement:

$$\max_{e} \quad u = w + v[h(e, s)] - c(e),$$

where effort and skill are complements in $h(e,s)$. To start, assume the school offers no incentive linked to $h(e,s)$, but that the teacher is a motivated agent, at least to some degree, $\partial u/\partial h > 0$. Imagine an exogenous increase in s. The teacher's optimal choice of effort, e, would (weakly) fall but her contribution, h, would rise and so too would her utility. That kind of potential gain should motivate the teacher to try to improve her skills, s, even without any evaluation or incentives. But learning new skills requires effort. In the short run, effort invested in skill development, f, trades off with effort invested in current production, e. We might rewrite the teacher's problem:

$$\max_{e,f} \quad u = w + v\big[h\big(e, s(f)\big)\big] - c(e+f). \tag{2}$$

The opportunity cost of improving teaching skills is a reduction in current contribution to student achievement.

Evaluation can shift the teacher's problem in favor of investing in skills in two ways. First, evaluation can reduce the effort required to learn new skills, that is, reduce the cost of skills. Evaluation creates new information: individualized feedback about current performance, often with comparison to coworkers' performance. For example, classroom observation rubrics typically score teachers on a dozen or more specific skills and tasks, with scope to reveal relative strengths and weaknesses. Acquiring that new information requires relatively little new effort from the teacher herself, and otherwise she would be left to self-assessment. That new information can direct development effort, f, toward skills which are higher return investments for the individual. Additionally, even before (or without) being scored, the evaluation instruments may help reduce the costs of skill development. For example, classroom observation rubrics typically have detailed written descriptions of what teacher actions and behaviors reflect low, average, and high performance. Those practical descriptions suggest what a

teacher might do differently in practice, and also what the school expects teachers to do. See Figure 1 for examples of rubric descriptions.

An evaluation as "feedback for improvement" rationale is common in the education sector (for example, Darling-Hammond 2015). Advocates for "formative evaluation" emphasize that extrinsic incentives linked to scores are not necessary for evaluation to generate improvement in teaching.

The second way evaluation changes the teacher's problem is by increasing the returns to learning new skills. But first consider the returns on skill investments without evaluation. The option to invest in skills—the choice of f—makes the teacher's problem an intertemporal problem. The cost of improving one's skills is a reduction in current students' achievement, while the return is an increase in future students' achievement. Borrowing the familiar features of human capital investment models (Becker 1962, Ben-Porath 1967, among others) we can write the teacher's problem as:

$$\max_{\mathbf{e},\mathbf{f}} \quad u = \sum_{t=0}^{T} \{w_t + v[h(e_t, s_t)] - c(e_t + f_t)\} \frac{1}{(1+r)^t} \qquad (3)$$

$$\text{where } s_t = g(s_{t-1}, f_{t-1})$$

In any given period, t, the marginal costs of e_t and f_t are the same. The marginal benefits of e_t occur only in period t, while the marginal benefits of f_t are a stream of returns accumulating over the teacher's remaining career, $t+1$ to T.[34] Taylor (2022) discusses this inter-temporal version of the teacher's evaluation and incentives problem.

Performance incentives increase the returns to learning new skills, if the teacher expects repeated evaluation and performance incentives over time. The potential stream of future rewards creates an incentive for the employee to invest

[34] In practice many measures of teacher performance, like value-added scores, are annual measures. If the periods in (3) are school years, then it may be possible for the returns on skill investments to begin during the same school year, that is, some return on f_t that shows up in h_t.

in skills. By contrast, if the school offered a performance reward for only one period, there would be no new incentive to invest more effort in skill development. The new stream of returns will only last while the performance incentive scheme lasts, and thus skill-investment incentive depends on the expected duration of the scheme. That incentive will be weakened by skepticism that the rewards program will last for very long.

Three insights are worth highlighting. First, extrinsic performance incentives can cause improvements in teaching skills, even without an assumption of "motivated agents," by creating a return on skill investments. Second, feedback from evaluation may also improve teaching skills, by lowering the cost of skill investments. The feedback (lower costs) and performance incentives (higher returns) mechanisms are complementary in motivating skill investments, but the feedback mechanism does not require extrinsic rewards. Third, if evaluation causes improvements in teaching skills, then the effects of evaluation on performance can persist after the end of any evaluation measurement or linked performance incentives, because skills persist over time.[35] By contrast, if teachers only increase their current effort to capture performance incentives (the agency theory rational), then effort and performance will return to prior levels once the incentives end.

Empirical evidence of skill development effects remains hard to come by. Skill development could contribute to the effects of most performance incentive programs, and thus could partly explain the reduced-form estimates of programs discussed already in this chapter. But evidence of skill effects, separate from effort effects, requires testing predictions specific to skill development.

[35] As skill depreciation accumulates over time, evaluation effects will weaken more and more. Dinerstein, Megalokonomou, and Yannelis (2020) tests for skill depreciation among teachers; even in that case, where teachers stopped teaching entirely for one or more years, skills persisted to some extent.

The first such prediction is that effects of evaluation on teacher performance will persist after evaluation measures and incentives end, because the skill gains will persist. Taylor and Tyler (2012) and Briole and Maurin (in-press) report evidence of persistent effects, studying teachers in Cincinnati, Ohio and France respectively. In both settings each teacher was scored during one school year, using a classroom observation rubric, but then not scored again until several years later. The timing of the evaluation year was determined by an administrative schedule, not by prior performance. In Cincinnati, teachers with very low scores were threatened with dismissal even if tenured; in France, scores partly determined salaries. Both papers measure teacher performance—contributions to student test scores (value-added estimates)—over several years before, during, and after the evaluation year. Though, notably, student test scores were not used in either evaluation program. Consistent with skill development effects, in both Cincinnati and France teacher value-added increases in the evaluation year and then remains higher in the years that follow.

Analysis of such long-run effects is a clear opportunity for future contributions. Taylor and Tyler (2012) and Briole and Maurin (in-press) show effects over several years post evaluation. Other studies include shorter post-evaluation periods and find mixed results on persistence. In the New Jersey evaluation program, described in section 3, the key incentive linked to performance is earning tenure and its employment protections. Ng (2021) finds that in the year after earning tenure teacher value-added declines in math, but not in English language arts. Santibanez and coauthors (2007) study a performance-based promotion program in Mexico called Carrera Magisterial, where the promotion came with a salary increase. Promotions were based on several measures, but, for a minority of teachers, additional effort in the classroom could make the difference between earning the promotion or not. For that minority, student test scores improved during the evaluation, but declined once the teacher

had been promoted. However, the evaluation in Mexico may have been too brief to generate skill investment incentives.

A second prediction is that evaluation can affect performance though skills, even without rewards or punishments linked to performance measures. Burgess, Rawal, and Taylor (2021) conduct a field experiment in England with new performance measurement but no formal incentives. Teachers were observed and scored by peers—other teachers working in the same school—using the Framework for Teaching observation rubric. Student test scores improved in treatment schools, consistent with a skill development mechanism. A possible alternative mechanism is that teachers gave more effort because they were incentivized by the social pressure of peer evaluation. However, teachers were only observed 1-3 times per year, and greater effort only during peer observations alone would not have improved student test scores. Steinberg and Sartain (2015) study teachers in Chicago who were scored with an observation rubric, but by their school principal. Student achievement improved in language but not math. One counter example is the feedback-only experiment in Andhra Pradesh, India. Muralidharan and Sundararaman (2010) find no effect on student test scores in schools that received low-stakes classroom observations, but no incentives, in their treatment bundle.

The New York City value-added score experiment, studied by Rockoff and coauthors (2012) and described in section 3, is also a case of new performance measures without linked incentives. When schools received value-added score reports, student test scores improved the following year. Teacher turnover cannot fully explain that improvement, suggesting teachers also changed their behavior in response to the new performance information. In a similar quasi-experiment, Bergman and Hill (2018) and Pope (2019) study the publication of individual teachers' value-added scores by the Los Angeles Times. Pope reports improvements in student achievement after publication for relatively low-

performing teachers, but no change among high-performing teachers. Bergman and Hill find evidence that the new information changed how students were assigned to teachers, and that sorting may partly explain the apparent improvements in Pope's analysis.[36]

A third prediction is that evaluation measures and performance incentives should be complements in improving job performance, if the mechanism for improvement is skill development. Testing this prediction requires (quasi-)experimental variation separately in both measures and incentives, making empirical opportunities scarce. Taylor (2022) finds evidence of that complementarity in data from Tennessee. All Tennessee teachers began receiving the same new performance measures starting in 2011-12, including classroom observation ratings using a detailed rubric. While the performance measures were the same, the consequences varied among teachers. For recently-hired teachers, new tenure rules required a teacher to score above a certain threshold in her fourth and fifth years of employment in order to earn tenure. Already-tenured teachers were not subject to the new rules and had no consequence attached to the new performance measures. Together these features create within-teacher over-time variation in performance measurement, and between hiring cohort variation in linked incentives. The new measures improved performance for all early-career teachers, as measured by their contributions to student achievement scores, both teachers with incentives to improve and those without formal incentives. But the performance gains were roughly twice as large for teachers facing the new tenure incentive linked to their scores.

A fourth prediction is that, if the mechanism is skill growth, the effects of evaluation should be heterogeneous—larger effects for lower-performing teachers

[36] Garet and coauthors (2017) report on a large-scale experiment, where the treatment condition included, among other things, new performance measures for teachers—both classroom observation ratings and value-added scores—but no linked incentives. Student test scores improved in treatment schools for math but not language.

who have more opportunity for skill growth. Matching this prediction, Taylor and Tyler (2012) find larger gains for Cincinnati teachers who had lower value-added scores before their evaluation year, and larger gains for teachers who had low initial observation ratings at the start of their evaluation year. Similarly, Burgess, Rawal, and Taylor (2021), using a quantile treatment effect test, also find larger gains for lower performing teachers. Hussain (2015) studies school inspections in England. Inspectors provide substantial feedback to all schools, whether they pass or fail the inspection. Hussain finds positive effects on student test scores following a failed inspection and its consequences, but no change after a passing inspection.[37]

Empirical tests of these predictions are scarce, leaving many opportunities for contributions. The first and fourth predictions, in particular, could be tested in many (quasi-)experiments on teacher evaluation and performance incentives. For example, table 1 includes several bonus programs that were short lived, but where it might be possible to obtain data following the treated teachers' performance in years after the program.

Some evidence suggests skepticism about the role of feedback in generating performance improvements. Specifically, the role of individualized feedback from performance measures like classroom observations. First, in the Burgess, Rawal, and Taylor (2021) experiment, schools were randomly assigned to the peer evaluation program, but, within treatment schools, teachers were randomly assigned to roles: observer and observee. Student test scores improved for observer teachers just as much as they did for observee teachers, even though observers were never scored themselves. One speculation is that observers' gains came from self-assessment and studying the rubric's descriptions of best

[37] A final fifth prediction is that skill investments should be larger when the teacher expects the performance incentive program to continue for a longer period of time. A longer program means a greater return on investment. In equation 3 terms, skill investments should be increasing in T. However, I am not aware of any current empirical evidence relevant to this prediction.

practices. However, the same opportunities were available to teachers in their pre-evaluation years in Cincinnati without any apparent effect (Taylor and Tyler 2012). A second speculation is that observers' gains came from watching other teachers teach, which suggests potential gains without evaluation per se. Second, in a recent field experiment, Kraft and Christian (2021) trained school principals to provide better feedback to teachers after classroom observations, but there were no treatment effects on teachers' value-added score performance.

A third source of skepticism is that estimated performance gains are not strongly related to the number of classroom observations. Rubric scores are noisy performance measures, just like value-added scores. Assume a teacher accepts her observation ratings as accurate feedback, ignoring the noise, and invests effort to improve in her low-rated tasks. If the ratings are noisy some of that effort will have little return; noise weakens any feedback mechanism. Averaging across multiple observation ratings should reduce the noise and strengthen any feedback mechanism. However, among studies of rubric-based evaluation, estimated effects are quite similar despite variation across studies from just one to four observations per year (Taylor and Tyler 2012, Steinberg and Sartain 2015, Briole and Maurin in-press). Burgess, Rawal, and Taylor (2021) randomly varied the number of observations, with a low dose of roughly 1.5 per teacher per year and a high dose of 3, but that difference had no effect on the estimated gains from peer observation.

Fryer (2014) and Dobbie and Fryer (2013) provide some optimistic evidence about feedback. Dobbie and Fryer (2013) collect and test several predictors of charter school performance; frequent teacher feedback is among the strongest individual predictors, at least as strong as things like more instructional time and high expectations for students. In a subsequent field experiment, Fryer (2014), student achievement improved substantially when schools adopted a bundle of best practices suggested by the Dobbie and Fryer (2013) results. Among

the bundle was a ten-fold increase in the quantity of classroom observations and feedback for teachers.

5. FORMAL TRAINING

Between-teacher differences in performance strongly suggest differences in relevant job skills, which in turn motivates interest in how teachers learn those skills. In the education sector, the traditional personnel strategy for improving skills is formal training programs, but there is little evidence that formal training improves teacher performance. Nevertheless, there is evidence that teachers learn from experience, peers, and other informal training opportunities.

5.1. Training for Prospective Teachers

A substantial investment in teacher training occurs "pre-service," most often as part of a college (or university) degree program required before beginning a teaching job. That investment combines both the prospective teacher's investment of her time and tuition, and public subsidies to the college. Research on pre-service training is scarce, and notably scarce given the size of the investments. So far, no convincing evidence shows a clear return on those investments—returns in the form of improved on-the-job performance.

An intuitive place to start is to compare on-the-job performance of teachers trained in different college programs. Boyd and coauthors (2009) compare average value-added scores across preparation programs for teachers working in New York City public schools, and there are similar comparisons from several other settings (Goldhaber, Liddle, and Theobald 2013, Koedel et al. 2015, von Hipple et al. 2016, von Hipple and Bellows 2018, Bardelli, Ronfeldt, and Papay 2021). Across settings the estimated differences are small; typically the between-program standard deviation is 0.01-0.03σ (σ = student test score standard deviations). Moreover, these studies generally find no statistically significant

differences between any two programs, except for comparisons of extremes. Koedel and coauthors (2015) discuss the statistical inference challenges.

Even if there are no differences between training programs, that does not necessarily mean such programs have zero benefit for future teacher performance. Thus, we would also like to know the difference between training programs and a "no pre-service training" counterfactual.

The nearest thing to a "no pre-service training" counterfactual are teachers allowed to begin teaching without a teaching certification.[38] Still, uncertified teachers typically receive some training, for example, in an abbreviated summer program or part-time program concurrent with the first year or two of teaching. Kane, Rockoff, and Staiger (2008) compare the performance of certified and uncertified teachers working in New York City public schools. They find no difference in the mean or variance of value-added scores, and no difference in growth with experience. Gordon, Kane, and Staiger (2006) report the same lack of differences in Los Angeles schools, as do Constantine and coauthors (2009) in an experiment that assigned students to teachers in several U.S. states. The lack of difference in performance outcomes contrasts the difference in costs, both to prospective teachers and in government subsidies.

Perhaps the best-known example of alternatively certified teachers are Teach for America (TFA) teachers. TFA recruits recent college graduates, often from selective universities, who have no prior teacher training, and commits them to two years working as a teacher. Comparisons of TFA and traditionally certified teachers, some with random assignment of students, consistently find little difference in student language test scores, but higher math scores for TFA teachers' students (Glazerman, Mayer, and Decker 2006, Kane, Rockoff, and

[38] For simplicity I use the term "uncertified" but the same kind of route into teaching is also described as "temporary certification" or "alternative certification." Also, "uncertified" teachers eventually do become certified. Finally, the terms "license" and "certification" are often used interchangeably, both by schools and researchers, so "uncertified" can be read "unlicensed."

Staiger 2008, Antecol, Eren, and Ozbeklik 2013, Chiang, Clark, and McConnell 2017). These estimates are also inconsistent with some clear benefit of formal conventional pre-service training programs. However, the TFA versus non-TFA comparisons are reduced-form estimates, which combine both training and selection effects. Perhaps TFA teachers are in fact worse off because of the limited training they receive, but those losses are (more than) offset by the benefits of positive selection into TFA.

Selection complicates the comparison of training programs in general. If there are any differences between training programs, those differences may well be explained by selection. First, selection at entry into the college or program itself. Second, selection among graduates into teaching and further into specific schools.[39] The potential selection is widely acknowledged in this literature, but little progress has been made on separating selection from effects of the program training itself (see especially Goldhaber, Liddle, and Theobald 2013).

The training versus selection distinction is not irrelevant. First, separating selection from training effects is necessary for the efficient allocation of public subsidies across training programs. Second, imagine schools begin hiring based on some signal from a ranking of training programs. Prospective teachers should respond by preferring to enroll in higher-ranked programs. But then that change in selection patterns will degrade the ranking's signal over time, if the ranking partly reflects selection. The actions of any one school would not create this kind of distortion, especially if each school had their own ranking. However, rankings are increasingly available from government education agencies, among other sources, which creates scope for many schools or school systems to act on the same rankings of training programs. The effects of making such rankings widely available has not been studied.

[39] Though, this second type of selection if downstream of the program's treatment, and thus may itself be partly a program effect.

One common, but little studied, feature of college training programs is "student teaching." Trainees spend several weeks or months working alongside an experienced teacher in that mentor's classroom, eventually taking the lead in teaching the students. Thus, while a feature of formal training programs, student teaching is in many ways more like the informal peer-to-peer learning and learning-by-doing discussed in section 6. Existing evidence on student teaching is mostly descriptive. Goldhaber, Krieg, and Theobald (2017, 2020a) find that new teachers perform better when they had a higher-performing mentor, and when the students they teach now are more like the students in their student-teaching placement. Boyd and coauthors (2009) test several training program features as predictors future value-added scores; student teaching and other opportunities to practice with students are among the strongest predictors. However, a related test in Harris and Sass (2011) seems to contradict that prediction. A clear opportunity for a contribution is a well-identified estimate of the causal effect of student teaching as an input. There are large returns to experience in teaching, as discussed in section 6, and student teaching may affect those returns.

5.2. Training for Working Teachers

Schools also invest substantial money and teacher time in "professional development" (PD)—formal training programs for working teachers, often required in teacher contracts. Public schools in the United States collectively spend $18 billion annually on PD courses (Gates Foundation 2014), which is roughly 6% of total salary spending (U.S. Census Bureau 2015 table 6). The average teacher in the United States spends nearly 70 hours in those courses each year (Gates Foundation 2014), and the average for OECD countries is not far behind (Jerrim and Sims 2019), which implies an opportunity cost paid by the teacher's current students while she is away for training.

Work on this topic remains relatively rare in economics but is ubiquitous in education research more generally. Yoon and coauthors (2009) reviewed more

than 1,300 studies of PD training programs, but selection bias is endemic in that literature, and just 9 of 1,300 had (quasi-)experimental designs suited to credible claims about causal effects. More fundamentally, Hill, Beisiegel, and Jacob (2013) highlight a key disconnect: designers of teacher training increasingly agree on what features they believe make PD effective, but, when tested rigorously, those consensus features do not produce improvements in teaching. This disconnect may partly explain the stagnation of work on this topic, given that economists generally take as given the PD designs of education specialists.

Angrist and Lavy (2001) and Jacob and Lefgren (2004) are two classic examples of empirical work on in-service teacher training. Jacob and Lefgren study a modest increase in training provided to teachers in Chicago schools and find no effects on student achievement. By contrast, Angrist and Lavy find meaningful achievement gains as a result of a teacher training program in Jerusalem schools. Each of the two treatments included a bundle of features, and the two bundles differed in several ways, making it difficult to pin down the reason for Jerusalem's relative success. In general, Chicago's program was a smaller dose of training and less structed than Jerusalem's program. Those differences would make the Chicago case more typical of PD programs in practice. Jerusalem's program included counseling and feedback components; atypical features which may explain the difference in outcomes, for reasons discussed elsewhere in this chapter.

The example of these two studies demonstrates the difficulty in making progress by estimating the effects of different training programs one by one, however well identified each estimate is. The typical training program is a bundle of features, sometimes further bundled with non-training inputs. More-recent examples include studies in the Philippines, Chile, and Uganda (Abeberese, Kumler, and Linden 2014, Lombardi 2019, and Kerwin and Thornton 2021). Two of the three suggest optimism about teacher training benefits, and one found

negative effects, but all three cases are complicated by a bundle of related treatments, including classroom materials and teacher evaluation. Clearer evidence would come from (quasi-)experimental variation in training features per se.

Two examples of experimental variation in training features: Loyalka and coauthors (2019a) study a national in-service training program in China. Math teachers from randomly-selected schools attended a 15 day training at a centralized location, and afterward were given accesses to supplemental online training. The content of the training, set by the Ministry of Education, covered a wide variety of topics: math knowledge, instructional skills, student behavior management, ethics, and others. The training had no effect on student math achievement scores, compared to control schools (-0.006σ, standard error 0.034). This result is an important test of training alone, separate from other features.

Additionally, Loyalka and coauthors (2019a) test two further features added to the 15-day training. Both features respond to a common criticism that such training is quickly forgotten without follow-up. In both conditions the treatment began with the same 15-day training. In a "follow-up" condition, teachers further received personalized text messages, nearly every week, tracking their use of the online resources. For an "evaluation" condition, at the end of the 15-day training, teachers were informed that they would be evaluated in two months time, and that passing was required to receive credit for the training. The evaluation required each teacher to prepare a lesson plan, present it to evaluators and coworkers, and take questions on the plan. Neither of these additional features changed the main result (follow-up 0.005σ, standard error 0.035; evaluation 0.011σ, standard error 0.032). Moreover, Loyalka and coauthors find no effect on teachers' math knowledge or teaching practices.

Kerwin and Thornton (2021) experimentally compare two approaches to training on the same material, studying primary schools in Northern Uganda. The

contrast partly involves a common practice in schools known colloquially as "train the trainers." Selected teachers are first trained themselves, and then return to their respective schools and become the trainers of their coworker teachers. The Uganda field experiment randomly assigned schools to one of two treatment alternatives or a control. In both treatment conditions, teachers received training in a literacy instruction program emphasizing students' literacy in the local language, Leblango. The training program included several formal training sessions throughout the school year, plus observations and feedback to teachers on their implementation. In the first treatment condition, teachers were trained directly by specialized trainers from the nonprofit organization which developed the program. Leblango literacy scores increased dramatically: 0.64σ in reading and 0.45σ in writing compared to the control schools (randomization inference *p*-values = 0.005 and 0.064, respectively). The second treatment condition used a "train the trainers" approach. The effects on literacy were at best null, and perhaps negative in some skills: 0.129σ (*p*-value 0.327) in reading, -0.159 (*p*-value 0.421) in writing compared to the control. These results suggest a degradation of the training in the train-the-trainers approach. Kerwin and Thornton describe and test several potential explanations, including some that involve other smaller differences in the two conditions.

Harris and Sass (2011) take a different approach made possible by extensive administrative panel data from Florida schools. Harris and Sass use within-teacher over-time variation in the number of PD training hours, to test whether student scores improve when their teacher completes more in-service training. A key difference from the studies cited already is that the Harris and Sass estimates pool across a wide variety of specific training programs and content, returning the effect of the average PD hour. In the end, they find no consistent effect of in-service training. A teacher's PD hours are not randomly assigned, but Harris and Sass control for several factors which may be correlated

temporally with training: changes in the characteristics of students assigned; changes in grade level, subject, or school; and increases in teaching experience.

Besides the formal training organized by their employer schools, many working teachers also complete graduate-level courses offered by colleges and universities. Often teacher contracts pay a higher salary to those with a master's degree or graduate-level coursework. Empirical comparisons consistently find that teachers with a master's degree are no more effective than those without (Rivkin, Hanushek, and Kain 2005, Clotfelter, Ladd, and Vigdor 2006, 2007, Chingos and Peterson 2011). Though teachers with a master's degree may score better on measures like classroom observation ratings (Jacob et al. 2018).

One final category of in-service training is formal mentoring or coaching programs, usually for novice teachers. Rockoff (2008) studies New York City's adoption of well-established novice mentoring program. The program had little if any effect on teacher retention or performance, as measured by contributions to student achievement. A field experiment with over 1,000 teachers across a dozen U.S. states similarly found little benefit (Glazerman et al. 2010). One notably intensive but successful coaching program is My Teaching Partner (for example, Allen et al. 2011).[40] Still, even if formal mentoring or coaching programs fail to help, teachers may nevertheless be learning a lot from their peer teachers, as section 6 discusses.

Cilliers and coauthors (2020) compare group training and one-on-one coaching experimentally in South Africa. Similar to the Uganda experiment, the goal in South Africa was improving literacy instruction in the students' home language, Setswana. Schools were randomly assigned to group training, coaching, or a control condition. In the training condition, two-day group trainings occurred at the beginning of each semester, with a 7:1 teacher to trainer ratio on average.

[40] For a review of research on teacher mentoring and coaching see Kraft, Blazar, and Hogan (2018).

Teachers received detailed lesson plans, classroom materials, and training on their use, then spent time practicing. The coaching condition used the same materials and was intended to achieve the same goals. However, coaches visited each trainee teacher every month, observed her teaching, provided feedback, and demonstrated practices. Student reading achievement grew faster with coaching than it did with group training: 0.24σ (standard error 0.08) over the control, compared to 0.12σ (standard error 0.08), though the difference between coaching and group training is not statistically significant at conventional levels.

6. INFORMAL TRAINING, LEARNING ON THE JOB

Teachers do have opportunities to learn new skills outside of formal pre-service and in-service training. Those informal training opportunities include things like learning from coworkers and learning by doing. Consistent with informal mechanisms, teacher performance improves with experience; returns to experience which occur despite the lack benefits from formal "professional development" training. Though informal learning mechanisms are of interest well beyond a teacher's early career years.

6.1. Returns to Experience

Teacher performance improves with experience, especially over the first few years of a teaching career. Student achievement grows substantially faster in a second-year teacher's class compared to the same teacher's first-year class. The average teacher's value-added score increases by one-quarter to one-half of a teacher standard deviation between year one and two, and between year two and three as much as one-quarter. Then year-over-year gains weaken until there is little change after the first several years. This stylized pattern is one of the more consistent results in the literature on teachers (Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Clotfelter, Ladd, and Vigdor 2007, Kane, Rockoff, and Staiger 2008, Harris and Sass 2011, Wiswall 2013, Papay and Kraft 2015). Recent papers

show a similar pattern using classroom observation ratings instead of test-score value added (Kraft, Papay, and Chi 2020, Bell et al. 2022). Despite the consistent results, there remain several opportunities for contributions to this literature, for example, what mechanisms generate the improvements.

Estimates of returns to experience are causal claims. "Returns to experience" is shorthand for the effect of additional experience on job performance, where "experience" is the treatment or a bundle of treatments. Consider the estimation task in light of that causal inference goal.

One immediate identification threat is selection—specifically, low-performing teachers are more likely to leave teaching.[41] Even if no one improved over time, such selection would create a positive correlation between experience and performance in cross-sectional data. Murnane and Phillips (1981) discuss this and other possible selection patterns. To address the selection threat, the conventional estimation strategy uses panel data and focuses on variation in performance within-teacher over-time (Rockoff 2004). The typical regression specification is:

$$A_{i,t} = f\left(A_{i,t-1}, X_{i,t}\right) + g\left(expr_{j(i),t}\right) + \mu_{j(i)} + \pi_t + \varepsilon_{i,t}, \quad (4)$$

where student i is taught by teacher j in school year t. The outcome, $A_{i,t}$, is the student's end-of-year test score in a given subject, like math or reading. Fixed effects for teacher, $\mu_{j(i)}$, and school year, π_t, are key to the within-teacher over-time strategy for addressing selection. Controls for prior-year achievement, $A_{i,t-1}$, are key to conditional independence claims about student assignment to teachers (see for example, Kane et al. 2013, Chetty, Friedman, and Rockoff 2014a,

[41] Alternatively, a low-performing teacher might leave her job in a specific school but continue as a teacher in a different school. The main threat to identification is attrition from the estimation data correlated with performance. A change of teaching jobs may or may not result in attrition from the data, though the change in schools may affect the returns to experience among switchers.

Bacher-Hicks and Koedel in-press).[42] The vector $X_{i,t}$ represents other student or class observables.

Teacher j has years of experience $expr_{jt}$. To avoid starting with strong parametric assumptions about g, most papers specify g as a series of indicator variables. A common approach is:

$$g\left(expr_{j(i),t}\right) = \sum_{e=1}^{\bar{E}} \delta_e \mathbf{1}\{expr_{j(i),t} = e\} + \delta_{\bar{E}} \mathbf{1}\{expr_{j(i),t} > \bar{E}\}. \quad (5)$$

Importantly, this approach assumes there are no returns to experience beyond some point in the career $\bar{E}$. Typically $\bar{E}$ = 5 or 10 years of experience. Under that assumption "veteran" teachers are appropriately represented by the single $\delta_{\bar{E}}$ coefficient. This assumption turns out to be quite plausible empirically and can be tested by varying the choice of $\bar{E}$. However, Papay and Kraft (2015) and Wiswall (2013) propose alternatives to this assumption, and find evidence that returns to experience may continue deeper into the average career. Still, some restriction on g is necessary. The age-period-cohort problem rules out a fully-saturated set of indicators for $expr_{jt}$ in g along with both $\mu_{j(i)}$ and π_t.[43] Using the approach to g in 5, the school year effects, π_t, are estimated in the sample of veterans, $expr_{jt} > \bar{E}$, and assumed to hold true for early-career teachers as well.

Return to the motivating threat: selection out of teaching correlated with job performance. The teacher fixed effects approach may successfully address selection on performance levels—measured by each teacher's average performance over time—but leaves the door open to selection on a teacher's

[42] These conditional independence claims are about the ignitability of student-to-teacher assignments, not the ignorable assignment of $expr_{j(i),t}$, which leaves open the possibility of omitted variable bias where the omitted variables are other teacher characteristics or resources correlated temporally with $expr_{j(i),t}$.

[43] In practice teachers do take leaves of absence, which breaks the strict age-period-cohort relationship. Thus, a fully-saturated g can return estimates, but those estimates will be identified using leaves of absence or other unusual features of administrative data. Among other relevant considerations, the results in Dinerstein, Megalokonomou, and Yannelis (2020) suggest teaching skills decay during years away from teaching.

changes in performance. For example, δ_1 is the change in performance between a teacher's first year, $expr_{jt} = 0$, and second year, $expr_{jt} = 1$. Using the specification in 4 and 5, our estimate of δ_1 will be a weighted average of teachers who leave teaching after just the first two years, and teachers who continue to the third year and beyond. Moreover, leavers are weighted more than their proportion in the sample since the weights are increasing in the within-teacher variance of $\mathbf{1}\{expr_{jt} = 1\}$. Such a weighted average is not meaningless but may be misleading for some inferences. Assume teachers who improve more slowly with experience are more likely to leave teaching, then the estimate of δ_1 will understate the true return for stayers. Recall, for example, that δ_1 was a critical parameter in the teacher dismissal simulations in section 3. To date, there is little empirical analysis of selection on changes (Henry, Fortner, Bastian 2012, and Bell et al. 2022 are partial exceptions).

Given the causal inference goal of "returns to experience" estimates, some will recognize specification 4 as an example of a two-way fixed effects difference-in-differences estimator. Early-career teachers are the treated group, but there are multiple treatments: the first year of teaching experience, the second, the third, etc. in g in 5. Veteran teachers, with $expr_{jt} > \bar{E}$, are the comparison group. Bell and coauthors (2022) discuss returns to experience estimation from this diff-in-diff perspective, and apply recent insights on diff-in-diff methods (for example, de Chaisemartin and D'Haultfœuille 2020, Goodman-Bacon 2021). Empirically, Bell and coauthors (2022) compare typical two-way FE estimates to the alternative estimator proposed in de Chaisemartin and D'Haultfœuille (2020), and find they yield similar results.

While the average returns-to-experience pattern is a well-established result, and it suggests skill growth, the actual mechanisms are largely unknown. In many papers "experience" is not further defined as a treatment. Yet, potential

features of an experience treatment include diverse things like: learning by doing, learning from other teachers, and formal training specifically for newly hired teachers, among others.[44] In general, there has been little work on variability in the returns. The next paragraphs describe some exceptions, but also demonstrate the many opportunities for contributions.

Evidence from Ost (2014) suggests early-career performance depends not just on general teaching experience, but also on grade-level specific experience. Teachers who switch grade levels improve less. Ost's specification of g includes total experience and grade-specific experience separately, and requires the additional assumption that changes in job assignment (switching grade levels) are uncorrelated with changes in potential outcomes. Cook and Mansfield (2016) report similar results for high school teachers who switch courses. Job specific effects are consistent with learning new skills, as compared to general maturation effects. However, these effects could reflect switching costs, without any change in skills; for example, teachers shifting more effort to lesson planning at the expense of effort during class.

Kraft and Papay (2014) show heterogeneity between teachers in the returns to experience, and that those differences are partly predicted by the school where the teacher works. Further, returns to experience are larger in schools with a better "professional environment," a composite measure of survey ratings for peer collaboration, supportive leadership, order and discipline, and other things. That correlation should motivate further study of the separate components of "environment." Also worth studying is an alternative explanation: that some schools are better at selecting (or attracting) new hires based on expected growth over time. Atteberry, Loeb, and Wyckoff (2015) also study between-teacher

[44] One common example of a formal program is mentors for novice teachers. Evidence on mentoring is covered later in this chapter. Bell et al. (2022) discuss some mechanisms specific to observation rating measures of performance, for example, perhaps raters are more lenient in their ratings with experienced teachers.

variation, and find a teacher's first-year performance level predicts growth over time. That correlation suggests, among other things, scope for pre-service mechanisms like student teaching, or scope for schools to select on predicted growth at the time of hire.

Finally, if returns to experience reflect skill growth through on-the-job learning, then unemployment spells will likely result in skill depreciation. Dinerstein, Megalokonomou, and Yannelis (in-press) show the negative effect of unemployment spells on teacher performance, consistent with skill depreciation. The paper's setting is Greece where an excess supply of teachers is quasi-randomly assigned to jobs using a wait list. The paper includes several tests. For example, imagine two teachers with the same experience—the same number of years actually teaching—but the first teacher previously had an unemployment spell while the second did not; students in the first teacher's class score lower on achievement tests.

6.2. Learning from Peers

Teachers may well learn new skills from other teachers they work with day to day. Whether and how coworkers learn from each other has been of interest in economics since at least Alfred Marshall. This section reviews evidence of peer effects among teachers, and whether those effects are the result of actually learning from peers or other mechanisms. Learning from peers may well contribute to the typical returns to experience among early-career teachers. But peer learning is of interest in its own right, and could have benefits beyond the early career years.

Jackson and Bruegmann (2009) test for peer learning among teachers, using data from North Carolina elementary schools. The test uses natural changes over time in coworkers, not assignment to a specific mentoring program. The core identification strategy follows a teacher's performance over time, while she works at one school, but as her exposure to peer teachers changes. Peers are coworkers

teaching the same grade level at the same school, and peers are characterized by their prior value-added scores. Peers change if the teacher switches grade level herself or if there is turnover in her current grade level.

Current peers do affect a teacher's current performance. The average teacher's value-added score increases 0.013σ (σ = student standard deviations) when one of her three grade-level peers is replaced by a new peer whose value-added score is one standard deviation higher than the replaced peer. A 0.013σ gain moves the teacher up one-tenth of a standard deviation or more in the teacher performance distribution. Though that result alone is not evidence of learning from peers. Teacher peer effects could arise through various mechanisms, as Jackson and Bruegmann (2009) discuss.

Learning directly from peers is one potential mechanism for the results in Jackson and Bruegmann (2009). In that case, a teacher can presumably learn more from a more skilled peer, consistent with the 0.013σ estimate. Results from Papay and coauthors' (2020) experiment, described next, further emphasize the value of peers with complementary skills. However, positive peer effects may reflect learning caused by better peers, but not learning directly from those peers. For example, a teacher may be motivated by a new high-performing peer, motivated to invest more effort in her own skill development. Other mechanisms do not involve learning new skills. First, positive peer effects, like the 0.013σ estimate, may have an explanation in joint production activities. If teachers trade tasks, like lesson planning, then the gains from trade should be greater with more skilled peers. Second, teachers may compete for resources from the school. If they compete as teams, a higher performing team member might increase resources for the entire team, a positive peer effect. If teachers compete individually, then the peer effects could be negative. Third, a new high-performing peer might motivate a teacher to give more effort at work, either through encouragement or social pressure to compare well.

One key prediction that differentiates peer learning from other mechanisms is persistence of effects into the future, paralleling the discussion in section 4 about learning from evaluation. Matching that prediction, Jackson and Bruegmann (2009) find that past peers also affect a teacher's current performance. A teacher's prior peers, measured by their value-added scores, explain about one-fifth of the teacher's own current value added. A second result consistent with learning is that peer effects are somewhat larger for novice teachers who have both the most to learn and the longest to capture returns on new skills. Though, Jackson and Bruegmann's estimates show some peer effects for veteran teachers as well.

Jackson and Bruegmann's (2009) estimates potentially capture a wide variety of different ways peers might interact or collaborate. The peer effects they find contradict the notion of the "egg crate" school, where each teacher works independently of her peers (Lortie 1975). Yet, the average school may still fall short of optimal teacher collaboration; schools may still be inefficiently egg crate. The field experiments discussed next show benefits of additional peer interaction among teachers, on top of the baseline interactions in Jackson and Bruegmann (2009). However, these experiments may find large marginal benefits because baseline peer interactions were infrequent in their samples.

Papay and coauthors (2020) show evidence of teachers learning from peers, during a field experiment in Tennessee. In treatment schools, teachers were paired up and asked to work together on improving their teaching. Each pair included a "target" teacher who had scored quite low in one or more of 19 skills during prior classroom observations, and a "partner" teacher who had scored high in (many of) the target's deficient skills. The skills included things like asking questions and responding to misbehavior, as measured by classroom observation rubrics.

The experiment's results are consistent with target teachers learning from their more-skilled partner. The main result is that student achievement improved 0.12σ in target teachers' classrooms.[45] Though that 0.12σ performance gain could reflect mechanisms other than peer learning, like the joint production or effort motivation hypotheses. The case for peer learning involves additional results. First, the target teacher's improvement was larger when more of her weak skill areas were matched by her partner's strong skills. By contrast, having a partner with higher value-added scores did not improve the effect on target teachers. In subsequent classroom observation ratings, target teachers scored higher in skills where her partner teacher was strong but not in other skill areas. Second, any joint production effect would likely be larger when target and partner are teaching the same grade level or subject, but the treatment effects are no larger or smaller when the pair shares grade or subject. Third, target teachers' gains persist in the following school year, after the partnerships had ended. Still, even if the mechanism is not entirely peer learning, the experiment's results suggest schools have not fully exploited the gains from teacher peer collaboration.

A second experiment in increasing peer interaction is the Burgess, Rawal, and Taylor (2021) study detailed in section 4. Those interactions were structured around peer observations and ratings. Recall both observers and observees benefited similarly suggesting possible peer effect mechanisms.

Murphy, Weinhardt, and Wyness (2021) report on a third experiment in teacher peer collaboration. In each treatment school, a group of three teachers worked together. The group chose what they would work on improving, observed each other teaching, and met to discuss feedback and strategies. The were no treatment effects on average. However, student test scores improved in larger

[45] Pairs were matched by an algorithm for all participating schools, before random assignment of schools to treatment and control, but only revealed by the researchers in treatment schools. Thus, the researchers observe the latent pairs in control schools.

schools, with more than three potential participants, but declined in smaller schools. The authors speculate that larger schools had scope to select participants based on potential gains. Additionally, the treatment program in Murphy, Weinhardt, and Wyness (2021) was less structured than the programs studied by Burgess, Rawal, and Taylor (2021) and Papay and coauthors (2020) which were organized around observation rubrics and specific skills described in the rubrics.

One important (potential) cost of peer mentoring programs is the opportunity cost of the mentor's time and effort. For example, a mentor who leaves her own classroom to spend time observing a peer's classroom, and thus the mentor's own students are worse off. In Papay and coauthors' (2020) experiment, partner teachers' students were not worse off, in the end, and perhaps slightly better off. The treatment effect for partner teachers was 0.03σ but far from statistically significant. Similarly, teachers randomly assigned to the observer role were no worse off in the Burgess, Rawal, and Taylor (2021) experiment. Goldhaber, Krieg, and Theobald (2020b) study the experienced teachers who host a student teacher. Achievement test scores do not decline in the hosting year, and, in the years after hosting, the mentor teacher's value-added score improves somewhat.[46]

7. CONCLUSION

Schools as employers must decide how to measure teacher job performance, whether to link compensation to those measures, and when to dismiss teachers based on performance, among other personnel matters. Schools'

[46] The evaluation program studied in Taylor and Tyler (2012) is described by some as a "peer assistance and review" program. The "peers" are not coworkers, however, but instead experienced high-performing teachers who leave classroom teaching jobs to work full-time as evaluators and mentors. Because these "peers" stop teaching, the opportunity cost is clear; the district must hire novices to teach the students whom these evaluators would have otherwise taught. Taylor and Tyler (2012) compare the likely student achievement losses of that opportunity cost to the gains from the evaluation program.

best decisions will anticipate how teachers' own choices are likely to respond to evaluation and incentives. This chapter summarizes empirical evidence and analysis relevant to understanding these employer-employee interactions and, importantly, the consequences for students.

Teacher effort responds to performance incentives. The literature now includes many (quasi-)experimental examples of increases in teacher performance measures caused by bonuses and other performance incentives. But those increases are not necessarily evidence of improvements in student welfare. For example, student improvements on tests that determine teacher rewards do not always generalize to other low-stakes non-incentivized tests of the same subject material. And there are other empirical examples of unintended distortions in effort, "teaching to the test," and manipulation of performance measures. However, the literature also includes examples consistent with improvements in student welfare. For example, teacher incentives for math test scores which have positive spillovers on science achievement, or positive effects on longer-run outcomes like college attendance and labor market earnings. Future work on this topic should prioritize an understanding of mechanisms—how teacher effort responds to incentives—which might reconcile the evidence of distortions and evidence of real improvements for students.

Loyalka and coauthors' (2019b) work in Shaanxi and Gansu is an example of thoughtful experimental variation and data collection designed to contribute understanding of mechanisms. Specifically in that example, how a teacher allocates her effort across her many students, and how that allocation changes in response to performance incentives. Studies described in section 2.1.3 suggest the teacher's allocation of effort across students is an important mechanism worthy of future analysis. Moreover, other potential mechanisms could also be studied experimentally. Motivated by the quasi-experimental examples in section 2.2, a thoughtfully designed study could experimentally vary the predictability of test

questions and topics, and thus experimentally vary the scope for test prep and teaching to the test.

If changing teacher effort is the most common rationale for performance incentives, then teacher self-selection is a close second. Teachers may self-select into or out of a school, acting on private information about their own ability or likely response to incentives. Evaluation programs also create (new) performance information which schools themselves can act on to select teachers. Yet, there is noticeable gap between the large potential benefits of teacher (self-)selection and the small empirical literature. For selection by schools, a limiting factor is empirical opportunities for quasi-experimental analysis. Few schools actually dismiss teachers based on performance measures. Rules which appear to threaten dismissal are more and more common, but the actual probability of dismissal (or loosing work as a teacher) is quite low. By contrast, work on teacher self-selection has produced more contributions in recent years. The creative experimental designs used by Leaver and coauthors (2021) and Brown and Andrbai (2021) are particularly notable. Taken together, these and other recent contributions suggest that teachers' private information likely varies between teachers contributing to heterogeneous effects of self-selection. In particular, it appears that prospective teachers may have little private information—specifically about their potential scores on the incentivized performance measures—before they begin working, leaving little scope for self-selection effects at entry into teaching. Further evidence on variation in private information, especially among prospective teachers, would help sharpen the selection rationale for teacher evaluation.

A third rational for evaluation is teacher skill development. More specifically, that evaluation measures and incentives can improve job performance through causal effects on teachers' job skills. This third rationale is quite common among education scholars outside of economics, where the

mechanisms are broadly described as "feedback for improvement," but there is economics here. Performance measurement can reduce the costs of teachers' skill investments, and performance incentives can increase the expected return on those investments. In a multiperiod agency-theory framework, if the teacher anticipates repeated evaluation and performance incentives over time, then the potential future rewards create an incentive for the teacher to invest in improving her relevant skills. The relationship between performance incentives and skill investments is a clear opportunity for future contributions, both theoretical and empirical.

Some evidence of skill development effects comes from Taylor and Tyler (2012) and Briole and Maurin (in-press), both of which track teachers over time for several years after their performance is no longer measured nor incentivized by the schools they work for. In both cases evaluation using classroom observations created lasting improvements in teachers' contributions to student achievement scores. This matches one key prediction that, if the mechanism is skill investment, then evaluation's effects can persist after the evaluation ends because skills persist. These are just two examples but testing this persistence prediction should be feasible in many other settings, including by following-up on prior (quasi-)experiments. A future study might experimentally vary teachers' expectations about how long an incentive program is expected to continue into the future.

This chapter has also raised several other topics which are promising areas for future contributions. I will highlight two. First, the potential complementarities between teacher performance incentives and other school inputs, as Mbiti and coauthors (2019) found in the Tanzania experiment. Second, the contrast between relatively-objective performance measures, like those based on student test scores, and more-subjective ratings by supervisors. Andrabi and Brown's (2022) experimental comparison of objective and subjective incentives

in Pakistan suggest, as some have argued, that subjective incentives may reduce the scope for distortions in effort.

Most schools' personnel strategies also include making investments in teachers' skills. The end of this chapter summarizes empirical analysis of formal training programs, as well as informal ways in which teachers learn new skills at work. A teacher's job performance can improve, even absent evaluation or new performance incentives. For example, average teacher performance improves quickly over the first five years of working as a teacher. That pattern of returns to experience is one of the most consistent results in the economics of education literature, and it strongly suggests teachers are learning new skills. What generates those returns to experience is mostly unknown. Possible mechanisms include learning by doing, learning from coworkers, and other informal methods, as well as formal training or mentoring programs. Sorting out these mechanisms is a clear opportunity for future contributions. As of this review, there remains little consistent evidence on the effects of formal teacher training, and often investments in formal training yield little return for schools. Informal mechanisms, like learning from other teachers, have received much less (quasi-)experimental analysis, but the small literature includes some encouraging results.

The employment relationship between teachers and schools is central to the economics of education, and often consequential for students' success. The topics discussed in this chapter—broadly, teacher evaluation and training—have received relatively more attention from economists than others—like, recruitment and hiring or job design. Yet, these various personnel decisions are often interrelated. For example, adjusting job design may reduce incentive distortions but at other costs. Considering the range of employer-employee interactions in schools will improve research on teachers and teaching, and contribute to understanding employer-employee relationships in other occupations and sectors.

REFERENCES

Abeberese, A. B., Kumler, T. J., & Linden, L. L. (2014). Improving reading skills by encouraging children to read in school: A randomized evaluation of the Sa Aklat Sisikat reading program in the Philippines. *Journal of Human Resources*, *49*(3), 611-633.

Adnot, M. (2016). Teacher evaluation, instructional practice and student achievement: Evidence from the District of Columbia Public Schools and the measures of effective teaching project. Doctoral dissertation, University of Virginia.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54-76.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034-1037.

Allen, R., & Burgess, S. (2012). How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. CMPO Working Paper No. 12-287.

Andrabi. C. & Brown, T. (2022). Subjective versus objective incentives and teacher productivity. Working paper.

Angrist, J. D., & Lavy, V. (2001). Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools. *Journal of Labor Economics*, *19*(2), 343-369.

Antecol, H., Eren, O., & Ozbeklik, S. (2013). The effect of Teach for America on the distribution of student achievement in primary school: Evidence from a randomized experiment. *Economics of Education Review*, *37*, 113-125.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, *131*(3), 1415-1453.

Araujo, M. D., Heineck, G., & Cruz-Aguayo, Y. (2020). Does test-based teacher recruitment work in the developing world? Experimental evidence from Ecuador. IZA Discussion Paper No. 13830.

Atkinson, A., Burgess, S., Croxson, B., Gregg, P., Propper, C., Slater, H., & Wilson, D. (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics*, *16*(3), 251-261.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4).

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., & Mozenter, Z. (2022). Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *The Economic Journal,* in-press.

Aucejo, E., Romano, T., & Taylor, E. S. (2022). Does evaluation change teacher effort and performance? Quasi-experimental evidence from a policy of retesting students. *Review of Economics and Statistics, 104*(2), in-press.

Bacher-Hicks, A., & Koedel, C. (in-press). Estimation and interpretation of teacher value-added in research applications. In Hanushek, E. A., Machin, S., & Woessmann, L. (Eds.), *Handbook of the Economics of Education* (in-press). Elsevier.

Bacher-Hicks, A., Kane, T. J., & Staiger, D. O. (2014). Validating teacher effect estimates using changes in teacher assignments in Los Angeles. NBER Working Paper No. 20657.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2019). An experimental evaluation of three teacher quality measures: Value-added, classroom observations, and student surveys. *Economics of Education Review*, *73*, 101919.

Balart, P., Oosterveen, M., & Webbink, D. (2018). Test scores, noncognitive skills and economic growth. *Economics of Education Review, 63*, 134-153.

Banerjee, A., & Duflo, E. (2006). Addressing absence. *Journal of Economic perspectives*, *20*(1), 117-132.

Bardelli, E., Ronfeldt, M., & Papay, J. (2021). Teacher preparation programs and graduates' growth in instructional effectiveness. EdWorkingPapers No. 21-450.

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review*, *102*(5), 1805-31.

Barr, A. S. (1928). An evaluation of items to observe in classroom supervision. *The Journal of Educational Research*, *18*(1), 53-65.

Barrera-Osorio, F., & Raju, D. (2017). Teacher performance pay: Experimental evidence from Pakistan. *Journal of Public Economics, 148*, 75-91.

Bates, M. (2020). Public and private employer learning: Evidence from the adoption of teacher value added. *Journal of Labor Economics*, *38*(2), 375-420.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *Journal of political economy*, *70*(5, Part 2), 9-49.

Behrman, J. R., Parker, S. W., Todd, P. E., & Wolpin, K. I. (2015). Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy*, *123*(2), 325-364.

Bell, C., James, J., Taylor, E. S., & Wyckoff, J. (2022). Measuring returns to experience using observations of teaching. EdWorkingPaper No. 22-526.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *Journal of Political Economy*, *75*(4, Part 1), 352-365.

Bergman, P., & Hill, M. J. (2018). The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers. *Economics of Education Review*, *66*, 104-113.

Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51.

Biasi, B. (2021). The Labor Market for Teachers Under Different Pay Schemes. *American Economic Journal: Economic Policy, 13*(2), 65-102.

Biasi, B., & Sarsons, H. (2022). Flexible wages, bargaining, and the gender gap. *Quarterly Journal of Economics, 137*(1), 215-266.

Biasi, B., Fu, C., & Stromme, J. (2021). Equilibrium in the market for public school teachers: District wage strategies and teacher comparative advantage. NBER Working Paper No. 28530.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416-40.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. *Education Finance and Policy, 6*(3): 439-54.

Brehm, M., Imberman, S. A., & Lovenheim, M. F. (2017). Achievement effects of individual performance incentives in a teacher merit pay tournament. *Labour Economics*, *44*, 133-150.

Briole, S., & Maurin, E. (in-press). There's always room for improvement: The persistent benefits of a large-scale teacher evaluation system? *Journal of Human Resources.*

Brown, C., & Andrabi, T. (2021). Inducing positive sorting through performance pay: Experimental evidence from Pakistani schools. Working Paper.

Brunner, E., Cowen, J. M., Strunk, K. O., & Drake, S. (2019). Teacher labor market responses to statewide reform: Evidence from Michigan. *Educational Evaluation and Policy Analysis*, *41*(4), 403-425.

Burgess, S., Greaves, E., & Murphy, R. (2022). Deregulating teacher labor markets. *Economics of Education Review, 88*, 102253.

Burgess, S., Propper, C., Slater, H., & Wilson, D. (2005). Who wins and who loses from school accountability? The distribution of educational gain in English secondary schools. CMPO Working Paper No. 5-128.

Burgess, S., Rawal, S., & Taylor, E. S. (2021). Teacher peer observation and student test scores: Evidence from a field experiment in English secondary schools. *Journal of Labor Economics, 39*(4), 1155-1186.

Burgess, S., Rawal, S., & Taylor, E. S. (2022). Teachers' use of class time and student achievement. NBER Working Paper No. 30686.

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences*, *110*(43), 17176-17182.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *Journal of Economic Perspectives, 20*(1), 91-116.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633-2679.

Chi, O. L. (in-press). A classroom observer like me: The effects of race-congruence and gender-congruence between teachers and raters on observation scores. *Education Finance and Policy*.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics, 93*(9-10), 1045-1057.

Chiang, H. S., Clark, M. A., & McConnell, S. (2017). Supplying disadvantaged schools with effective teachers: Experimental evidence on secondary math teachers from teach for America. *Journal of Policy Analysis and Management*, *36*(1), 97-125.

Chingos, M. M., & Peterson, P. E. (2011). It's easier to pick a good teacher than to train one: Familiar and new results on the correlates of teacher effectiveness. *Economics of Education Review*, *30*(3), 449-465.

Chingos, M. M., & West, M. R. (2012). Do more effective teachers earn more outside the classroom? *Education Finance and Policy*, *7*(1), 8-43.

Cilliers, J., Fleisch, B., Prinsloo, C., & Taylor, S. (2020). How to improve teaching practice? An experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources, 5*5(3), 926-962.

Cilliers, J., Kasirye, I., Leaver, C., Serneels, P., & Zeitlin, A. (2018). Pay for locally monitored performance? A welfare analysis for teacher attendance in Ugandan primary schools. *Journal of Public Economics*, *167*, 69-90.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources*, *41*(4), 778-820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682.

Cohodes, S. R. (2016). Teaching to the student: Charter school effectiveness in spite of perverse incentives. *Education Finance and Policy, 11*(1), 1-42.

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification. Final report.* NCEE 2009-4043. U.S. Department of Education.

Cook, J. B., & Mansfield, R. K. (2016). Task-specific experience and task-specific talent: Decomposing the productivity of high school teachers. *Journal of Public Economics*, *140*, 51-72.

Cullen, J. B., Koedel, C., & Parsons, E. (2021). The compositional effect of rigorous teacher evaluation on workforce quality. *Education Finance and Policy*, *16*(1), 7-41.

Cullen, J.B., & Reback, R. (2006). Tinkering toward accolades: School gaming under a performance accountability system. In Gronberg, T. J., & Jansen, D. J. (Eds.) *Improving School Accountability*. Emerald Group Publishing Limited.

Danielson, C. (1996). *Enhancing professional practice: a framework for teaching.* Alexandria, VA:ASCD.

Darling-Hammond, L. (2015). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Teachers College Press.

De Chaisemartin, C., & D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, *110*(9), 2964-96.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*(3), 418-446.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267-297.

Dee, T. S., James, J., & Wyckoff, J. (2021). Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools. *Education Finance and Policy, 16*(2), 313–346.

Deming, D. J., Cohodes. S., Jennings, J. & Jencks, C. (2016). School accountability, postsecondary attainment, and earnings. *Review of Economics and Statistics, 98* (5): 848-862.

Dinerstein, M., Megalokonomou, R., & Yannelis, C. (in-press). Human capital depreciation and returns to experience. *American Economic Review.*

Dinerstein, M., & Opper, I. M. (2022). Screening with Multitasking. NBER Working Paper No. 30310.

Dixit, A. (2002). Incentives and organizations in the public sector: An interpretative review. *Journal of Human Resources, 37*(4), 696-727.

Dobbie, W., & Fryer Jr, R. G. (2013). Getting beneath the veil of effective schools: Evidence from New York City. *American Economic Journal: Applied Economics*, *5*(4), 28-60.

Dohmen, T., & Falk, A. (2010). You get what you pay for: Incentives and selection in the education system. *The Economic Journal*, *120*(546), F256-F271.

Donaldson, M. L., & Papay, J. P. (2015). Teacher evaluation for accountability and development. In Ladd, H. F., & Goertz, M. E. (Eds.) *Handbook of Research in Education Finance and Policy*. Routledge.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review*, *101*(5), 1739-74.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *American Economic Review*, *102*(4), 1241-78.

Estrada, R. (2019). Rules versus discretion in public service: Teacher hiring in Mexico. *Journal of Labor Economics*, *37*(2), 545-579.

Figlio, D. N. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4-5), 837-851.

Figlio, D. N., and Getzler, L.S. (2006). Accountability, ability and disability: Gaming the system? In Gronberg, T. J., & Jansen, D. J. (Eds.) *Improving School Accountability*. Emerald Group Publishing Limited.

Figlio, D. N., & Winicki, J. (2005). Food for thought: the effects of school accountability plans on school nutrition. *Journal of Public Economics, 89*(2-3), 381-394.

Fryer, R. G. (2014). Injecting charter school best practices into traditional public schools: Evidence from field experiments. *Quarterly Journal of Economics*, *129*(3), 1355-1407.

Fryer, R. (2018). The "pupil" factory: Specialization and the production of human capital in schools. *American Economic Review, 108*(3), 616-56.

Fryer, R. G. (2013). Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics*, *31*(2), 373-407.

Fryer, R. G., Levitt, S. D., List, J., & Sadoff, S. (2022). Enhancing the efficacy of teacher incentives through framing: A field experiment. *American Economic Journal: Economic Policy, 14*(4), 269–299.

Garet, M. S., Wayne, A. J., Brown, S., Rickles, J., Song, M., & Manzeske, D. (2017). *The impact of providing performance feedback to teachers and principals.* NCEE 2018-4001. U.S. Department of Education.

Gates Foundation. (2014). *Teachers know best: Teachers views on professional development.* Seattle, WA: Bill & Melinda Gates Foundation.

Gibbons, R., & Roberts, J. (2013). Economic theories of incentives in organizations. In Gibbons, R., & Roberts, J. (Eds.), *Handbook of Organizational Economics*, Princeton, NJ: Princeton University Press.

Gill, B., Shoji, M., Coen, T., & Place, K. (2016). *The content, predictive power, and potential bias in five widely used teacher observation instruments.* REL 2017–191. U.S. Department of Education.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. (2010). *Impacts of comprehensive teacher induction: Final results from a randomized controlled study.* NCEE 2010-4027. U.S. Department of Education

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75-96.

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3), 205-27.

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291-308.

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

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

Goldhaber, D., Krieg, J. M., & Theobald, R. (2017). Does the match matter? Exploring whether student teaching experiences affect teacher effectiveness. *American Educational Research Journal*, *54*(2), 325-359.

Goldhaber, D., Krieg, J., & Theobald, R. (2020a). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics, 63*, 101792.

Goldhaber, D., Krieg, J., & Theobald, R. (2020b). Exploring the impact of student teaching apprenticeships on student achievement and mentor teachers. *Journal of Research on Educational Effectiveness*, *13*(2), 213-234.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics, 225*(2), 254-277.

Goodman, S. F., & Turner, L. J. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics*, *31*(2), 409-420.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job.* The Hamilton Project Policy Brief No. 2006-01. Washington, DC: Brookings Institution.

Grissom, J. A., & Bartanen, B. (2022). Potential race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management, 41*(1), 131-161.

Hanushek, E. A. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data." *American Economic Review, 61*(2), 280-288

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, *30*(3), 466-479.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*(2), 297-327.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-71.

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

Henry, G. T., Fortner, C. K., & Bastian, K. C. (2012). The effects of experience and attrition for novice high-school science and mathematics teachers. *Science*, *335*(6072), 1118-1121.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476-487.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Holmstrom, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 324-340.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, *7*(Special), 24-52.

Hoxby, C. M., & Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review*, *94*(2), 236-240.

Hudson, S. (2010). The effects of performance-based teacher pay on student achievement. SIEPR Discussion Paper No. 09-023.

Hussain, I. (2015). Subjective performance evaluation in the public sector evidence from school inspections. *Journal of Human Resources*, *50*(1), 189-221.

Imberman, S. A., & Lovenheim, M. F. (2015). Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics*, *97*(2), 364-386.

Jackson, C. K. (2010). A little now for a lot later a look at a Texas Advanced Placement incentive program. *Journal of Human Resources, 45*(3), 591-639.

Jackson, C. K. (2014). Do college-preparatory programs improve long-term outcomes? *Economic Inquiry, 52*(1), 72-99.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*(5), 2072-2107.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85-108.

Jackson, C. K., Rockoff, J.E., & Staiger, D.O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6*(1), 801-825.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5-6), 761-796.

Jacob, B. A. (2011). Do principals fire the worst teachers? *Educational Evaluation and Policy Analysis*, *33*(4), 403-434.

Jacob, B. A. (2013). The effect of employment protection on teacher effort. *Journal of Labor Economics*, *31*(4), 727-761.

Jacob, B. A., & Lefgren, L. (2004). 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., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-36.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843-877.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources*, *45*(4), 915-943.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics*, *166*, 81-97.

Jerrim, J., & Sims, S. (2019). *The teaching and learning international survey (TALIS) 2018: Research report.* London: UCL, Institute for Education.

Johnston, A. C. (2021). Preferences, selection, and the structure of teacher pay. IZA Discussion Paper No. 14831.

Kane, T. J., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle, WA: Bill & Melinda Gates Foundation.

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

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic perspectives*, *16*(4), 91-114.

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587-613.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591-598.

Kerwin, J. T., & Thornton, R. L. (2021). Making the grade: The sensitivity of education program effectiveness to input choices and outcome measures. *Review of Economics and Statistics, 103*(2), 251-264.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* Santa Monica, CA: RAND.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher preparation programs and teacher quality: Are there real differences across programs? *Education Finance and Policy*, *10*(4), 508-534.

Koretz, D. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources, 37*(4), 752-777.

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.

Koretz, D. (2017). *The testing charade: Pretending to make schools better*. Chicago, IL: University of Chicago Press.

Koretz, D., & Barron, S. I. (1998). *The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND.

Kraft, M. A. (2015). Teacher layoffs, teacher quality, and student achievement: Evidence from a discretionary layoff policy. *Education Finance and Policy*, *10*(4), 467-507.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1-36.

Kraft, M. A., & Blazar, D. (2017). Individualized coaching to improve teacher practice across grades and subjects: New experimental evidence. *Educational Policy*, *31*(7), 1033-1068.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234-249.

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476-500.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588.

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315-347.

Krueger, A. B., & Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, *111*(468), 1-28.

Ladd, H. F. (1999). The Dallas school accountability and incentive program: An evaluation of its impacts on student outcomes. *Economics of Education Review*, *18*(1), 1-16.

Lavy, V. (2002). Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy*, *110*(6), 1286-1317.

Lavy, V. (2009). Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review*, *99*(5), 1979-2011.

Lavy, V. (2020). Teachers' pay for performance in the long-run: The dynamic pattern of treatment effects on students' educational and labour market outcomes in adulthood. *Review of Economic Studies*, *87*(5), 2322-2355.

Lazear, E. P. (2006). Speeding, terrorism, and teaching to the test. *Quarterly Journal of Economics, 121*(3), 1029-1061.

Lazear, E., & Oyer, P. (2013). Personnel economics. In Gibbons, R., & Roberts, J. (Eds.), *Handbook of Organizational Economics*, Princeton, N.J.: Princeton University Press.

Leaver, C., Ozier, O., Serneels, P., & Zeitlin, A. (2021). Recruitment, effort, and retention effects of performance contracts for civil servants: Experimental evidence from Rwandan primary schools. *American Economic Review, 111*(7), 2213-2246.

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

Liu, J., & Loeb, S. (2021). Engaging teachers measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*, *56*(2), 343-379.

Loeb, S., Miller, L. C., & Wyckoff, J. (2015). Performance screens for school improvement: The case of teacher tenure reform in New York City. *Educational Researcher*, *44*(4), 199-212.

Lombardi, M. (2019). Is the remedy worse than the disease? The impact of teacher remediation on teacher and student performance in Chile. *Economics of Education Review*, *73*, 101928.

Lortie, D. C. (1975). *Schoolteacher: A sociological study*. Chicago, IL: University of Chicago Press.

Loyalka, P., Popova, A., Li, G., & Shi, Z. (2019a). Does teacher training actually work? Evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics, 11*(3), 128-54.

Loyalka, P., Sylvia, S., Liu, C., Chu, J., & Shi, Y. (2019b). Pay by design: Teacher performance pay design and the distribution of student achievement. *Journal of Labor Economics*, *37*(3), 621-662.

Luginbuhl, R., Webbink, D., & De Wolf, I. (2009). Do inspections improve primary school performance? *Educational Evaluation and Policy Analysis*, *31*(3), 221-237.

Macartney, H. (2016). The dynamic effects of educational accountability. *Journal of Labor Economics, 34*(1), 1-28.

Macartney, H., McMillan, R., & Petronijevic, U. (2018). Teacher value-added and economic agency. NBER Working Paper No. 24747.

Macartney, H., McMillan, R., & Petronijevic, U. (2021). A quantitative framework for analyzing the distributional effects of incentive schemes. NBER Working Paper No. 28816.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., & Epstein, S. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses: Final evaluation report.* Santa Monica, CA: RAND.

Martins, P. S. (2009). Individual teacher incentives, student achievement and grade inflation. IZA Discussion Paper No. 4051.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2019). Inputs, incentives, and complementarities in education: Experimental evidence from Tanzania. *Quarterly Journal of Economics*, *134*(3), 1627-1673.

Mulhern, C. & Opper, I. (2021). Measuring and summarizing the multiple dimensions of teacher effectiveness. EdWorkingPaper No. 21-451.

Muralidharan, K. (2012). Long-term effects of teacher performance pay: Experimental evidence from India. Working Paper.

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *The Economic Journal, 120*(546), F187-F203.

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *Journal of Political Economy*, *119*(1), 39-77.

Murnane R. J. (1975). *The impact of school resources on the learning of inner city children.* Cambridge, MA: Ballinger.

Murnane, R. J., & Phillips, B. R. (1981). Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance. *Economics of Education Review*, *1*(4), 453-465.

Murphy, R., Weinhardt, F., & Wyness, G. (2021). Who teaches the teachers? A RCT of peer-to-peer observation and feedback in 181 schools. *Economics of Education Review*, *82*, 102091.

Nagler, M., Piopiunik, M., & West, M. R. (2020). Weak markets, strong teachers: Recession at career start and teacher effectiveness. *Journal of Labor Economics*, *38*(2), 453-500.

NCES. (2016). *The condition of education 2016.* U.S. Department of Education.

Neal, D. (2011). The design of performance pay in education. In Hanushek, E. A., Machin, S., & Woessmann, L. (Eds.), *Handbook of the Economics of Education*. Elsevier.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics, 92*(2), 263-283.

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

New York Times. (2013b). Ed-schools chief in Atlanta is indicted in testing scandal. March 29, 2013.

Ng, K. (2021). The effects of teacher tenure on productivity and selection. Working paper.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics, 6*(2), 127–51.

Oyer, P., & Schaefer, S. (2011). Personnel economics: Hiring and incentives. In Ashenfelter, O., & Card, D. (Eds.), *Handbook of Labor Economics.* Elsevier.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105-119.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359-88.

Petek, N., & Pope, N. (2021). The multidimensional impact of teachers on students. Working paper.

Phipps, A. R., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy, 16*(2), 283-312.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS): Dimensions overview*. Baltimore, MD: Paul H Brookes Publishing.

Pope, N. (2019). The effect of teacher ratings on teacher performance. *Journal of Public Economics*, *172*, 84-110.

Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, *37*(1), 7-63.

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2021). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics*, *46*(2), 138-167.

Reback, R. (2008). Teaching to the rating: School accountability and the distribution of student achievement. *Journal of Public Economics, 92*(5-6), 1394-1415.

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

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

Rockoff, J. E. (2008). Does mentoring reduce turnover and improve skills of new employees? Evidence from teachers in New York City. NBER Working Paper No. 13868.

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review*, *100*(2), 261-66.

Rockoff, J. E., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *American Economic Review*, *102*(7), 3184-3213.

Ross, E. & Walsh, K. (2019). *State of the States 2019: Teacher and Principal Evaluation Policy.* Washington, DC: National Council on Teacher Quality.

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

Rothstein, J. (2015). Teacher quality policy when supply matters. *American Economic Review*, *105*(1), 100-130.

Rothstein, J. (2017). Measuring the impacts of teachers: Comment. *American Economic Review*, *107*(6), 1656-84.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. NBER Working Paper No. 13681.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2013). Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure. *American Economic Journal: Economic Policy, 5*(2), 251-81.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for

educational evaluation and research. *Journal of Personnel Evaluation in Education*, *12*(3), 247-256.

Santibanez, L., Martinez, J. F., Datar, A., McEwan, P. J., Setodji, C. M., & Basurto-Davila, R. (2007). *Breaking ground: Analysis of the assessment system and impact of Mexico's teacher incentive program.* Santa Monica, CA: RAND Corporation.

Sartain, L., & Steinberg, M. P. (2016). Teachers' labor market responses to performance evaluation reform: Experimental evidence from Chicago public schools. *Journal of Human Resources*, *51*(3), 615-655.

Schwartz, A. L., (2021). Accuracy versus incentives: A trade-off for performance measurement. *American Journal of Health Economics, 7*(3), 333-360.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438-1457.

Sojourner, A. J., Mykerezi, E., & West, K. L. (2014). Teacher pay reform and productivity panel data evidence from adoptions of Q-Comp in Minnesota. *Journal of Human Resources*, *49*(4), 945-981.

Speroni, C., Wellington, A., Burkander, P., Chiang, H., Herrmann, M., & Hallgren, K. (2020). Do educator performance incentives help students? Evidence from the teacher incentive fund national evaluation. *Journal of Labor Economics, 38*(3), 843-872.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review, 27*(5), 556-563.

Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V. N., Pepper, M., Lockwood, J. R., & Stecher, B. M. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching.* Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives*, *24*(3), 97-118.

Steinberg, M. P., & Donaldson, M. L. (2016). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy, 11*(3), 340-359.

Steinberg, M. P., & Sartain, L. (2015). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching project. *Education Finance and Policy*, *10*(4), 535-572.

Summers, A. A., & Wolfe, B. L. (1977). Do schools make a difference? *American Economic Review*, *67*(4), 639-652.

Taylor, E. S. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162-181.

Taylor, E. S. (2022). Employee evaluation and skill investments: Evidence from public school teachers. NBER Working Paper No. 30687.

Taylor, E. S. & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Vigdor, J. L. (2009). Teacher salary bonuses in North Carolina. In Springer, M. (Ed.), *Performance Incentives: Their Growing Impact on American K-12 Education.* Washington, D.C.: Brookings Institution Press.

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, *64*, 298-312.

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31-45.

Warren C, S., & Balfour S, T. (1934). An evaluation of differences in teaching ability. *The Journal of Educational Research*, *28*(1), 10-15.

Washington Post. (2022). The Atlanta schools cheating scandal isn't over. February 1, 2022.

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

Winters, M. A., & Cowen, J. M. (2013). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher*, *42*(6), 330-337.

Winters, M. A., Trivitt, J. R., & Greene, J. P. (2010). The impact of high-stakes testing on student proficiency in low-stakes subjects: Evidence from Florida's elementary science exam. *Economics of Education Review, 29*(1), 138-146.

Winters, M., Greene, J., Ritter, G., & Marsh, R. (2008). The effect of performance pay in Little Rock, Arkansas on student achievement. In Springer, M. (Ed.),

Performance Incentives: Their Growing Impact on American K-12 Education. Washington, D.C.: Brookings Institution Press.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61-78.

Woessmann, L. (2011). Cross-country evidence on teacher performance pay. *Economics of Education Review*, *30*(3), 404-418.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement.* REL 2007-33. U.S. Department of Education.

Table 1—Teacher and school performance incentives, (quasi-)experimental evidence

Study	Place	Performance measure		Performance incentive
		Measure type	Individual or team measure	
Allen and Burgess (2012), Hussain (2015)	England	School inspections	School	School accountability
Atkinson et al. (2009)	England	Subjective	Teacher	Salary
Barrera-Osorio and Raju (2017)	Punjab, Pakistan	Student tests, enrollment, test participation	School	Bonus
Behrman et al. (2015)	Mexico	Student tests	Teacher or school	Bonus
Briole and Maurin (in-press)	France	Classroom observations, subjective	Teacher	Salary
Brown and Andrabi (2021), Andrabi and Brown (2022)	Pakistan	Student tests or subjective	Teacher	Bonus
Cilliers et al. (2018)	Uganda	Teacher attendance	Teacher	Bonus
Cohodes (2016)	Massachusetts	Student tests	School	School accountability
Dee and Jacob (2011)	United States	Student tests	School	School accountability
Dee and Wyckoff (2015)	Washington, DC	Student tests, classroom observations, subjective	Teacher	Bonus, salary, dismissal
Duflo, Hanna, and Ryan (2012)	Rajasthan, India	Teacher attendance	Teacher	Bonus
Fryer (2013), Goodman and Turner (2013)	New York City, NY	Several types including student tests	School	Bonus
Fryer et al. (2022)	Chicago Heights, IL	Student tests	Teacher or team	Bonus
Glewwe, Ilias, and Kremer (2010)	Kenya	Student tests	School	Prizes
Goldhaber and Walch (2012)	Denver, CO	Several types including student tests	Teacher	Bonus
Hanushek and Raymond (2005)	United States	Student tests	School	School accountability
Hudson (2010)	United States	Student tests, classroom observations	Teacher, school	Bonus

Table 1 (cont.)—Teacher performance incentives, (quasi-)experimental evidence

Study	Place	Performance measure		Performance incentive
		Measure type	Individual or team measure	
Imberman and Lovenheim (2015), Brehm, Imberman, and Lovenheim (2017)	Houston, TX	Student tests	Teacher, team	Bonus
Jackson (2010, 2014)	Texas	Student tests	Team	Bonus
Jacob (2005), Neal and Schanzenbach (2010)	Chicago, IL	Student tests	School	School accountability
Klein et al. (2000), Reback (2008), Deming et al. (2016)	Texas	Student tests, attendance, dropout rates	School	School accountability
Koretz and Barron (1998)	Kentucky	Student tests	School	School accountability
Ladd (1999)	Dallas, TX	Student tests	School	Bonus
Lavy (2002)	Israel	Several student outcomes	School	Bonus
Lavy (2009, 2020)	Israel	Student tests	Teacher	Bonus
Leaver et al. (2021)	Rwanda	Student tests, classroom observations, teacher attendance, other	Teacher	Bonus
Loyalka et al. (2019b)	Shaanxi and Gansu, China	Student tests	Teacher	Bonus
Martins (2009)	Portugal	Subjective	Teacher	Salary
Mbiti et al. (2019)	Tanzania	Student tests	Teacher	Bonus
Muralidharan and Sundararaman (2011), Muralidharan (2012)	Andhra Pradesh, India	Student tests	Teacher or school	Bonus
Ng (2021)	New Jersey	Student tests, classroom observations	Teacher	Tenure, dismissal
Rouse et al. (2007, 2013), Chiang (2009), Winters, Trivitt, and Greene (2010)	Florida	Student tests	School	School accountability

Table 1 (cont.)—Teacher performance incentives, (quasi-)experimental evidence

Study	Place	Performance measure		Performance incentive
		Measure type	Individual or team measure	
Santibanez et al. (2007)	Mexico	Student tests, subjective, teacher tests, other	Teacher	Salary
Sojourner, Mykerezi, and West (2014)	Minnesota	Student tests, classroom observations, other	Teacher, school	Bonus
Speroni et al. (2020)	United States	Student tests, classroom observations	Teacher	Bonus
Springer et al. (2010)	Nashville, TN	Student tests	Teacher	Bonus
Taylor (2022)	Tennessee	Student tests, classroom observations	Teacher	Tenure, dismissal
Taylor and Tyler (2012)	Cincinnati, OH	Classroom observations	Teacher	Dismissal, promotion
Vigdor (2009), Macartney (2016), Macartney, McMillan, and Petronijevic (2018, 2021), Aucejo, Romano, and Taylor (2022)	North Carolina	Student tests	School	Bonus, school accountability
Winters et al. (2008)	Little Rock, AK	Student tests	Teacher	Bonus

	Distinguished (4)	Proficient (3)	Basic (2)	Unsatisfactory (1)
Standard 2.3: The teacher manages and monitors student behavior to maximize instructional time.				
A. Monitoring of Student Behavior and Response to Misbehavior	▪ Teacher monitors behavior in a manner that anticipates and prevents student misbehavior, and that allows for students to monitor their own and/or their peers' behavior, which promotes individual, group, and/or whole class time on task. ▪ Teacher response to misbehavior is appropriate, consistent, and sensitive to students' individual needs. The desired behavior is attained. -or- ▪ Student misbehavior is not evident.	▪ Teacher monitors student behavior at all times which promotes individual, group, and/or whole class time on task. ▪ Teacher response to misbehavior is appropriate and consistent.	▪ Teacher monitors student behavior in a manner which results in a loss of individual, group, and/or whole class time on task. ▪ Teacher does not respond or does not respond appropriately to some off-task or disruptive behavior.	▪ Teacher does not consistently monitor student behavior and/or teacher is unaware of student behaviors, which result in considerable loss of individual, group and/or whole class time on task. ▪ Teacher does not respond to off task or disruptive behavior. -or- ▪ Teacher response to student misbehavior is inconsistent and/or has minimal results.
Standard 3.4: The teacher engages students in discourse and uses thought-provoking questions aligned with the lesson objectives to explore and extend content knowledge.				
B. Thought-Provoking Questions	▪ Teacher routinely asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. ▪ Teacher seeks clarification and elaboration through additional questions. ▪ Teacher provides appropriate wait time.	▪ Teacher asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. ▪ Teacher seeks clarification through additional questions. ▪ Teacher provides appropriate wait time.	▪ Teacher asks questions that are relevant to the objectives of the lesson. ▪ Teacher asks follow-up questions. ▪ Teacher is inconsistent in providing appropriate wait time.	▪ Teacher frequently asks questions that are inappropriate to objectives of the lesson. ▪ Teacher frequently does not ask follow-up questions. ▪ Teacher answers own questions. ▪ Teacher frequently does not provide appropriate wait time.

Figure 1a—Cincinnati FFT observation rubric, two example items

Source: Cincinnati Public Schools (2005). *Teacher Evaluation System. Revised 8/10/05.*

Behavior Management			
	Low (1,2)	Mid (3,4,5)	High (6,7)
Clear Behavior Expectations ▪ Clear expectations ▪ Consistency ▪ Clarity of rules	Rules and expectations are absent, unclear, or inconsistently enforced.	Rules and expectations may be stated clearly but are inconsistently enforced.	Rules and expectations for behavior are clear and consistently enforced.
Proactive ▪ Anticipates problem behavior or escalation ▪ Rarely reactive ▪ Monitoring	The teacher is reactive, and monitoring is absent or ineffective.	Teacher uses a mix of proactive and reactive responses; sometimes she monitors and reacts to early indicators of behavior problems but at other times misses or ignores them.	Teacher is consistently proactive and monitors the classroom effectively to prevent problems from developing.
Redirection of Misbehavior ▪ Effective reduction of misbehavior ▪ Attention to the positive ▪ Uses subtle cues to redirect ▪ Efficient redirection	Attempts to redirect misbehavior are ineffective; the teacher rarely focuses on positives or uses subtle cues. As a result, misbehavior continues and/or escalates and takes time away from learning.	Some of the teacher's attempts to redirect misbehavior are effective; particularly when he or she focuses on positive or uses subtle cues. As a result, misbehavior rarely continues, escalates, or takes time away from learning.	The teacher effectively redirects misbehavior by focusing on positives and making use of subtle cues. Behavior management does not take time away from learning.
Student Behavior ▪ Frequent compliance ▪ Little aggression and defiance	There are frequent instances of misbehavior in the classroom.	There are periodic episodes of misbehavior in the classroom.	There are few, if any, instances of student misbehavior in the classroom.

Figure 1b—CLASS observation rubric, one example item

Source: Pianta, La Paro, and Hamre (2008). *Classroom Assessment Scoring System: Dimensions Overview.*