Teacher Evaluation

Eric S. Taylor
K-12 Education
Workforce: Teachers

Introduction

How should schools measure a teacher's (or a team of teachers') job performance? How should job performance affect compensation, promotion, or other rewards and sanctions for teachers? How should schools provide performance feedback to teachers? Should teachers be evaluated individually or as teams? Answers to these questions span a variety of policies and management practices often grouped together as *teacher evaluation.*

The goal of evaluation policies and programs is to improve the quality of teaching in schools. There are large differences between teachers in how successful they are at promoting student achievement and development. Measuring each teacher's performance is the first but not the only step. Evaluation results inform performance feedback, rewards and sanctions, selection decisions, and investments in skill development.

Common performance *measures* for teachers include: (i) Ratings based on classroom observations. These ratings assess a teacher's effectiveness in specific tasks, such as asking questions or responding to misbehavior, using a structured rating rubric. (ii) Measures of a teacher's (or team of teachers') contribution to their students' achievement test scores or other outcomes. These measures are often called "value-added" measures. (iii) Other measures based on student test scores, such as the percentage of students who score proficient or higher. (iv) Other kinds of subjective ratings by principals or students. Many evaluation programs combine several measures.

Common *incentives* for teachers, linked to performance measures, include: earning tenure, eligibility for promotion, increases in salary, annual bonuses, and the threat of dismissal, among others. Though, state and local policies vary widely in whether and how performance measures determine these rewards and sanctions.

Teacher evaluation is fundamental to the operation of schools, and has become a central theme of education policy innovation and debate over the last two decades. Most notably, many states adopted new teacher evaluation policies with the encouragement of incentives from the Obama administration. Yet, teacher evaluation did not begin in the Obama era. School districts have been innovating locally for decades and continue to innovate today. "School accountability" policies are teacher evaluation policies. A school is a team of teachers (and other adults) measured, rewarded, and sanctioned as a team. Moreover, school leaders will always be assessing teacher performance—with or without formal rules for measuring. Those assessments will shape teachers' careers and compensation—with or without formal rules for incentives. "No teacher evaluation" is not a policy option. Informal evaluation will occur in the absence of a formal evaluation policy.

Does teacher evaluation improve teaching? There are three overlapping rationales for how evaluation works to improve teaching in schools. The first rationale: Teachers change their behavior at work—that is, their choices, attention, and effort—when their job performance is evaluated and rewarded. For example, a teacher may spend more time in class on tested subjects, such as math and reading, or spend more time outside class on lesson planning. A teacher may increase her total effort or allocate her effort in a different way. Some people reject this first rationale. Teachers are intrinsically motivated by seeing their students

learn and flourish, they argue, and thus teachers already give their best effort at work. Extrinsic incentives will not change what teachers do, they argue. Other people do not see a contradiction. They argue that everyone, including teachers, chooses leisure and family over work at some point each day or each week. Intrinsic motivation is not infinite. Extrinsic incentives may motivate teachers to work a little more, even if their primary motivation is intrinsic.

Nevertheless, if evaluation does change what teachers do, those changes may be unintended or unwanted. Rewarding higher math test scores may succeed in students learning more math, but possibly at the expense of learning less about the arts or history. Teachers may give more attention to some students and less attention to others. Additionally, teacher performance measures are imperfect. All evaluators make mistakes, and some may inject their personal biases. All performance scores include some measurement error. These imperfections weaken the incentives for teachers to change.

The second rationale: Evaluation can improve teaching by changing the composition of the teaching workforce. A school system might use performance measures to selectively hire, retain, or dismiss teachers, and thus improve the effectiveness of the system's teacher workforce, even if individual teachers do not change. Teachers may also self-select—choosing where to work because of the school system's evaluation measures and incentives.

The third rationale: Evaluation can improve teaching skills. Performance measures, and the associated feedback, provide guidance on how a teacher can improve her skills, thus reducing the costs of investing in her skills. Performance incentives create new rewards for improved skills.

Key findings

Key finding #1: *Modern evaluation measures reveal meaningful differences in effectiveness between teachers, but available measures are not without shortcomings.*

The most common performance measures are based on student test scores or based on observing teaching in the classroom. "Value-added" scores measure a teacher's contribution to her students' achievement test score growth, and predict her students' success years in the future. Value-added scores isolate a teacher's causal effect, separate from the backgrounds of the students she is assigned to teach. But value-added scores are noisy: A value-added score equals a teacher's true effect plus or minus some amount of statistical error. Modern classroom observations rate a teacher's concrete, observable teaching actions using detailed rubrics. Observation scores differ between teachers partly because of the different students in each classroom. Observation scores are also noisy, like value-added scores, and observation scores may be systematically biased. Value-added scores, observation ratings, and other available measures do not capture all of a teacher's responsibilities and contributions.

Key finding #2a: *Teachers generally do change their behavior—choices, attention, effort—when their performance is evaluated, and measurable student outcomes improve.*

For example, when students' math scores are used to evaluate teachers, student math scores go up. When teacher or school performance measures emphasize certain students, like students at risk of failing, those students do better. Teachers self-report that they change their behavior in response to evaluation. Researchers have now documented many examples of changes in teacher behavior in a variety of places and program designs. Still, there are also some notable cases where researchers found no change in teacher behavior.

Key finding #2b: *Those changes in behavior require teachers to make tradeoffs, sometimes intended, but other times unexpected or costly to students.*

Student math scores may improve, but whether students are better off generally is often more difficult to assess. The costs can be lower achievement in other subjects, for some students, or in the future. Researchers have documented many examples of these costs and distortions.

Key finding #3: *Schools could improve student achievement by selecting teachers—hiring, retaining, tenuring, and firing—based on their performance scores.*

Policy simulations demonstrate large improvements in student achievement under a policy of probationary screening, that is, dismissing newly hired teachers if their performance on the job falls below a defined cutoff. One policy evaluation, in Washington, DC, documents improvements empirically. In practice, however, few school systems select teachers based on performance evaluations.

Key finding #4: *Teachers themselves may self-select—choosing which schools to work at, or whether to work in teaching at all—based on their expectations about how they will be evaluated and incentivized.*

Self-selection affects the composition of the teacher workforce and, thus, affects the quality of teaching in schools. However, at least for U.S. schools, there is little empirical research on patterns of self-selection and how self-selection affects students.

Key finding #5: *Performance evaluation also improves teachers' skills by informing and incentivizing investments in skill growth.*

Those improvements in teaching skills result in greater student achievement growth. Some evidence suggests that evaluation and feedback alone, without incentives, may be sufficient to boost teachers' skills; teachers can use the new information to motivate or direct their own efforts to improve their skills. However, extrinsic rewards linked to evaluation scores, such as earning tenure, can further incentivize teachers to invest effort in their own skill growth.

Evidence

Key finding #1: *Modern evaluation measures reveal meaningful differences in effectiveness between teachers, but available measures are not without shortcomings.*

Teacher value-added scores based on student tests.

Teacher value-added scores measure a teacher's *contribution* to her students' achievement test score growth. A teacher's value-added score is a statistical estimate. First, calculate (a) the average test score achieved by the teacher's students, in a given year and subject, and then subtract (b) the average expected score for those same students. The expected scores, (b), are multivariate regression estimates based on the scores of other students who had different teachers, but have similar backgrounds and began the year with similar achievement. The difference (a) – (b), however, is caused partly by the teacher, but also partly by peer effects and other idiosyncratic differences from class to class. Second, isolate the teacher's contribution to (a) – (b) by examining what is consistent about (a) – (b) across classes and years as the teacher teaches many different classes. A value-added score captures that consistent component.

Using statistics jargon, researchers often say: "Value-added scores are unbiased but noisy." What does this mean? Ideally, value-added scores would measure each teacher's true causal effect on the achievement scores of her students. In practice, value-added scores are statistical estimates; each teacher's value-added score is equal to (i) her true causal effect (ii) plus some amount of error. "Unbiased" means that the error, (ii), is not systematically positive or negative for some teachers. Thus, on average, teacher value-added scores do equal the true causal effect. Value-added scores would be "biased" if, for example, teachers assigned students who were behind grade level always had lower value-added scores. A number

of empirical studies, both experimental and quasi-experimental, support the conclusion that value-added scores are unbiased.[1] For an explanation of those studies see the Handbook's main chapter on value added. While the empirical evidence is consistent with unbiasedness, researchers continue to work on important methodological questions.[2]

"Noisy" means that the error, (ii), creates noticeable differences between teachers in their value-added estimates. Two teachers with the same true causal effect might have different value-added scores because one teacher's score had more error than the other teacher's score. These differences due to noise are not informative. Empirically, perhaps half or more of the variation in one-year teacher value-added scores is noise.[3]

Value-added scores reveal educationally and economically meaningful differences between teachers in how successful they are at promoting student achievement. Researchers have now estimated teacher value-added in many schools, districts, states, and countries. The differences between teachers are notably consistent. One standard deviation in teacher value-added scores is typically 0.10–0.20 in student test score standard deviations (s.d.) units.[4] Thus, a teacher whose value-added score is one standard deviation above the average teacher will have students who score 0.10–0.20 s.d. higher. Alternatively, imagine two classrooms of students which start the year with similar achievement levels. The first class is taught by a teacher who is in the top quartile of teachers, as judged by value-added scores, and the second class is taught by a bottom-quartile teacher. At the end of the year, the first class will score 0.13–0.27 s.d. higher, which is equivalent to 5–10 percentile points.

In the short run, students learn more when they are assigned to a teacher with higher value-added scores. But those same students are also more successful in the long run. In their seminal work on teacher value-added, Raj Chetty, John Friedman, and Jonah Rockoff followed students from elementary school into adulthood. Students with a higher value-added teacher in elementary school were more likely to graduate from college, less likely to have children as teenagers, earn more in the labor market, and save more for retirement, among other positive outcomes. They estimate that a classroom of students whose teacher was one standard deviation above the value-added average would go on to earn, collectively, $200,000 more in present value than students in the average teacher's classroom.[5] The long-run effects of higher value-added teachers is an area of ongoing research.[6]

For more information, see the Handbook's main article on teacher value added or other summaries of the methods and related literature.[7]

Scores from classroom observations.

Modern classroom observation systems score a teacher's observable actions while she teaches. Observers rate a teacher on many separate tasks or skills using a detailed rubric. The rubric describes in practical terms what the rater must observe to assign each rating level. As an example, Figure 1 reproduces two tasks from the *Framework for Teaching* rubric, "Using Questioning and Discussion Techniques" and "Managing Student Behavior." In this example, teachers are rated on a scale from 1 to 4 corresponding to labels of Unsatisfactory, Basic, Proficient, or Distinguished. States and school districts typically create their own rubric, but many are based on the Framework for Teaching or similar rubrics. Most rubrics have 4 or 5 rating categories with descriptive labels.

For some teachers and administrators, the practical concrete nature of the rubrics makes observation ratings seem less prone to bias or error compared to value-added scores. This is not the case in reality. First, observation ratings typically have just as much "noise" as value-added scores, perhaps more. The Measures of Effective Teaching (MET) Project studied five different rubrics. The reliability of a single observation score, averaging across many rubric items, was 0.14–0.37 (the maximum is 1.0). Reliability reached 0.65 only with four different observations from four different raters.[8] The noise in observation

scores arises in part because any one classroom visit by a rater is only a small sample of a teacher's performance, sometimes just 15 minutes. Second, a growing body of evidence finds important sources of bias in observation ratings. Although most rubrics focus on teacher behaviors, differences between teachers in their observation scores are partly caused by differences in the students they are teaching. For example, researchers have documented how teachers assigned many students who are behind grade level or many students of color consistently receive lower observation scores.[9] Evaluation systems do not attempt to adjust observation scores based on student background or prior achievement, in contrast to value-added scores which do make adjustments. Additionally, raters tend to give higher scores to teachers with whom they share the same race or gender.[10] In short, observation scores do not consistently measure a teacher's true performance in the classroom.

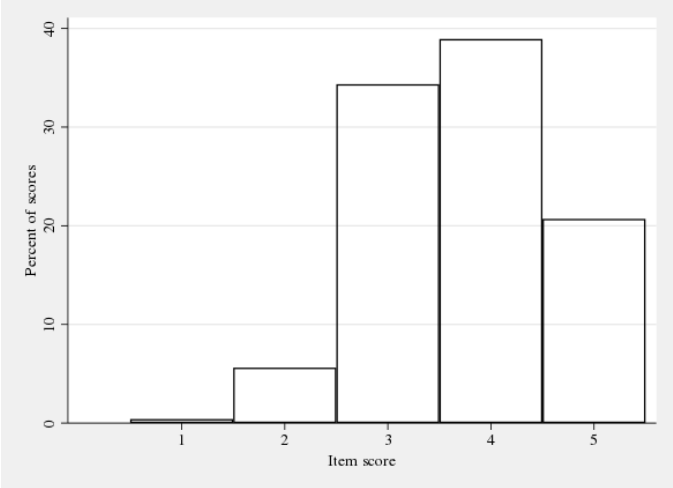
Figure 1. Example classroom observation rubric items

Domain: 3. Instruction Component: 3b. Using Questioning and Discussion Techniques Element: Quality of Questions	
(4) Distinguished	Teacher's questions are of uniformly high quality, with adequate time for students to respond. Students formulate many questions.
(3) Proficient	Most of the teacher's questions are of high quality. Adequate time is provided for students to respond.
(2) Basic	Teacher's questions are a combination of low and high quality, posed in rapid succession. Only some invite a thoughtful response.
(1) Unsatisfactory	Teacher's questions are virtually all of poor quality, with low cognitive challenge and single correct responses, and they are asked in rapid succession.
Domain: 2. The Classroom Environment Component: 2d. Managing Student Behavior Element: Response to Student Misbehavior	
(4) Distinguished	Teacher response to misbehavior is highly effective and sensitive to students' individual needs, or student behavior is entirely appropriate.
(3) Proficient	Teacher response to misbehavior is appropriate and successful and respects the student's dignity, or student behavior is generally appropriate.
(2) Basic	Teacher attempts to respond to student misbehavior but with uneven results, or there are no major infractions of the rules.
(1) Unsatisfactory	Teacher does not respond to misbehavior, or the response is inconsistent, is overly repressive, or does not respect the student's dignity.

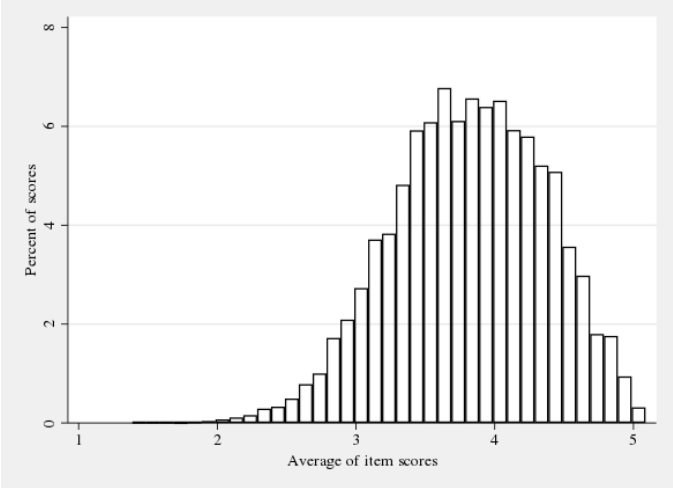
Note: These are 2 example elements drawn from 76 total elements which are organized into 22 components and 4 domains in the Framework for Teaching. The text is quoted directly quoted from: Danielson, Charlotte. 2007. Enhancing Professional Practice: A Framework for Teaching (2nd ed.). Association for Supervision and Curriculum Development.

Figure 2. Example distribution of classroom observation scores

(a) Item level scores



(b) Year average of item scores



Note: Classroom observation scores from several school years for teachers teaching grades 4–8 math and English language arts, and in years 1–6 of their teaching career. Panel (a) shows item-level scores—one score for each time a task was scored, with a sample of over 565,000 item-score observations. Panel (b) shows a teacher's annual average of item scores, with a sample of over 15,000 teacher-by-year observations.

Reproduced from: Taylor, Eric S. 2024. Employee Evaluation and Skill Investments: Evidence from Public School Teachers (NBER Working Paper No. w30687). National Bureau of Economic Research.

One common criticism of observation scores is that raters rarely give teachers low scores. For example, panel (a) of Figure 2 shows a histogram of over 500,000 observation ratings from one state over several years. Each of the ratings is for one of 19 specific items scored during a specific classroom visit; each item is a skill or task similar to the two examples in Figure 1. Only 5–6% of item ratings were "below expectations" (2) or lower.[11] The left skew in Figure 2 is often called "leniency bias" in performance evaluation settings. (This is a different notion of bias from the bias in value-added scores discussed above.) Some policy advocates predicted that modern classroom observation systems, adopted during the 2010s, would not exhibit leniency bias. This did not turn out to be true. The pattern in Figure 2 is found in data from many other districts and states, although systems with only 4 rating levels often have stronger ceiling effects.[12] Moreover, classroom observation ratings from trained, impartial raters in the MET Project also showed leniency bias.[13] Furthermore, leniency bias is not unique to teachers and schools; similar skew occurs in many occupations and sectors.[14] In short, some left skew is likely inevitable and is not necessarily a sign of a failed evaluation system. Finally, some of the apparent problem is created by evaluation systems when they round off (or round up) evaluation ratings. Panel (b) of Figure 2 shows a histogram of teacher observation scores, where each score is the average of a teacher's many item-level ratings in one school year. Panels (a) and (b) are the same data, just represented differently.

Other teacher performance measures.

Performance measures based on classroom observations and based on student tests, like value-added scores, do not capture all of a teacher's responsibilities and contributions. Researchers are developing new performance measures, for example, measures of a teacher's value-added contributions to student social–emotional growth, as reflected in attendance, behavior, and grades.[15] Some school systems also ask students to rate their teachers or ask principals to provide their subjective evaluations. Finally, teachers are often evaluated as teams—most often as a school—using the percentage of students who score proficient on a state test. In other words, the accountability metrics of No Child Left Behind (NCLB) and the Every Student Succeeds Act are themselves performance evaluation measures for teachers.

Key finding #2a: *Teachers generally do change their behavior—choices, attention, effort—when their performance is evaluated, and measurable student outcomes improve.*

Empirical research on teachers now includes many examples of evaluation and incentive programs causing changes in teacher behavior—changes in teachers' choices, attention, and effort at work. Teachers respond to being evaluated, both with and without rewards and sanctions determined by evaluation scores. Many empirical examples show improvement in student achievement scores as a result of these changes in teacher behavior. However, the empirical research literature also includes examples where evaluation causes no effect on student achievement (or at least no statistically significant effect). The many examples have been discussed in detail in several reviews of the literature over the last decade.[16]

Individual teacher evaluation.

Evaluation programs in Washington, DC Public Schools (DCPS) and Cincinnati Public Schools are two notable policy case studies. The DCPS evaluation program, known as IMPACT, began in 2009. Some features have changed over the years, but the basic components remain the same. The details and research discussed here are from the first years of the program. Each teacher is evaluated using a combination of performance measures, including rubric-based classroom observations conducted by principals or master teachers, value-added contributions to student achievement tests (for some teachers), and other ratings by the school principal. These component measures are combined into an overall weighted average known as the teacher's annual IMPACT score. IMPACT scores correspond to four groups based on predetermined cut points: "Ineffective," "Minimally Effective," "Effective," and "Highly Effective." Teachers rated highly effective (approximately the top 15%) received a cash bonus that year, and a permanent salary increase if they were rated highly effective in two consecutive years. By contrast, teachers rated ineffective (approximately the bottom 1%) were dismissed immediately, as were teachers

rated minimally effective (approximately the bottom 10%) in two consecutive years. Many teachers rated minimally effective chose to leave the district voluntarily.

The IMPACT program—including its performance measures, rewards, and sanctions—improved student achievement in Washington, DC, in two ways. First, some DCPS teachers changed their behavior at work. Most notable were two groups: teachers whose IMPACT score was near the cutoff for earning a bonus or a salary increase and teachers who had a score near the cutoff for being fired. The next year, both groups improved their teaching by roughly one quarter of a standard deviation.[17] Second, other DCPS teachers left the district and were replaced by more effective teachers. This turnover in the teacher workforce improved student math achievement scores by 0.08 student standard deviations (s.d.), on average. In classrooms that would have had an ineffective or minimally effective teacher, but that teacher left DCPS because of IMPACT, student scores increased by 0.21 s.d. in math and 0.14 s.d. in reading.[18]

Cincinnati's evaluation program—Peer Assistance and Review (PAR)—was developed and operated jointly by the Cincinnati Public Schools and Cincinnati Federation of Teachers beginning in the late 1990s. The Cincinnati program was similar to PAR programs in several other districts. Each teacher was paired with a peer evaluator for an entire school year. The peer evaluator scored the teacher's performance using rubric-based classroom observations and provided feedback and support for improvement. Tenured teachers were evaluated every five years and pre-tenure teachers every year. A teacher could be dismissed if she received persistently low evaluation scores, although dismissal was rare in practice.

Cincinnati's PAR program improved teaching and student achievement. Researchers studying the program used panel data to track teachers' value-added contributions to student achievement over several years: before, during, and after each teacher's PAR evaluation year. The students a teacher taught during her PAR evaluation year scored 0.05 test-score standard deviations (s.d.) higher in math compared to the students she taught in the years before being evaluated. The students she taught in the years after being evaluated scored 0.11 s.d. higher. In other words, the positive benefits of evaluation persisted for years after the teacher was no longer actively being evaluated (for more, see the section on key finding #5). These higher scores were gains in teacher value-added, but, notably, value-added scores were not part of Cincinnati's evaluation program at the time. Evaluation using rubric-based observations improved teachers' value added.[19]

A U.S. Department of Education field experiment in eight districts across the United States also documented achievement gains caused by similar classroom observation measures and feedback.[20]

Several state and district teacher evaluation programs have improved student achievement scores. Beyond Washington, DC, and Cincinnati, researchers have documented achievement gains in Chicago,[21] Dallas,[22] Houston,[23] Little Rock,[24] Minnesota,[25] New York City,[26] North Carolina,[27] and several districts in the Teacher Incentive Fund (TIF) program's experimental evaluation.[28] The evidence in these examples is not uniformly positive. For example, the Houston and New York City programs involved team-based evaluation and rewards. When teams of teachers were larger or otherwise more difficult to coordinate, the student achievement gains were smaller or even zero, on average.[29] Additionally, increases in student test scores do not always mean that students' mastery of the tested subject improved (for more, see the section on key finding #3).

Researchers have also documented unsuccessful programs, with little to no change in student achievement. Notable programs in Denver[30] and New York City[31] saw no gains for the average teacher's students. The well-known Project on Incentives in Teaching (POINT) experiment in Nashville provided bonuses to top value-added teachers, but produced no changes in teacher value added.[32] POINT promised $5,000, $10,000, and $15,000 bonuses to the top 20%, 10%, and 5% of teachers, respectively, based on their annual value-added scores. One explanation for the lack of effects in Nashville is that the cutoffs for winning were set too high. Moreover, the noise in value-added scores, discussed earlier, makes the effective cutoffs even higher.[33] More recently, the Gates Foundation funded teacher evaluation and personnel management reforms in Hillsborough County, Memphis, Pittsburgh, and four California charter

school networks. Researchers from RAND and the American Institutes for Research (AIR) concluded that student achievement in the Gates-supported schools did not grow any faster than did student achievement in a group of comparison schools in the same states. However, the comparison schools were also adopting similar teacher evaluation reforms at the same time, contaminating the comparison.[34]

Nearly all states (all but six) adopted new teacher evaluation rules during the Obama administration. These state-level policy reforms were the most visible and widespread action on teacher evaluation in the last decade plus. Still, whether these reforms benefited student achievement or not remains unclear. Researchers compared student test scores before and after these state-level policy changes, using difference-in-differences methods with the six holdouts as a comparison group. In this comparison, the average school district's test scores did not improve or decline after its state adopted new evaluation rules. Though it is unclear whether teacher evaluation practices changed meaningfully in that average U.S. school district as a result of the new state policy. The researchers did find that student achievement improved in some select states and districts thought to be examples of successful change.[35] The same researchers have also shown that state-level reforms affected self-selection into teaching, in both positive and negative ways.[36]

These examples from the United States are complemented by examples from other countries. Teacher evaluation policies and experiments have produced student achievement gains in settings as diverse as China, England, France, India, Israel, Kenya, Mexico, Pakistan, Portugal, Rwanda, and Tanzania.[37] Among the Organisation for Economic Co-operation and Development (OECD), countries with more pay for performance score higher on the Programme for International Student Assessment (PISA) exams.[38]

School accountability.

School accountability policies are teacher evaluation policies. Each school is a team of teachers (and other adults) measured, rewarded, and sanctioned as a team. Teachers change their behavior in response to school accountability policies and student achievement improves. Scores on the National Assessment of Educational Progress (NAEP) improved when states adopted school accountability with meaningful potential sanctions in the 1990s and 2000s.[39] Large gains occurred in the schools most at risk of sanctions and for students at risk of failing the exams used to evaluate the schools.[40]

Key finding 2b: *Those changes in behavior require teachers to make tradeoffs, sometimes intended, but other times unexpected or costly to students.*

Reallocating effort across subjects, students, or other responsibilities of teachers. Teachers have many responsibilities at work. They must decide how to allocate their effort—energy, time, attention—across the many tasks they are assigned.

One concern, often raised in education policy debates, is that evaluating teachers based on math, reading, and English test scores will reduce class time and achievement growth in other subjects such as science, social studies, or the arts. The cost of higher math scores might be lower social studies scores. When asked, teachers self-report reallocating time from untested subjects to tested subjects.[41] To date, however, researchers have found limited evidence that teacher evaluation based on math and language causes achievement losses in other subjects. Scores in science, social studies, etc. may not improve, but they generally do not decline. Perhaps teachers are reducing time for art or physical education.

In the 1990s, the Chicago Public Schools began evaluating schools based on students' math and reading scores from the Iowa Test of Basic Skills (ITBS). Math and reading scores improved as a result of the new school accountability program. Science and social studies ITBS scores also improved, on average, contrary to the often-raised concerns. The gains in math and language were much larger, consistent with the new incentives. Additionally, there may have been more of a tradeoff across subjects for Chicago's lowest-achieving students.[42] Similarly, in the 2000s, Florida schools were graded based on students' math and language scores. Schools that received a failing grade subsequently improved those math and language

test scores, but they also improved in science.[43] Beyond these two case studies, math and language scores on the National Assessment of Educational Progress (NAEP) improved when states adopted school accountability programs in the 1990s and 2000s. Science scores, which were not part of those accountability programs, did not improve but did not decline.[44]

Additionally, in separate field experiments in Andhra Pradesh, India and Tanzania, treatment teachers were evaluated and rewarded based on student scores in math and language. Teacher evaluation improved math and language scores, but also improved science and social studies scores for the Indian and Tanzanian students.[45]

Academic subjects are not the only things students learn in school. Teachers contribute to students' social–emotional development as well. Teachers scored based on math and reading tests may neglect students' social–emotional skills. That kind of tradeoff has not been studied directly, though an experiment in Pakistan offers suggestive results. Students' self-reported social–emotional outcomes were worse when their teachers were evaluated based on test scores alone compared to when broader subjective evaluations were used.[46]

A second concern is that evaluating teachers may benefit some students but harm others. Some teachers must allocate effort across different subjects. All teachers must choose how to allocate their effort across different students. Performance measures can (unintentionally) cause teachers to give more energy, time, and attention to some of their students and less to their other students.

One widely known example of reallocation even has a slang name: "bubble kids." This concern applies to a specific but ubiquitous performance measure: when schools and teachers are evaluated based on the percentage of students who score proficient or higher on some test (synonymously, the percentage of students who pass the test). Students who are "on the bubble" are those who may or may not pass the test. Some students will certainly pass, and others have little chance of passing. When teachers and schools are evaluated based on the "percent proficient," students on the bubble receive more attention from their teachers. Chicago's 1990s program, mentioned earlier, used this percent-passing-style evaluation measure. When the program began, achievement scores increased the most among students in the middle of the distribution, near the passing cutoff. In Chicago, the scores of "bubble kids" increased much more than those of their higher- and lower-performing classmates. Researchers found the same pattern in the 2000s when Chicago adopted new "percent proficient" measures under NCLB.[47] Similar quasi-experiment results have been documented in several settings, including North Carolina,[48] Texas,[49] and England.[50]

A third concern is that evaluating teachers may improve student test scores in the short run, but at the expense of students' flourishing in the long run. Studies that follow students into adulthood are rare, but the studies we do have suggest long-run benefits, not costs. Texas was one of the first states to adopt a test-based accountability policy in the 1990s. When a high school was under pressure to improve math and English test scores, students' scores did improve. More importantly, those same students were more likely to finish college and were earning more in the labor market in their late 20s. All because their high school teachers were evaluated based on math and English test scores.[51] The Texas program was team evaluation (school accountability), not individual evaluation; the effects of individual evaluation on long-run outcomes for students is an important opportunity for future research contributions. There is reason to be optimistic: An experiment in Israel provides similar evidence of college completion and labor market benefits caused by individual teacher evaluation.[52]

Teacher actions which change or corrupt the conclusions we make about student achievement based on student test scores.

Administrators and policymakers designing teacher evaluation programs should carefully consider how performance measurement, rewards, and sanctions might motivate teachers to act in unwanted ways. In other words, remember Campbell's Law: "The more any quantitative social indicator is used for social

decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[53]

Some actions by teachers and schools directly corrupt test scores. Allegations of outright cheating by teachers sometimes appear in the popular press. Researchers found compelling evidence of cheating in the Chicago Public Schools—cheating caused by new evaluation measures and sanctions introduced in the 1990s.[54] Some actions are more subtle than changing students' test answer sheets. For example, a school may prevent low-achieving students from taking the test, thus raising the school's average test score. Researchers have documented a variety of empirical examples: encouraging certain students to be absent,[55] suspending students during the test period,[56] and designating students as special education students.[57]

The consequences of cheating are straightforward. We use math test scores to make conclusions about students' actual mastery of math concepts and skills (and use other subject test scores similarly). When teachers and schools cheat, math test scores increase without any change in students' true mastery of math. Consequently, we will make incorrect conclusions about students' achievement growth and about teachers' contribution to that achievement growth. Cheating is an extreme case. However, teachers may take other actions in response to evaluation pressure; actions which are not obviously corrupt, but actions which nevertheless result in incorrect conclusions about students' true mastery of math and other subjects.

What other actions should we be concerned about? Discussions about test-based evaluation often involve debates about "narrowing instruction," "teaching to the test," "test prep," "score inflation," etc. A teacher feeling evaluation pressure may choose to spend more class time on the math concepts or skills that she knows are more likely to be on the test. Perhaps focusing on regular polygons instead of irregular polygons, for example. This behavior is sometimes called "narrowing instruction," in response to evaluation, or "teaching to the test." If a teacher narrows instruction to match the test, students' true mastery of *some* math concepts will still improve. The improvements will be on the tested skills (regular polygons) but come at the cost of less mastery of other non-tested math skills (irregular polygons). These dynamics suggest we should exercise some restraint when making conclusions based on test scores. The topic of "narrowing instruction" and "teaching to the test" could occupy several paragraphs alone. The conditions under which narrowing instruction might be beneficial (or harmful) to students depend on several factors. Daniel Koretz, Edward Lazear, and others have written thoughtfully about this issue and the tradeoffs for students.[58]

The conclusions we make about teacher performance should acknowledge the possibility of "narrowing instruction." However, there is mixed evidence on whether teachers do, in fact, narrow instruction in response to evaluation. The most common research design requires two different tests: the "high-stakes" test used to evaluate teachers and determine rewards or sanctions and a second "low-stakes" test covering the same subject area. Evaluation creates an incentive for teachers to narrow instruction to match the high-stakes test, not the low-stakes test. Imagine scores improve on the high-stakes test, but not on the low-stakes test; that pattern would be consistent with narrowing instruction.

Minnesota's Q-Comp program paid cash bonuses to teachers based on student test scores. Students took both the state's Minnesota Comprehensive Assessment (MCA) and a Northwest Evaluation Association (NWEA) test. In some districts, Q-Comp bonuses were based on MCA scores, making MCA high stakes and NWEA low stakes. In other districts the reverse was true. In all districts, Q-Comp raised student scores on both the high- and low-stakes tests. The improvements on both tests is evidence that is inconsistent with narrowing instruction.[59] By contrast, an experiment in Tanzania found the opposite result. Student scores increased on the test used to determine teacher bonuses, but the same students' scores did not change on a low-stakes test.[60]

Several studies of statewide evaluation programs use the National Assessment of Educational Progress (NAEP) as the low-stakes test. In the 2000s, North Carolina's ABCs program gave bonuses to teachers based on state tests in math and reading. Scores on the state tests (high stakes) improved in math and reading as a result of the ABCs program. However, concurrent with those state test improvements, North Carolina's NAEP (low stakes) scores improved in math but not in reading.[61] School accountability programs in Texas and Kentucky, during the 1990s, improved state test scores but did not always improve NAEP scores.[62] A national study found the same mixed evidence for the average state. Researchers estimated the causal effects of the school accountability programs created to satisfy the requirements of No Child Left Behind (NCLB). The researchers documented improvements in NAEP math scores, but not reading scores, caused by NCLB accountability. No teacher was "teaching to the NAEP," and yet NAEP math scores improved because teachers and schools were evaluated based on student achievement measures.[63]

A second research design, less often available, involves comparing students' answers to different questions (synonymously, test items) on the same test. After Chicago's reforms, student math scores improved on both basic skills questions (e.g., computation, number concepts) and more complex questions (e.g., estimation, word problems). But the improvements were twice as large on basic skills items. What explains the difference? Teachers familiar with the ITBS might predict—correctly, it turns out—that basic skills items appear on ITBS more often, and choose to narrow instruction to those basic skills when they are evaluated. Alternatively, basic skills may simply be easier to teach, or easier to learn, and thus teachers might choose to invest more class time on those skills even if both basic and complex skills appear on the ITBS with equal frequency.[64]

Teachers may devote class time to "test prep" in response to being evaluated based on student scores. For example, teachers might have students take practice tests or spend class time just before the test dates to review relevant material. Teachers themselves self-report spending more time on test prep, for example, in pay-for-performance experiments conducted by researchers in Nashville and Mexico.[65] And researchers observing classrooms also see more "test prep" activities when teachers are randomly assigned to be evaluated and rewarded based on student tests.[66] One common prediction is that test prep, especially intensive review before the test, may raise students' math or language skills in the short run, but students will forget what they learned quickly. Student test score gains from many educational interventions often "fade out" (synonymously, decay) over time. Gains induced by teacher evaluation programs also fade out, but not faster or slower than the gains caused by other interventions.[67]

Finally, teachers may teach test-taking skills, for example, strategies for multiple-choice questions. An experiment in Kenya provided rewards to teachers based on a multiple-choice test. Student scores did improve on that high-stakes incentivized test. Their scores also improved on the multiple-choice questions of a low-stakes test, but scores did not improve on the open-ended questions.[68]

This possibility—that teachers teach test-taking skills—should affect the conclusions we make about teacher performance based on student test scores. Improvements in scores may not reflect math and language skills; although test-taking skills themselves are valuable skills for success in life.

Key finding #3: *Schools could improve student achievement by selecting teachers—hiring, retaining, tenuring, and firing—based on their performance scores.*

Policies that tie selection decisions—hiring, firing, tenure—directly to a teacher's measured performance are still uncommon, despite the growing availability of evaluation measures. Notable exceptions include the IMPACT program in Washington, DC, described above, and several states and districts conditioning tenure on measured performance, described below.

Given the scarcity of existing policies to evaluate, researchers have used policy simulations to demonstrate the potential benefits of performance-based selection. Consider a school system which commits to a policy of probationary screening. Each year the schools hire some number of novice teachers. At the end of the school year, all the novices are ranked based on their value-added scores, the lowest scoring x% are

fired, and the highest scoring (100–x)% are retained and given tenure. What rate of dismissal, x%, will maximize student achievement? How much will student achievement improve?

Douglas Staiger and Jonah Rockoff answer these questions with simulations that take into account two key features of this problem. First, dismissing a teacher requires hiring a novice. Some classrooms will be taught by a novice instead of a second-year teacher, those students will (likely) learn less as a result, and that lost achievement is the main (potential) cost of dismissing a teacher. Even a less-effective second-year teacher may well be more effective than the average novice. However, if the new hire is highly effective and granted tenure, she will more than compensate for that one-year cost over a long career. Second, value-added scores are "unbiased but noisy," as discussed earlier in this chapter. When the school system dismisses teachers with value-added estimates in the bottom x%, it will dismiss some teachers whose true value added is in the top (100-x)% but who had an unlucky draw of negative measurement error. After building in these two features, among others, what does the policy simulation return? The optimal choice is to dismiss 80% of novice new hires each year. The gain in student achievement is 0.08 standard deviations (s.d.), in the steady state.[69]

Jessee Rothstein adds a third important feature to the simulation: The school system will need to raise teacher salaries to compensate prospective hires for the risk that they will work only one year. With this third feature added, the optimal dismissal rate falls to 40% and the achievement gains are 0.023 s.d.[70] Notably, even a 40% dismissal rate is much higher than what is typical in schools today. Only 10–20% of novice teachers leave after their first year, and they are not the lowest-performing 10–20%.

In several states and districts, earning tenure now requires scoring above some performance cutoff. Examples in the research literature include Michigan,[71] New Jersey,[72] New York City,[73] and Tennessee.[74] However, these actual policies differ in important ways from the hypothetical policies considered in the simulations. Most notably, while the simulation policies are strict—a teacher is fired if her performance is too low—actual policies typically allow teachers to continue working as teachers until they score high enough to earn tenure or they choose to quit. In the cases that have been analyzed empirically, teacher performance improves in response to the new tenure rules, at least pre-tenure, but there has been little differential selection of teachers.

One example of a strict dismissal policy is the IMPACT program in Washington, DC. As described above, DCPS fired any teacher rated minimally effective in two consecutive years (in the bottom 10% twice). Teachers rated ineffective (the bottom 1%) in just one year were fired immediately. In addition to the district's selection choices, teachers also self-selected out of DCPS. Many teachers left DCPS after being rated minimally effective only once. The teachers who were fired or left were replaced by more effective teachers. In classrooms that would have had an ineffective or minimally effective teacher, but that teacher left DCPS because of IMPACT, student scores increased by 0.21 s.d. in math and by 0.14 s.d. in reading.[75]

Selection policies may cause more self-selection by teachers, as the DCPS case suggests. Even without formal selection policies, schools may select teachers informally by encouraging low-performing teachers to leave. Research framed around self-selection is often helpful for thinking about selection by schools, and vice versa.

Key finding #4: *Teachers themselves may self-select—choosing which schools to work at, or whether to work in teaching at all—based on their expectations about how they will be evaluated and incentivized.*

Self-selection affects the composition of the teacher workforce, and thus affects the quality of teaching in schools. Consider the influence of an evaluation program: One teacher may prefer to work in a school with a system of performance evaluation and incentives because she expects to succeed and be rewarded. While a different teacher may avoid the same school because she expects to score poorly in evaluations and face sanctions. Some teachers may prefer a school with evaluation because they expect other high-

performing teachers will also prefer that school. However, at least for U.S. schools, there is little empirical research on patterns of self-selection and how self-selection affects students.

Evaluation can cause low-performing teachers to quit. In the 2010s, New York City provided school principals with reports detailing the value-added scores of their teachers. The pilot phase of the program included a field experiment, with treatment schools receiving value-added reports while control schools did not. In control schools, teacher turnover was unrelated to value added. By contrast, in treatment schools, revealing value-added scores caused many low-value-added teachers to leave the school. The teachers who left were choosing to leave (self-selection), though perhaps with some encouragement from their principal (selection by schools). Notably, receiving value-added reports caused principals to change their opinions; principals' ratings of their teachers became more aligned with the teachers' value-added scores. Student achievement improved in treatment schools the following school year, partly because of the turnover.[76]

In a related quasi-experiment, Guilford County Schools provided value-added reports to principals long before other districts in North Carolina provided them. As in New York City, low-value-added teachers often left their schools, but so did some high-value-added teachers. Unlike in New York City, all school principals in Guilford could view the value-added reports of teachers they were considering hiring away from another school in the county. The high-value-added teachers moved to more desirable schools in Guilford County. The low-value-added teachers moved to other school districts.[77] Perhaps the low-performing teachers left Guilford to avoid being evaluated with value-added scores, or perhaps because principals outside Guilford could not access the value-added data.

In a second field experiment, treatment schools in Chicago piloted a new teacher evaluation program which used rubric-based classroom observations to score teacher performance. While treatment and control schools had similar turnover rates, in treatment schools lower-performing teachers were more likely to leave their schools.[78] The pilot program also improved student achievement in treatment schools.[79] In the research literature, Houston and Washington, DC, are two other examples of evaluation programs increasing self-selection out of teaching among lower-performing teachers.[80]

Empirical evidence on teacher self-selection, in response to evaluation, remains scarce and indirect for schools in the United States, as these examples demonstrate. However, two recent and novel field experiments outside the United States provide some insights.

An experiment in Rwanda began by randomly assigning labor markets—defined by geographic district and subject taught—to one of two conditions. In one condition, advertisements for open teaching positions offered a conventional salary ("fixed wage" or FW). In the other condition, advertisements offered a smaller salary plus bonus pay based on performance ("pay for performance" or P4P). Prospective teachers applied, and schools hired teachers. At the start of the new school year, schools were randomly assigned to pay teachers either a conventional salary (FW) or a smaller salary plus bonus pay based on performance (P4P). The first and second randomizations were independent, and teachers and schools did not know about the second randomization until it happened. Thus, for example, some teachers who took a new job expecting P4P ended up actually working in an FW job, and vice versa.

Offering pay-for-performance, at the time of hiring, had little to no effect on who self-selected into teaching in Rwanda. If a teacher was recruited with P4P, her students scored just 0.01 standard deviations (s.d.) higher than students whose teacher was recruited with FW (and the difference is not statistically significant). However, teachers did behave differently when working at a school with pay-for-performance. Students scored 0.12 s.d. higher in P4P schools compared to FW schools, and that effect of actual P4P did not differ between classrooms where the teacher was recruited with P4P or FW.[81]

A second study in Pakistan also randomly assigned schools to either P4P or FW. Prior to random assignment, the researchers conducting the study asked the teachers whether they would prefer P4P or FW compensation, with incentives designed to elicit their true preferences. Among teachers who

preferred P4P, students scored 0.09 s.d. higher in P4P schools compared to FW schools. Among teachers who preferred FW, that same P4P–FW difference was just 0.01 s.d. This notable difference in the treatment effect of P4P is in contrast to the lack of difference in P4P effects in Rwanda. One potential explanation is that novice teachers, like the new hires in Rwanda, have limited information about how effective they will be in the classroom, especially how effective they will be at raising student test scores. Experienced teachers, like those in Pakistan, have more information on which to self-select. Indeed, in Pakistan the effects of P4P did depend on a teacher's preferred contract, P4P or FW, but did not depend on her prior value added. This result suggests teachers had private information about their own potential behavior under a program of evaluation and performance incentives.[82]

The Handbook's chapter on compensation includes additional examples relevant to self-selection and selection by schools.

Key finding #5: *Performance evaluation also improves teachers' skills by informing and incentivizing investments in skill growth.*

Several experiments and quasi-experiments suggest teacher evaluation can improve a teacher's skills. One notable example is Cincinnati's PAR program, described earlier, where a teacher was evaluated only every five years. Researchers used panel data to track a teacher's value-added contributions to student achievement over several years: before, during, and after the teacher's PAR evaluation year. The students a teacher taught during her PAR evaluation year scored 0.05 standard deviations (s.d.) higher in math compared to students she taught in the years before being evaluated. An improvement in teacher value added. Importantly, that improvement in value added did not go away after the evaluation year ended. The students scored 0.11 s.d. higher in the years after evaluation. If teachers simply work harder while being evaluated, then we would expect a teacher's value added after evaluation to be similar to her value added before evaluation. Contrary to that prediction, teacher value added remained higher after evaluation. Additionally, the improvements were largest for teachers who were the least effective prior to evaluation. The likely explanation for this pattern of results is that Cincinnati's PAR program caused growth in teachers' skills.[83]

Teacher value added also improved in France after a classroom-observation-based evaluation. As in Cincinnati, the evaluation occurred only every five or so years, allowing researchers to track performance before and after. As in Cincinnati, teacher value added in France remained higher in the years after evaluation.[84]

How might evaluation improve teaching skills? First, performance evaluation can reduce the costs of investing in skill development. Some advocates of evaluation emphasize the value of "feedback" as a shorthand. In more concrete terms, performance measures create new information about an individual teacher's strengths and weaknesses, her performance relative to other teachers, and sometimes explicit suggestions from the evaluator about how to improve. In particular, classroom observation ratings and rubrics, such as the example in Figure 1, provide practical descriptions of what teachers might do to improve. Teachers can use those valuable resources (synonymously, feedback) to help decide where to invest their effort in improving their skills. Absent an evaluation program, those resources would be costly for any individual teacher to produce on her own. Second, performance incentives—rewards or sanctions tied to evaluation scores—can increase the return to investing in skill development. Potential rewards such as cash bonuses, increases in salary, and the job protections of tenure are valuable incentives for a teacher to invest in improving her skills.[85]

A teacher evaluation program or policy could improve teacher skills through one or both mechanisms. The Cincinnati and France programs combined performance measures and feedback with rewards and sanctions; both mechanisms might have contributed to the gains in teachers' skills and student achievement. In other evaluation programs—for example, in Chicago and England—teacher value added improved as a result of performance measurement and feedback without any formal rewards or

sanctions.[86] A U.S. Department of Education field experiment in eight districts across the United States also documented achievement gains caused by classroom observation measures and feedback alone.[87]

Quasi-experimental evidence from Tennessee demonstrates the potential for extrinsic rewards to incentivize teachers' investments in their own skill development. Beginning in the 2011-12 school year, all Tennessee teachers were evaluated annually using rubric-based classroom observations, value added (when available), and other performance measures. All these measures were combined into a weighted average "Level of Effectiveness" (LOE) score at the end of the school year. The statewide reforms in 2011-12 also included new tenure rules. A teacher would earn tenure after her 5th year teaching if her LOE score was "above expectations" or higher (empirically, above the 33rd percentile) in both her 4th and 5th years.

Both the new performance measures and the new tenure rules caused improvements in student achievement in Tennessee by improving teacher value added. First, consider teachers who already had tenure before the 2011-12 reforms. These teachers were evaluated annually, using the new performance measures, but were not subject to the new tenure rules. Among tenured, but early-career teachers, value added increased by 0.024 student standard deviations (s.d.) in the first year of the new program. By comparison, teachers who were subject to the new tenure rules improved twice as much; their value added increased by 0.047 s.d. in the first year of the new program. That doubling of the effect was caused by the potential reward of earning tenure. Further, these improvements in value added had already begun in a teacher's 2nd and 3rd years. That is, improvements before a teacher's performance would "count" for earning tenure in her 4th and 5th years. And teacher value added did not decline after earning tenure. This pattern of anticipation effects—improvement before scores count—and persistent effects—continued higher performance after scores no longer count—is consistent with skill growth.[88]

Endnotes and references

[1] Bacher-Hicks, Andrew, and Cory Koedel. 2023. Estimation and Interpretation of Teacher Value Added in Research Applications. In Handbook of the Economics of Education. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier; Kane, Thomas J., and Douglas O. Staiger. 2008. Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation (NBER Working Paper No. w14607). National Bureau of Economic Research; Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment (MET Project Research Paper). Bill & Melinda Gates Foundation; Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. American Economic Review 104(9): 2593–2632; Bacher-Hicks, Andrew, Thomas J. Kane, and Douglas O. Staiger. 2014. Validating Teacher Effect Estimates Using Changes in Teacher Assignments in Los Angeles (NBER Working Paper No. w20657). National Bureau of Economic Research.

[2] Rothstein, Jesse. 2017. Measuring the Impacts of Teachers: Comment. American Economic Review 107(6): 1656–1684; Staiger, Douglas O., Thomas J. Kane, and Brian D. Johnson. 2024. Why Does Value-Added Work? Implications of a Dynamic Model of Student Achievement.

[3] Bacher-Hicks and Koedel (2023).

[4] Ibid.

[5] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. American Economic Review 104(9): 2633–2679.

[6] Jackson, C. Kirabo. 2018. What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. Journal of Political Economy 126(5): 2072–2107; Backes, Ben, James Cowan, Dan Goldhaber, and Roddy Theobald. 2024. How to Measure a Teacher: The Influence of Test and Non-Test Value-Added on Long-Run Student Outcomes. Journal of Human Resources.

[7] Bacher-Hicks and Koedel (2023); Harris, Douglas N. 2011. Value-Added Measures in Education: What Every Educator Needs to Know. Harvard Education Press.

[8] Kane, Thomas J., and Douglas O. Staiger. 2012. Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains (MET Project Research Paper). Bill & Melinda Gates Foundation; Ho, Andrew D., and Thomas J. Kane. 2013. The Reliability of Classroom Observations by School Personnel (MET Project Research Paper). Bill & Melinda Gates Foundation.

[9] Campbell, Shanyce L., and Matthew Ronfeldt. 2018. Observational Evaluation of Teachers: Measuring More than We Bargained For? American Educational Research Journal 55(6): 1233–1267.

[10] Chi, Olivia L. 2023. A Classroom Observer Like Me: The Effects of Race-Congruence and Gender-Congruence between Teachers and Raters on Observation Scores. Education Finance and Policy 18(3): 442–466.

[11] Taylor, Eric S. 2024. Employee Evaluation and Skill Investments: Evidence from Public School Teachers (NBER Working Paper No. w30687). National Bureau of Economic Research.

[12] Steinberg, Matthew P., and Matthew A. Kraft. 2017. The Sensitivity of Teacher Performance Ratings to the Design of Teacher Evaluation Systems. Educational Researcher 46(7): 378–396.

[13] Kane and Staiger (2012).

[14] Prendergast, Canice. 1999. The Provision of Incentives in Firms. Journal of Economic Literature 37(1): 7–63.

[15] Jackson (2018); Backes et al. (2024); Liu, Jing, and Susanna Loeb. 2021. Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. Journal of Human Resources 56(2): 343–379.

[16] Taylor, Eric S. 2023. Teacher Evaluation and Training. In Handbook of the Economics of Education. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier; James, Jessalynn, and James Wyckoff. 2021. Teacher Labor Markets: What Have We Learned over the Last Decade? In Routledge Handbook of the Economics of Education. Edited by Brian P. McCall. Routledge; Rowan, Brian, and Stephen W. Raudenbush. 2016. Teacher Evaluation in American schools. In Handbook of Research on Teaching. Edited by Drew Gitomer and Courtney Bell. American Educational Research Association; Neal, Derek. 2011. The Design of Performance Pay in Education. In Handbook of the Economics of Education. Edited by Eric A. Hanushek, Stephen Machin, and Ludger Woessmann. Elsevier.

[17] Dee, Thomas S., and James Wyckoff. 2015. Incentives, Selection, and Teacher Performance: Evidence from IMPACT. Journal of Policy Analysis and Management 34(2): 267–297.

[18] Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. Educational Evaluation and Policy Analysis 39(1): 54–76.

[19] Taylor, Eric S., and John H. Tyler. 2012. The Effect of Evaluation on Teacher Performance. American Economic Review 102(7): 3628–3651.

[20] Garet, Michael S., Andrew J. Wayne, Seth Brown, Jordan Rickles, Mengli Song, and David Manzeske. 2017. The Impact of Providing Performance Feedback to Teachers and Principals (NCEE 2018-4001). National Center for Education Evaluation.

[21] Steinberg, Matthew P., and Lauren Sartain. 2015. Does Teacher Evaluation Improve School Performance? Experimental Evidence from Chicago's Excellence in Teaching Project. Education Finance and Policy 10(4): 535–572.

[22] Hanushek, Eric A., Jin Luo, Andrew J. Morgan, Minh Nguyen, Ben Ost, Steven G. Rivkin, and Ayman Shakeel. 2023. The Effects of Comprehensive Educator Evaluation and Pay Reform on Achievement (NBER Working paper No. w31073). National Bureau of Economic Research.

[23] Imberman, Scott A., and Michael F. Lovenheim. 2015. Incentive Strength and Teacher Productivity: Evidence from a Group-based Teacher Incentive Pay System. Review of Economics and Statistics 97(2): 364–386.

[24] Winters, Marcus, Jay Greene, Gary Ritter, and Ryan Marsh. 2008. The Effect of Performance Pay in Little Rock, Arkansas on Student Achievement. National Center on Performance Incentives at Vanderbilt University.

[25] Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota. Journal of Human Resources 49(4): 945–981.

[26] Goodman, Sarena F., and Lesley J. Turner. 2010. 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.

[27] Vigdor, Jacob L. 2009. Teacher Salary Bonuses in North Carolina. In Performance Incentives: Their Growing Impact on American K-12 Education. Edited by Matthew G. Springer. Brookings Institution Press.

[28] Speroni, Cecilia, Alison Wellington, Paul Burkander, Hanley Chiang, Mariesa Herrmann, and Kristin Hallgren. 2020. Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation. Journal of Labor Economics 38(3): 843–872.

[29] Taylor (2023); Neal (2011).

[30] Goldhaber, Dan, and Joe Walch. 2012. Strategic Pay Reform: A Student Outcomes-based Evaluation of Denver's ProComp Teacher Pay Initiative. Economics of Education Review 31(6): 1067–1083.

[31] Fryer, Roland G. 2013. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. Journal of Labor Economics 31(2): 373–407.

[32] Springer, Matthew G., Laura Hamilton, Daniel F. McCaffrey, Dale Ballou, Vi-Nhuan Le, Matthew Pepper, J. R. Lockwood, and Brian M. Stecher. 2010. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives at Vanderbilt University.

[33] Taylor (2023); Neal (2011).

[34] Brian M. Stecher, Deborah J. Holtzman, Michael S. Garet, Laura S. Hamilton, John Engberg, Elizabeth D. Steiner, Abby Robyn, Matthew D. Baird, Italo A. Gutierrez, Evan D. Peet, Iliana Brodziak De Los Reyes, Kaitlin Fronberg, Gabriel Weinberger, Gerald Paul Under, and Jay Chambers. 2018. Improving Teaching Effectiveness: Final Report: The Intensive Partnerships for Effective Teaching through 2015–2016. RAND Corporation.

[35] Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew Springer. 2024. Taking Teacher Evaluation to Scale: The Effect of State Reforms on Achievement and Attainment (EdWorkingPaper No. 21-496). Annenberg Institute at Brown University.

[36] Kraft, Matthew A., Eric J. Brunner, Shaun M. Dougherty, and David J. Schwegman. 2020. Teacher Accountability Reforms and the Supply and Quality of New Teachers. Journal of Public Economics 188: 104212.

[37] Taylor (2023).

[38] Woessmann, Ludger. 2011. Cross-Country Evidence on Teacher Performance Pay. Economics of Education Review 30(3): 404–418.

[39] Dee, Thomas S., and Brian Jacob. 2011. The Impact of No Child Left Behind on Student Achievement. Journal of Policy Analysis and Management 30(3): 418–446; Hanushek, Eric A., and Margaret E. Raymond. 2005. Does School Accountability Lead to Improved Student Performance? Journal of Policy Analysis and Management 24(2): 297–327.

[40] Neal, Derek, and Diane Whitmore Schanzenbach. 2010. Left Behind by Design: Proficiency Counts and Test-based Accountability. Review of Economics and Statistics 92(2): 263–283; Rouse, Cecilia Elena, Jane Hannaway, Dan Goldhaber, and David Figlio. 2013. Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure. American Economic Journal: Economic Policy 5(2): 251–281.

[41] Koretz, Daniel, and Sheila I. Barron. 1998. The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS). RAND.

[42] Jacob, Brian 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.

[43] Rouse et al. (2013).

[44] Dee and Jacob (2011).

[45] Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. Journal of Political Economy 119(1): 39–77; Mbiti, Isaac, Karthik Muralidharan, Mauricio Romero, Youdi Schipper, Constantine Manda, and Rakesh Rajani. 2019. Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania. Quarterly Journal of Economics 134(3): 1627–1673.

[46] Andrabi, Tahir, and Christina Brown. 2022. Subjective versus Objective Incentives and Teacher Productivity (RISE Working Paper Series No. 22/092). RISE.

[47] Neal and Schanzenbach (2010).

[48] Macartney, Hugh, Robert McMillan, and Uros Petronijevic. 2021. A Quantitative Framework for Analyzing the Distributional Effects of Incentive Schemes (NBER Working Paper No. w28816). National Bureau of Economic Research.

[49] Reback, Randall. 2008. Teaching to the Rating: School Accountability and the Distribution of Student Achievement. Journal of Public Economics 92(5–6): 1394-1415.

[50] Burgess, Simon, Carol Propper, Helen Slater, and Deborah Wilson. 2005. Who Wins and Who Loses from School Accountability? The Distribution of Educational Gain in English Secondary Schools (CMPO Discussion Paper No. 5248). Centre for Market and Public Organisation, University of Bristol.

[51] Deming, David J., Sarah Cohodes, Jennifer Jennings, and Christopher Jencks. 2016. School Accountability, Postsecondary Attainment, and Earnings. Review of Economics and Statistics 98(5): 848–862.

[52] Lavy, Victor. 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.

[53] Campbell, Donald T. 1979. Assessing the Impact of Planned Social Change. Evaluation and Program Planning 2(1): 67–90.

[54] Jacob, Brian A., and Steven D. Levitt. 2003. Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating. Quarterly Journal of Economics 118(3): 843–877.

[55] Cullen, Julie Berry, and Randall Reback. 2006. Tinkering toward Accolades: School Gaming under a Performance Accountability System. In Improving School Accountability. Edited by Timothy J. Gronberg and Dennis W. Jansen. Emerald Group Publishing Limited.

[56] Figlio, David N. 2006. Testing, Crime and Punishment. Journal of Public Economics 90(4–5): 837–851.

[57] Jacob (2005); Cullen and Reback (2006); Figlio, David N., and Lawrence S. Getzler. 2006. Accountability, Ability and Disability: Gaming the System? In Improving School Accountability. Edited by Timothy J. Gronberg and Dennis W. Jansen. Emerald Group Publishing Limited; Deming et al. (2016).

[58] Koretz, Daniel. 2008. Measuring Up: What Educational Testing Really Tells Us. Harvard University Press; Lazear, Edward P. 2006. Speeding, Terrorism, and Teaching to the Test. Quarterly Journal of Economics 121(3): 1029–1061.

[59] Sojourner, Mykerezi, and West (2014).

[60] Mbiti et al. (2019).

[61] Vigdor (2009).

[62] Koretz and Barron (1998); Klein, Stephen P., Laura S. Hamilton, Daniel F. McCaffrey, and Brian M. Stecher. 2000. What Do Test Scores in Texas Tell Us? RAND.

[63] Dee and Jacob (2011).

[64] Jacob (2005).

[65] Springer et al. (2010); Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 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.

[66] Muralidharan and Sundararaman (2011); Andrabi and Brown (2022).

[67] Taylor (2023).

[68] Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. Teacher Incentives. American Economic Journal: Applied Economics 2(3): 205–227.

[69] Staiger, Douglas O., and Jonah E. Rockoff. 2010. Searching for Effective Teachers with Imperfect Information. Journal of Economic Perspectives 24(3): 97–118.

[70] Rothstein, Jesse. 2015. Teacher Quality Policy When Supply Matters. American Economic Review 105(1): 100–130.

[71] Brunner, Eric, Joshua M. Cowen, Katharine O. Strunk, and Steven Drake. 2019. Teacher Labor Market Responses to Statewide Reform: Evidence from Michigan. Educational Evaluation and Policy Analysis 41(4): 403–425.

[72] Ng, Kelvin. 2024. The Effects of Teacher Tenure on Productivity and Selection. Economics of Education Review 101: 102558.

[73] Loeb, Susanna, Luke C. Miller, and James Wyckoff. 2015. Performance Screens for School Improvement: The Case of Teacher Tenure Reform in New York City. Educational Researcher 44(4): 199–212; Dinerstein, Michael, and Isaac M. Opper. 2022. Screening with Multitasking (NBER Working Paper No. 30310). National Bureau of Economic Research.

[74] Taylor (2024).

[75] Adnot et al. (2017).

[76] Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2012. Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. American Economic Review 102(7): 3184–3213.

[77] Bates, Michael. 2020. Public and Private Employer Learning: Evidence from the Adoption of Teacher Value Added. Journal of Labor Economics 38(2): 375–420.

[78] Sartain, Lauren, and Matthew P. Steinberg. 2016. Teachers' Labor Market Responses to Performance Evaluation Reform: Experimental Evidence from Chicago Public Schools. Journal of Human Resources 51(3): 615–655.

[79] Steinberg and Sartain (2015).

[80] Cullen, Julie Berry, Cory Koedel, and Eric Parsons. 2021. The Compositional Effect of Rigorous Teacher Evaluation on Workforce Quality. Education Finance and Policy 16(1): 7–41; Dee and Wyckoff (2015); Adnot et al. (2017).

[81] Leaver, Claire, Owen Ozier, Paul Serneels, and Andrew Zeitlin. 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.

[82] Brown, Christina, and Tahir Andrabi. 2023. Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools (RISE Working Paper Series No. 23/123). RISE.

[83] Taylor and Tyler (2012).

[84] Briole, Simon, and Eric Maurin. 2022. There's Always Room for Improvement: The Persistent Benefits of a Large-Scale Teacher Evaluation System? Journal of Human Resources.

[85] Taylor (2024).

[86] Steinberg and Sartain (2015); Burgess, Simon, Shenila Rawal, and Eric S. Taylor. 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.

[87] Garet et al. (2017).

[88] Taylor (2024).

Suggested citation

Eric S. Taylor (2025). "Teacher Evaluation," in Live Handbook Live Handbook of Education Policy Research, in Douglas Harris (ed.), Association for Education Finance and Policy, viewed MM/DD/YYYY, <WEB LINK>.