EdWorkingPaper No. 22-686

Employee evaluation and skill investments: Evidence from public school teachers

Eric S. Taylor
Harvard University and NBER

When employees expect evaluation and performance incentives will continue (or begin) in the future, the potential future rewards create an incentive to invest in relevant skills today. Because skills benefit job performance, the effects of evaluation can persist after the rewards end or even anticipate the start of rewards. I provide empirical evidence of these dynamics from a quasi-experiment in Tennessee schools. New performance measures improve teachers' value-added contributions to student achievement. But improvements are twice as large when the teacher also expects future rewards linked to future scores. Value-added remains at the now higher level after performance incentives end.

VERSION: May 2024

Suggested citation: Taylor, Eric S.. (2024). Employee evaluation and skill investments: Evidence from public school teachers. (EdWorkingPaper: 22-686). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/h36d-0j56

Employee evaluation and skill investments: Evidence from public school teachers[†]

Eric S. Taylor
Harvard University and NBER

May 2024

When employees expect evaluation and performance incentives will continue (or begin) in the future, the potential future rewards create an incentive to invest in relevant skills today. Because skills benefit job performance, the effects of evaluation can persist after the rewards end or even anticipate the start of rewards. I provide empirical evidence of these dynamics from a quasi-experiment in Tennessee schools. New performance measures improve teachers' value-added contributions to student achievement. But improvements are twice as large when the teacher also expects future rewards linked to future scores. Value-added remains at the now higher level after performance incentives end.

JEL No. I21, J24, J45, M5

[†] Taylor: eric_taylor@gse.harvard.edu, Gutman 469, 6 Appian Way, Cambridge, MA, 02138. Special thanks to the Tennessee Department of Education and Tennessee Education Research Alliance for access to the data. Thanks also to those who provided helpful feedback on earlier drafts: Hedvig Horváth, Kirabo Jackson, Dick Murnane, Amanda Pallais, Paul Peterson, Jan Sauermann, Jim Wyckoff, and anonymous reviewers; as well as seminar participants at AEFP, Arizona State, Brown, CESifo, FEW, Harvard, IFAU/Uppsala, and the Northeast Economics of Education Workshop. The Spencer Foundation provided generous financial support.

Job performance measures and linked incentives are familiar features of the workplace. The motivation is also familiar. Performance incentives tie an employee's compensation to their employer's success, thus inducing more effort at work, better allocation of effort across tasks, or self-selection out of the job. In the literature this agency-theory view of employee evaluation and incentives dates to Holmström (1979) among others, and today we have empirical tests from many occupations and sectors. However, that large empirical literature largely ignores how performance incentives might affect the employee's skill investments.[1]

When an employee expects evaluation and performance incentives will continue into the future (or will begin in the future), the potential future rewards create an incentive to invest in relevant skills today. In many ways this is just a special case of the familiar human capital investment models which began with Becker (1962) and others, combined with agency theory features. Employer-provided incentives usually come to an end, at some point in time, but the employee retains the skills gained. Thus, the effects of incentives on job performance can persist after the incentives end because skills persist. Skill investments, and thus effects on performance, can also anticipate the start of incentives. Additionally, new performance measures themselves, separate from any linked rewards, can reduce the employee's cost of investing in skills.

In this paper I present evidence of anticipation, persistence, and other effects on job performance. I study public school teachers in Tennessee. The state of Tennessee adopted a new teacher evaluation strategy in 2012. The policy required new performance measures and performance incentives (tenure) linked to those measures.[2]

[1] For general reviews see Oyer and Schaefer (2011) and Gibbons and Roberts (2013). For teachers specifically see Neal (2011) and Taylor (2023). The theoretical literature on incentives and skill investments is also comparatively small; see Prendergast (1993), Gibbons (1998), Cisternas (2018).

[2] Following the simplifying convention, I refer to school years by the spring number. Thus the 2011-12 school year is "2012".

I use a difference-in-differences strategy to estimate various effects of incentives and evaluation measures on teacher performance. The first difference is the change in value-added between year $(e - 1)$ and e. The time dimension, e, is years of employment, with $e = 1$ for teachers in their first year of teaching. "Value-added" is shorthand for a teacher's contributions to student achievement test scores. The second difference is between groups of teachers. The treated group experienced some new treatment in year e—announcement of new tenure rules, new performance measures, the start and end of performance incentives. The comparison group is teachers who completed their eth year of teaching before Tennessee's evaluation reforms in 2012. Causal claims require a parallel trends style assumption: Absent the evaluation reforms, teacher value-added would have improved with experience, from $(e - 1)$ to e, at the same rate observed prior to 2012.

Tennessee's evaluation reforms improved teacher performance. And the pattern of effects is consistent with teachers making investments in their own skills in response to the new incentives and evaluation measures. The paper details five main results.

First, teacher performance began improving before any rewards or consequences were linked to a teacher's measured performance. Tennessee adopted and announced new tenure rules just before the start of the 2012 school year. To earn tenure, a teacher must score above a cutoff (empirically about the 33rd percentile of performance scores) in both her fourth and fifth year of employment. Thus, first-, second-, and third-year teachers knew their future performance would determine tenure, but not their current performance.

Anticipation of future incentives increased teacher value-added by 0.023σ or more (σ = student test-score standard deviations, for this estimate e = 2–3). An improvement of 0.02σ is roughly 10–20 percent of the between-teacher standard deviation in value-added performance, or roughly $1,000 per student in net present earnings.

Second, new performance measures boosted value-added as well, separate from incentives. Tennessee began using new measures in 2012, including ratings of teaching skills and practices in classroom observations. All teachers were scored, but the new tenure rules (incentives) did not apply to teachers who had already earned tenure before 2012. The new measures alone increased value-added by 0.024σ at their introduction, among teachers who were already tenured but still early in their career (*e* = 4–7).

The new measures also likely benefited pre-tenure teachers too. The combined effect of the two treatments—anticipating future incentives and new measures—was 0.047σ for pre-tenure teachers (*e* = 2–3). The earlier 0.023σ anticipation effect estimate is actually a triple-difference estimate: 0.047σ minus the effect of new measures estimated with tenured teachers, 0.024σ.

Third, when incentives formally began any additional performance improvements were small at best. A teacher's explicit performance incentives begin in her fourth year. Year 4 evaluation scores count for tenure, but year 3 scores do not. I estimate a 0.013σ effect in year 4, but that estimate is relatively imprecise (standard error 0.009), and thus there may well be zero additional effect when incentives begin.

Fourth, teacher performance persisted at higher levels after performance incentives ended. Two-thirds of teachers successfully earn tenure on time, at the end of year 5. In year 6, newly-tenured teachers no longer have rewards or consequences linked to their evaluation scores. The effect of prior tenure incentives on value-added in year 6—the effect after those incentives had ended—is 0.025σ (standard error 0.014).

If teachers simply boosted their effort when scores counted for earning tenure, we would predict a decline in performance after year 5, as effort returns to its unincentivized level. Contrary to that prediction, I can rule out declines in value-added larger than -0.002σ, which is only 1–2 percent of the between-teacher standard deviation in performance. In other words, the gains accumulated before earning tenure (reflected in year 5 value-added) persisted after earning tenure (year 6 value-added).

Fifth, the new performance incentives had little to no effect on teachers quitting or changing jobs. Teachers with lower (higher) value-added were more likely to leave (stay in) their jobs, but those patterns did not change under the new tenure incentives. (Self-)selection is a common motivation for employee evaluation; for example, Staiger and Rockoff (2010) and Rothstein (2015) address selection and teacher tenure rules specifically.[3] However, selection (attrition) effects cannot explain the teacher performance improvements in Tennessee.

In summary, teacher performance began improving before any scores counted for earning tenure, and performance remained at the new higher level after teachers earned tenure. These anticipatory and persistent effects are consistent with skill improvements caused by the evaluation program and its performance incentives.

Moreover, ignoring potential anticipatory and persistent effects is consequential to estimating the benefits of the program. Consider a naïve difference-in-differences design: years e = 1–3 are the pre period and e = 4–5 the post period, teachers subject to the new tenue rules are the treated group. That naïve estimate is 0.015σ (standard error 0.007), substantially short of the total effect of the tenure performance incentives.

This paper contributes most directly to the literature on how employee evaluation affects the performance of teachers. Teachers do respond to performance incentives; distinct empirical examples include Neal and Schanzenbach (2010),

[3] Unlike the tenure rules considered by Staiger and Rockoff (2010) and Rothstein (2015), in Tennessee teachers who failed to earn tenure were allowed to continue working (Section 2).

Muralidharan and Sundararaman (2011), Duflo, Hanna, and Ryan (2012), Dee and Wyckoff (2015), Deming et al. (2016), Lavy (2009, 2020), Mbiti et al. (2019), Leaver et al. (2021), Aucejo, Romano, and Taylor (2022), among many others. In contemporaneous work, Dinerstein and Opper (2023) and Ng (2022) study teachers' response to performance-based tenure rules specifically. The empirical literature spans a variety of performance measures and incentives, including monetary bonuses and dismissal threats (see reviews by Neal 2011 and Taylor 2023). However, as with the literature from other sectors and occupations (see reviews by Oyer and Schaefer 2011 and Gibbons and Roberts 2013), the empirical tests focus on the conventional prediction: changes in performance should coincide temporally with changes in incentives. This paper contributes empirical evidence of anticipation effects and persistent effects.

First, this paper's empirical evidence of anticipation effects on individual job performance is, to my knowledge, a novel contribution. The potential for anticipation effects suggests existing estimates may understate the total effect of performance measures and linked incentives.

Second, the paper contributes empirical evidence of persistent effects. Tests of persistence remain scarce. Two prior papers, Taylor and Tyler (2012) and Briole and Maurin (in-press), also document improvements in job performance which persist long after performance measures and linked incentives end.[4] However, in those papers the treatment is a bundle of both measures and incentives. Leaving open the possibility that good performance measurement alone is sufficient to generate skill investments, by reducing the costs of skill investments for otherwise motivated agents. This paper's setting includes identifying variation to unbundle

[4] Griffith and Neely (2009) find similar persistent effects, though noisily estimated, among employees in retail sales.

measures and incentives. I document persistent effects of performance incentives, separate from the effects of measures.[5]

Finally, teacher performance evaluation has become a central theme of education policy in the United States. The policy motivation begins by pointing out the large differences between teachers in how much their students' learn during a school year, and that those differences carry into future outcomes in college and the labor market.[6] From there some proposals focus on stronger selection of teachers based on observed performance (e.g., Gordon, Kane, and Staiger 2006, Hanushek 2011), while others emphasize the potential for feedback to benefit skill development (e.g., Darling-Hammond 2015). Selection proposals have been much more carefully considered in the economics literature (Staiger and Rockoff 2010, Rothstein 2015, Dinerstein and Opper 2023). Teacher pay-for-performance schemes have had little success in practice, at least in the United States (for reviews see Neal 2011, Taylor 2023). The policy debates about teacher evaluation are wide ranging. However, missing from those debates is the potential for performance rewards, like tenure, to create incentives for skill investments.

Section 1 briefly summarizes the theoretical ideas which motivate the empirical analysis. Sections 2 and 3 detail the empirical setting and identification strategy, respectively. Results on anticipation effects are described in Section 4, effects when incentives begin in Section 5, and effects after incentive end in Section 6. Section 7 concludes.

[5] A third contribution is evidence that performance measures alone can generate improvements teacher performance. See the second of five main findings summarized in the introduction. Rockoff et al. (2012) and Burgess, Rawal, and Taylor (2021) are also examples where teacher performance improved without formal incentives attached to evaluation scores.

[6] The large literature on "teacher value added" includes seminal papers by Kane and Staiger (2008), Rothstein (2010), Chetty, Friedman, and Rockoff (2014a,b), and Jackson (2018). For reviews see Hanushek and Rivkin (2010), Jackson, Rockoff, and Staiger (2014), and Bacher-Hicks and Koedel (2023).

1. Agency Models and Skill Investments

In this section I briefly describe some familiar features of agency models and human capital investment models—and the interaction of the two—to motivate this paper's empirical analysis. Agency models have a long history in the study of employee evaluation and performance incentives, dating to Holmström (1979) and Holmström and Milgrom (1991) among others. More recent examples include Lazear (2000), Baker (2002), and, specific to teachers, Barlevy and Neal (2012).[7] Examples of human capital models often focus on investments early in life, before entering the workforce, but individuals' skill investments while employed are central to the analysis in Becker (1962), Mincer (2962), Ben-Porath (1967), and others. While both frameworks are familiar, combining the two provides useful insights.

1.1 Performance Incentives and Skill Investments

Start with some basic components from agency theory. Employers and employees make simultaneous choices. The employee chooses effort at work. Her optimal choice balances her own marginal return from that effort against the increasing marginal cost of effort. Effort, combined with skill, produces the employee's contribution to the employer's objectives. The employer chooses how compensation will depend on that contribution. The employer's options include explicit performance incentives—bonuses or other rewards—linked to measured job performance. Well-designed rewards will increase effort—effort chosen by the employee—by increasing the employee's own marginal return from that effort.[8]

[7] The discussion in this section focuses on the skill investments and performance of incumbent employees, matching the paper's empirical setting. Performance incentives may also affect (self-)selection of employees, into or out of a particular employer, thus changing average performance through composition changes. A well-known example is Lazear (2000). For theoretical discussions of incentives and selection see Holmström and Milgrom (1987), Levin (2003), and others. For (self-)selection of teachers see Staiger and Rockoff (2010), Rothstein (2015), Leaver et al. (2021), Brown and Andrabi (2023), and Dinerstein and Opper (2023).

[8] See Section 2.1 for ways in which compensation can depend on performance without explicit performance incentives. Also, "increase effort" is shorthand for a change in effort which improves

At this point, the prediction is straightforward: Effort increases when new performance incentives begin. Effort decreases when those incentives end. Thus, any effect of incentives on job performance will begin and end the same way. Only performance which is explicitly linked to rewards will be affected by those rewards. That conventional prediction focuses on effort.

But performance incentives can also affect skills. Linking rewards to performance creates an incentive for the employee to invest in improving her own skills (human capital).

Consider a basic human capital investment model, without any performance incentives. An employee chooses how much effort (and other resources) to invest in improving her skills. Her optimal choice balances her own marginal return from the skills gained against the marginal cost. The return to skills, at least at work, is improved job performance and compensation.

If the employer introduces a performance incentive program, the new rewards increase the return on skill investments. The human capital investment model is especially apt when the employee expects repeated evaluation and rewards over time into the future. An employee's performance depends on both her effort and her skills. Thus, incentives can affect job performance because the employee gained new skills, not just because she increased her effort in current production.

The skill investment mechanism raises two important possibilities: anticipation effects and persistent effects. First, improvements in employee performance can persist even after performance incentives end because skills persist. When performance is no longer linked to rewards, the employee's incentive to increase her current effort goes away; that conventional prediction remains true. However, the persistent effects arise not from current effort but from past effort invested in skills. Because skills persist over time (to some degree), an incentive

the employer's objective. That could be an increase in total effort, or a reallocation of effort across tasks.

program's past positive effects on skills can increase future performance even after the performance incentives end.

Performance improvements are unlikely to persist fully or forever. One, skills depreciate over time. As that depreciation accumulates, evaluation effects will weaken.[9] Two, the employee should reoptimize her effort choice. Improved skills make the employee's effort more productive. Compared to the counterfactual where the performance incentive program never occurred, the employee can choose to give less effort and still have higher performance. But that new performance level would likely be lower than the counterfactual where the incentive program continued.

The second possibility: A new evaluation program can improve employee performance before its performance incentives begin, because employee skill investments can anticipate the future rewards. Better skills will improve (future) performance directly. Also, skills and effort are complements. Investing in skills makes (future) increases in effort more productive.

Such anticipation effects require some amount of time lag between when the employee is aware of the future incentives and when performance actually determines rewards. That lag exists for the public-school teachers I study in this paper. As detailed in Section 2, teachers earn tenure if they score above a performance cutoff in their fourth and fifth year of employment, but teachers are aware of this tenure incentive in the years before their performance scores count. Earning tenure increases job security and thus increases the present value of a teacher's expected future earnings.

Performance could initially fall—a negative anticipation effect in the short run. Gaining new skills requires effort (and perhaps other resources). Thus, effort

[9] Dinerstein, Megalokonomou, and Yannelis (2022) 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.

devoted to skill building may crowd out effort devoted to current production tasks.[10]

Anticipation and persistence effects are important considerations for an employer calculating the costs and benefits of a performance incentive program. Imagine we ignore these (potential) effects, assuming that only performance which is explicitly linked to rewards will be affected by those rewards. A difference-in-differences estimate of the effect on performance, for example, will understate the true benefits.

1.2 Performance Measures and Skill Investments

The discussion so far has focused on how performance incentives can increase the employee's return on investments in her skills. But performance measures can also reduce the cost of skill investments, by reducing the effort required. First, performance measures create new information: feedback about an individual's current performance, advice on how to improve, comparisons to coworkers' performance. The cost of creating that new information is borne largely by the employer. Absent an evaluation program, the employee is left to her own self-assessment and data gathering. Second, the new information can make skill investments more efficient by directing the employee's effort toward specific skills.

A reduction in costs could generate new skill investments even if there are no new performance incentives. Recall that the employee's optimal choice balances her own marginal return from the skills gained against the, now lower, marginal cost. The marginal return is not constant, even without explicit incentives linked to measured performance. For example, many employees derive some intrinsic

[10] That crowd out (or tradeoff) is not inevitable. The employee can increase total effort, and such an increase is plausibly motivated by the benefits of higher performance even without explicit performance incentives, for example, career concerns and intrinsic rewards. Moreover, for some types of skills there may be relatively little tradeoff. Some skills improve through "learning by doing," that is, an employee can become more efficient at completing some tasks simply through repeating the task over and over in the normal course of work.

rewards from their contributions at work, and intrinsic rewards may be especially relevant for public school teachers (Dixit 2002).[11] Performance is also relevant to career concerns, and thus expected future earnings (Fama 1980, Holmström 1999, Lazear 2000, Lazear and Oyer 2013).

2. Setting and Data

I study public school teachers in Tennessee. In the 2012 school year (synonymously, 2011-12) Tennessee began a new performance evaluation program for teachers. As I detail in this section, the new program included both new performance measures and new incentives attached to those measures. I use data from 2008-2015, four years before and after the start of the new program in 2012.[12]

This paper focuses on a subset of Tennessee's teachers defined by two criteria. First, teachers who teach math or English language arts (ELA) or both to students in grades 4-8. These are the subjects and grades where students are tested annually, all taking the same state administered test, and those testing details are important for identifying a teacher's contribution to student achievement. Second, teachers who are in the early years of their career, specifically in years 1-7. This constraint is primarily motivated by identification, as I describe in Section 3. But this early-career period is also when the evaluation program's incentives are most salient, as I describe shortly.

Table 1 describes the teachers and their students. My estimation sample in column 1 includes over 11,000 teachers and 720,000 students. The teachers are observably similar to others in the state, except, by construction, they are earlier in

[11] Strong intrinsic motivations do not rule out a response to extrinsic rewards, of the kind discussed in Section 1.1. The cost of effort is (likely) increasing and convex, thus all employees will dislike work and prefer leisure at some margin. New extrinsic rewards can shift that margin further out.

[12] Additional details on the setting and data are provided in Appendix Section A2. The policy details in this section are drawn from Tennessee Code § 49-5-504, State Board Rule 0520-02-01, and State Board Policy 5.201, see https://team-tn.org/statute-and-policy/, as well as the summaries in Tennessee Department of Education (2014) and Hunter (2018).

their careers. The students they teach are also similar to other grade 4-8 students in Tennessee. All data used in this paper are administrative data provided by the Tennessee Department of Education through the Tennessee Education Research Alliance at Vanderbilt University.

2.1 New Performance Measures

Tennessee's current teacher evaluation program began at the start of the 2012 school year, just over a year after the state won a federal Race to the Top grant to support the new program. While all public-school teachers were evaluated, the description of measures and incentives here applies to grade 4-8 math and ELA teachers, during the years 2012-2015.[13]

Each teacher's evaluation includes three performance measures: a classroom observation rating, a value-added score, and an additional student test score measure selected by the teacher. Broadly speaking, the classroom observation rating measures inputs, and the student test-score components measure outputs. All three measures make use of a 5-point expectations scale: (1) "significantly below expectations," (2) "below expectations," (3) "at expectations," (4) "above expectations," and (5) "significantly above expectations."

2.1.1 Classroom Observation Scores

Tennessee's new classroom observations measure a teacher's performance of several teaching tasks. The tasks include things like managing student behavior, use of assessment, questioning, and lesson structure and pacing. The school principal (or other school administrator) visits a teacher's class and scores each task separately. Possible scores are the five integer expectations-scale scores. Scoring is guided by a rubric which describes specific teacher behaviors and decisions that must be observed to warrant a given score. Figure 1 shows an example of one task "Questioning" from the rubric. A teacher is scored 1–3 times on each task every

[13] For a thorough description covering all teachers and all years see Hunter (2018).

year, depending on experience and prior performance. Then the task-specific scores are averaged for the final observation rating. Additionally, after each visit the observer provides feedback on how the teacher can improve.[14]

Observation scores do vary. The most common ratings are "at expectations" (3) and "above expectations" (4), each accounting for one-third of task-level scores. The top score (5) is given 20 percent of the time, but low scores are rare (see score histograms in Appendix Figure A1). In other words, the scores do show leniency bias—as is common in employee evaluations across sectors and occupations—but less leniency bias than is often suggested in policy discussions of teacher evaluation (Weisberg et al. 2009, New York Times 2013, Kraft and Gilmour 2017).

Prior to 2012, classroom observation measures were more limited in scope and frequency. During a teacher's first three years of work, her school principal would observe and score her 2-3 times per year, a frequency similar to the new program. However, after year 3 the next observation and scoring would not occur until year 8 for the typical teacher; the state only required evaluation every five years after the probationary period. The pre-2012 process also used a rubric which covered several items (or teaching tasks), and for each item described three levels of performance. These basic features of the rubric were similar to the new rubric, but the types and specificity of tasks covered were different. For example, Figure 1 shows the new rubric for "Questioning." Contrast the level of specificity in Figure 1 with the pre-2012 rubric which for the top score simply says: "Activities, including higher order questioning, are used to develop higher order thinking processes." Moreover, in the pre-2012 rubric questioning is grouped with several

[14] This paragraph describes details of the TEAM system which applies to more than 80 percent of teachers in Tennessee. And the results in this paper are robust to limiting the analysis sample to just TEAM school districts. Details on the other systems are provided in Appendix Section A2.

other tasks on lesson pacing, communication, etc. into one single scored item for "Teaching Strategies."[15]

2.1.2 Teacher Value-Added Scores

Each teacher's evaluation also includes a "value-added score," which measures the teacher's contribution to her students' test score growth. Tennessee's value-added scores are estimated by the SAS Institute and known locally as TVAAS scores.[16] The TVAAS approach is distinctive, but conceptually similar to more-familiar value-added estimation methods (compare SAS Institute 2021 to Jackson, Rockoff, and Stagier 2014 or Bacher-Hicks and Koedel 2023). When describing TVAAS to teachers, Tennessee emphasizes the growth characteristic and that students are compared to peers who scored similarly in prior years. TVAAS scores are reported to teachers in the 5-point expectations scale, and it is often referred to as the "student growth score."

Tennessee principals and teachers have had access to TVAAS reports with teacher value added scores since the early 1990s, long before the new 2012 program. However, prior to the new evaluation rules in 2012, the TVAAS scores were not used for personnel decisions, at least not formally or explicitly.[17]

2.1.3 Achievement Score

The third performance measure is known as the "student achievement score." This measure is also based on student test scores, but typically focuses on the level of student achievement as opposed to growth. Each teacher defines this measure for herself, in collaboration with her school principal. Together they, first, choose a student assessment from a state-approved list. That list includes the state-administered tests and several commercially available assessments. Then, second,

[15] A side-by-side comparison of the two rubrics is provided in Appendix B.

[16] What is now SAS EVAAS began with William Sanders and colleagues' work in Tennessee in the 1990s (Sanders and Horn 1998).

[17] While this paper estimates effects on teacher value-added, my estimation strategy does not make use of the state's TVAAS score data or the "Achievement Score" data as outcome variables.

they set the criteria that will map from student scores onto the 5-point expectations scale. For example, a 7th grade math teacher's 1-5 rating might be determined by the percent of students who pass the 7th grade math test in her school (where "pass" is synonymous with scoring "proficient" or higher). Alternatively, it might be the pass rate for just her class or for all grade-levels in the school. In my estimation sample, 45 percent of teachers take an option like this example, where the 1-5 rating is determined by pass rates on the state tests. For another 40 percent the teacher's rating is determined by her school's TVAAS score. The remaining teachers choose some other commercial assessment.

These achievement scores vary much less than the other evaluation measures. Nearly two-thirds of teachers receive the top score of (5) "significantly above expectations." But the low scores of (1) and (2) are somewhat more common than they are for observation scores (Appendix Figure A2).

2.1.4 Final LOE Score

At the end of the school year, the three performance measures are combined to determine the teacher's "Level of Effectiveness" (LOE) score. First the three measures are averaged together with weights 0.50 for observation, 0.35 for value-added, and 0.15 for achievement. Then that average is discretized into the 5-point expectations scale using pre-determined cut points.[18] Figure 2 is a histogram of LOE scores for the teachers in my analysis sample; the solid line bars are all teachers, and the dashed line bars are teachers in the first five years of teaching.

2.2 New Performance Incentives

Along with the new performance measures in 2012, Tennessee also adopted new rules linking teacher tenure to those measures. Beginning in 2012, LOE scores

[18] This description of LOE calculation and weights here applies to teachers with individual TVAAS scores, which includes this study's sample of grade 4-8 math and ELA teachers. Additional details of LOE scoring are provided in Appendix Section A2.

determine who earns tenure. But the new rules did not apply to teachers who had earned tenure before July 2011.

Under the new rules, a teacher is first eligible for tenure after teaching for five school years. To earn tenure the teacher's annual LOE score must be "above expectations" (4) or "significantly above expectations" (5) in both year 4 and year 5. Teachers who miss the LOE cutoff can continue on a probationary contract in year 6 and beyond, but earn tenure only after scoring LOE ≥ 4 in two consecutive years (Tennessee Code § 49-5-504).

These new rules are a real constraint on tenure. As shown in Figure 2 top panel, two-thirds of teachers score LOE ≥ 4 in any given year, and that proportion is not larger or smaller for early-career teachers. Over any two consecutive years, 57 percent of teachers score LOE ≥ 4 in both years. But in years four and five 63 percent meet the requirement (dashed bars in bottom panel).

Teachers can also lose tenure under the new rules, though empirically losing tenure is unlikely. Tenure is revoked when a teacher scores "below expectations" (2) or lower in two consecutive school years. In practice, however, teachers rarely lose tenure. Fewer than 5 percent of teachers score LOE ≤ 2 in two consecutive years (Figure 2 bottom panel). A teacher can regain tenure after scoring LOE ≥ 4 in two consecutive years.

Notably, these new tenure rules apply only to new cohorts of teachers—only to teachers who began working in 2010 or later. The new rules do not apply to teachers already tenured before the 2012 school year. Under the old rules, teachers were eligible for tenure after three years. Thus, teachers who began working in the 2009 school year earned tenure at the end of the 2011 school year. Teachers who began working in the 2010 school year were the first cohort subject to the new tenure rules. The 2010 cohort would have earned tenure after 2012 under the old rules, but instead had to wait until 2014 at the earliest. And, recall, the new cohorts also had to meet the new LOE score requirements. The 2010 and 2011 cohorts are distinctive because they began working before the 2012 changes but were nevertheless subject the new tenure rules. Figure 3 summarizes these tenure incentives as a function of cohort and years of employment.[19]

2.3 Summarizing Treatments

One way to summarize the many details in this section is to think in terms of treatments applied in this quasi-experiment. The first treatment is the change in performance measures for all teachers. Starting in 2012 all teachers were scored in classroom observations every year. Before 2012 teachers were scored in years 1-3 but not again until years 8, 13, etc. The new observation program also used a new and improved rubric. For many years prior to 2012, teachers had received informational reports showing their value-added scores. Beginning in 2012, those scores were formally used in teacher performance evaluations.

The second type of treatment is the change in performance incentives attached to the new measures. Only teachers hired in 2010 or later received this treatment. The new incentives began in a teacher's fourth year of employment. Under the post-2012 rules, earning tenure required scoring above a cutoff in both year 4 and year 5. The new incentives ended in a teacher's six year, if they had successfully met the score requirements. By contrast, teachers hired after 2010 had already earned tenure before the 2012 school year began; those already-tenured teachers were treated with the new performance measures, but no rewards or consequences were attached to their scores.

[19] Appendix Section A2 describes pay for performance programs in Tennessee. Beginning in the 2012 school year, 10 percent of Tennessee school districts (14) began paying some teachers based partly on evaluation scores. This pay-for-performance treatment is confounded with the evaluation treatment, however, as demonstrated in Appendix Section A4 the paper's results are robust to excluding pay for performance districts from the estimation sample. Tennessee's pay for performance programs include the well-known POINT experiment in Nashville (Springer et al. 2012) but the POINT treatment teachers represent less than 0.5 percent of my sample.

3. Identification Strategy

I use a difference-in-differences strategy to estimate the effects of various evaluation program features—anticipation of future incentives, the start of incentives, the end of incentives, and the start of new performance measures—on teacher job performance. The first difference is the change in a teacher's value-added performance between her $(e-1)$th and eth years of employment. The second difference is between treated and comparison teachers. Treated teachers experienced a change in some evaluation program feature(s) between year $(e-1)$ and e. Comparison teachers were in their $(e-1)$ and e years in 2011 or earlier school years, before the new evaluation program began in 2012.

Let D_{je} be an indicator variable equal to 1 if teacher j is treated in her eth year of employment. Assume, for a moment, that we observe μ_{je}—teacher j's value-added contribution to student achievement scores in each year e. Then, for a given value of e, my difference-in-differences estimate would be:

$$\hat{\delta}_e = \left[\frac{1}{N_e} \sum_{\substack{j:D_{j,e-1}=0, \\ D_{je}=1}} \left(\mu_{je} - \mu_{j,e-1}\right)\right] - \left[\frac{1}{M_e} \sum_{\substack{j:D_{j,e-1}=0, \\ D_{je}=0}} \left(\mu_{je} - \mu_{j,e-1}\right)\right] \qquad (1)$$

where N_e is the number of teachers in the treated group, and M_e the number of comparison teachers. To estimate the average effect, $\hat{\delta}$, combining two or more $\hat{\delta}_e$, I weight by the number of treated teachers, $\hat{\delta} = \frac{1}{N}\sum_e N_e \hat{\delta}_e$. This strategy is an application of the estimator proposed by de Chaisemartin and D'Haultfœuille (2020).[20]

[20] I set $e = 1$ for j's first year working as a teacher in Tennessee, and then I mechanically increment $e + 1$ with each successive school year. This definition of e is an intent-to-treat approach, which avoids bias from endogenous leaves of absence. While the student test score data begin in 2007, the state's administrative data go back many more years which allows me to identify a teacher's first year in Tennessee with confidence.

Take for example the estimate $\hat{\delta} = 0.047\sigma$ in Table 2 column 1 row 1 (discussed in Section 4). For that estimate, e is a teacher's second or third year of employment, $e \in \{2,3\}$. And $\hat{\delta}$ is the average of $\hat{\delta}_2$ and $\hat{\delta}_3$ weighted by treated sample size. For the treated group, year $(e-1)$ occurred in 2011 and year e occurred in 2012, the first year of the new evaluation program. For the comparison group, both years e and $(e-1)$ occurred in 2011 or earlier school years. In other words, treated teachers were hired in 2010 or 2011, and comparison teachers were hired in 2009 or earlier. When treated, $D_{je} = 1$, teachers were scored using new performance measures, and teachers knew that their scores in future years would determine tenure, but there were no incentives linked to teachers' current scores. When not treated, $D_{je} = 0$, teachers had no scores, no incentives, and no anticipation of incentives.

The treatment effect, $\hat{\delta} = 0.047\sigma$, is growth in performance between the first and second (or second and third) year of a teacher's career. But, importantly, it is additional growth on top of the typical growth between the first and second (or second and third) year. That typical growth is the counterfactual estimate from the comparison group, shown in the furthest right set of brackets in 6.[21]

I use the same difference-in-differences strategy for all the paper's estimated effects on value-added performance. For example, for the estimates in Table 4 panel A, I follow the same 2010 and 2011 hire cohorts into their fourth year of employment—year four is the first year that evaluation scores count for tenure. I apply the same strategy described in equation (1) but with $(e-1) = 3$ and $e = 4$. For the estimates in Table 4 panel B, $(e-1) = 5$ and $e = 6$. A teacher's sixth year is the first year she can be tenured under the new rules.

[21] Returns to experience among early career teachers have been well documented (Rockoff 2004, Papay and Kraft 2015, Bell et al. in-press, and Taylor 2023 for a recent review).

Because teacher value-added, μ_{je}, is not directly observable, I use student-level data to fit a regression-based version of (1). The basic specification is:

$$A_{ist} = \delta D_{je} + \alpha_j + \gamma_e + f\left(A_{is,t-1}\right) + X_{it}\beta + \varepsilon_{ist} \qquad (2)$$

where A_{ist} is the test score for student i taught by teacher $j = j(ist)$ in subject s and school year t.[22] I fit specification (2) repeatedly, once to obtain each $\hat{\delta}_e$. For each $\hat{\delta}_e$ the estimation sample is limited to teachers who are in year e or $(e-1)$ of their teaching career, who are observed in the data in both years e and $(e-1)$, and for whom either $\{D_{j,e-1} = 0, D_{je} = 1\}$ or $\{D_{j,e-1} = 0, D_{je} = 0\}$. These sample constraints reproduce the conditions in equation (1). The α_j and γ_e terms are teacher and year-of-employment fixed effects, respectively, though γ_e is equivalent to a single indicator variable for year e. Notably, a given teacher j can contribute observations to more than one $\hat{\delta}_e$. Thus, I stack the several cases into a simple set of seemingly unrelated regressions, and report cluster-corrected standard errors with teacher clusters across regressions.[23,24]

Other features of (2) are more typical of the literature. Student test scores, A_{ist}, are measured in student standard deviation units. Scores are standardized (mean 0, standard deviation 1) within each grade-by-year-by-subject cell using the statewide distribution. The specification controls for a quadratic in prior test scores, $f\left(A_{is,t-1}\right)$, where the parameters are allowed to vary by grade and subject, and several other student and peer characteristics in the vector X_{it}.[25] This "lagged test

[22] All test scores come from the state's TCAP standardized tests, administered to all students in Tennessee schools, not tests chosen locally by individual teachers or schools.

[23] For clarity, there are no cross-equation restrictions on coefficients, only the cross-equation clusters for the standard errors. Thus, for example, $\hat{\alpha}_j$ are specific to each $\hat{\delta}_e$ as are all other parameters.

[24] Appendix Table A1 reports randomization inference p-values. Hire cohorts—groups of teachers who were first-year teachers in the same school year—are randomly assigned to treatment conditions, in each permutation, to construct the null distribution. Matching the statistical inference reported in Tables 2 and 5, anticipation and persistence estimates are statistically significant at conventional levels with randomization inference.

[25] The vector X_{it} includes indicator variables for (i) female; (ii) black, Hispanic, and other race or ethnicity, with white omitted; (iii) eligible for free or reduced-price lunch; (iv) English language

score" specification is common in the study of teachers and has a strong theoretical motivation (Todd and Wolpin 2007). Perhaps more importantly, (quasi-)experimental tests show that the assignment of students to teachers is plausibly ignorable conditional on prior test scores, and thus it is plausible to assume $E[\varepsilon_{ist}] = E[\varepsilon_{ist}|j]$.[26]

A causal interpretation of the 0.047σ estimate, and the other estimates in this paper, requires a parallel trends style assumption: Absent the new evaluation program, teacher value-added would have improved with experience, from $(e-1)$ to e, at the same rate observed in cohorts prior to the new program. Note that the trends in this case are over years of employment not calendar time. Econometrically this assumption is clear in equation (1). Substantively, rapid performance growth early in the teaching career—the returns to experience—is a first order feature of teacher contributions to student achievement scores (Rockoff 2004, Papay and Kraft 2015, Taylor 2023). Thus, the importance of a counterfactual estimate which includes the counterfactual returns to experience, especially since the useful treatment variation here occurs during the first several years of a teacher's career. Threats to this identifying assumption would be changes over time in the rate of returns to experience. Perhaps, for example, the selection or training of new teachers is improving over time in Tennessee, specifically, in a way that

learner; and (v) special education. The vector X_{it} also includes peer measures: the classroom mean and standard deviation of $A_{is,t-1}$, and classroom mean of (i)-(v). Additionally, approximately 17 percent of the time, a student will have two or more teachers in a given subject and year. In those cases, I duplicate the ist observation for each teacher j, j', etc. and weight each by the proportion of responsibility assigned by the state to the teacher. Given the low proportion of multiple teachers, the results are robust to assigning all students to the teacher with the highest proportion of responsibility.

[26] For (quasi-)experimental tests see Kane and Staiger (2008), Kane et al. (2013), Chetty, Friedman, and Rockoff (2014a), and Bacher-Hicks et al. (2019). For a more skeptical assessment see Rothstein (2010, 2017).

makes the returns to experience steeper or shallower as each cohort begins their career.[27]

Figure 4 provides two pieces of evidence supporting the plausibility of the identifying assumption. The top panel shows a time series of performance for first-year teachers. The y-axis measures average first-year value-added relative to the average experienced teacher. The x-axis is the year hired. There is little evidence that Tennessee's cohorts of newly hired teachers are systematically improving or declining over this period. There is some noise, but we cannot reject a flat trend line, and the differences from year to year are generally less than 0.01σ. The bottom panel summarizes the returns to experience in Tennessee over time. The estimation procedure follows equations (1) and (2) above, except that $\hat{\delta}_e$ is estimated for each school year t, $\hat{\delta}_{et}$, and then I average across e for a given year to get $\hat{\delta}_t$. The y-axis then measures the improvement from $(e-1)$ to e. The solid line is for $e \in \{2,3\}$, the returns to experience between year 1 and 2 or 2 and 3. There is a clear trend break in 2012 when the new evaluation program begins. The dashed line is for $e \in \{4,5,6,7\}$ and discussed later.

Finally, I use a corresponding difference-in-differences strategy to estimate treatment effects on teacher attrition. In this setting, teachers attrit from the estimation sample either by leaving Tennessee public schools entirely, or by switching jobs from a grade and subject where students are tested to some other teaching job. In equation (1), the first difference term $(\mu_{je} - \mu_{j,e-1})$ is replaced with an indicator for attrition: $\Delta T_{je} = 1$ if teacher j was present in the estimation

[27] Different cohorts of teachers began their careers in different calendar years. Changes over time in the outcome measure, student test scores, are also potentially relevant. To control for those changes I standardize test scores (mean 0, standard deviation 1) within each cell defined by test year, grade level, and subject using the statewide distribution in each cell. This standardization will address any secular trends which are unrelated to teacher experience. I cannot control directly for year effects in specification (2) because of the age-period-cohort problem. This approach is common in the literature (see for example Rockoff 2004, Taylor 2023, Bell et al. in-press).

sample in year $(e-1)$ but absent in year e, and $=0$ if teacher j was present in both $(e-1)$ and e.[28] Estimating δ_e for attrition simplifies to regressing ΔT_{je} on an indicator for treatment. Just as before, I stack the several δ_e cases into a set of seemingly unrelated regressions, and report cluster-corrected standard errors.

4. Effects of Future Performance Incentives

4.1 Performance

The performance of pre-tenure teachers began improving in the first year of Tennessee's new evaluation program—improvements consistent with teachers acting in anticipation of the newly-announced but not-yet-in-effect performance incentives. The new program increased average value-added by 0.047σ among teachers anticipating those future incentives (Table 2 column 1 row 1). Both math teachers and language teachers improved (columns 2–3), though math teachers improved twice as much as language teachers.

Two distinctive cohorts provide the empirical opportunity to test for anticipation effects. Teachers who began their teaching jobs in 2010 or 2011 were already working in Tennessee schools when the new program began. Thus, we can measure each teacher's value-added both before and after the new program starts in 2012. Additionally, in 2012 teachers hired in 2010 and 2011 were in their second and third year of employment, respectively. Thus, in 2012 these teachers received scores and feedback using the new performance measures, but there were no rewards or consequences linked to scores yet. Under the new rules, tenure incentives began only in a teacher's fourth year of employment (see Section 2).[29]

[28] $\Delta T_{je}=\left(T_{je}-T_{j,e-1}\right)$. $T_{je}=0$ if j is teaching in a tested grade and subject in year e, and $=1$ if not. $T_{j,e-1}=0$ in all cases because this defines the baseline sample.

[29] Recall from Section 2 that, under the new rules, earning tenure required scoring above a cutoff (about the 33rd percentile of teachers empirically) in both the fourth and fifth year of employment.

For those two distinctive cohorts, Tennessee's new program increased value-added by 0.047σ in the program's first year. That 0.047σ gain is the treatment effect on top of the normal "returns to experience" improvements in value-added we would expect between a teacher's first and second (or second and third) year of teaching. As explained in Section 3, I estimate that counterfactual normal improvement using data from the school years just before the new evaluation began, 2008-2011.

These effects are educationally and economically meaningful in magnitude. The between-teacher standard deviation in value-added—total contribution to student achievement—is typically estimated at 0.10–0.20σ (see reviews in Hanushek and Rivkin 2010, Jackson, Rockoff, and Staiger 2014, Bacher-Hicks and Koedel 2023). Treatment effects in the range of 0.02–0.04σ are then 10 to 40 percent of the standard deviation in teacher performance. Effects of 0.02–0.04σ are also similar to the gain from adding one or two weeks of additional class time to the school year (Fitzpatrick, Grissmer, and Hastedt 2011, Aucejo and Romano 2016, Aucejo et al. 2022). Finally, a back-of-the-envelope application of estimates from Chetty, Friedman, and Rockoff (2014b) suggests that a 0.02–0.04σ gain may be worth $1,000–2,000 per student in net present earnings.

4.2 Selection, Turnover

An alternative explanation for the 0.047σ estimate is dynamic selection. Inexperienced teachers often quit the profession or change jobs. Tennessee's new performance measures and tenure rules might have caused more (or fewer) teachers to quit or change jobs. Further, a teacher's choice to stay or leave might depend even more on her performance than it did before the new evaluation program. For example, a novice teacher who expects her performance will not improve enough to meet the new tenure requirements might quit even before her scores officially count for tenure decisions. But only teachers working in both year $(e-1)$ and e

are included in the estimation sample. Thus, differential attrition could (partly) explain the diff-in-diff estimates in Table 2.

However, the empirical patterns of quits and job changes in Tennessee leave little scope for differential attrition to meaningfully bias the 0.047σ estimate. The attrition relevant to that estimate comes from quits and changes after a teacher's first or second year of employment. In the comparison years, 7.3 percent of teachers attrited by quitting. A further 24.1 percent attrited by changing jobs—in year $(e-1)$ they were teaching 4–8 grade math or language where students are tested annually, but in year e they had switched to a non-tested grade or subject. In total the attrition rate was 31.4 percent in the years before the new evaluation program. When the new performance measures and tenure incentives began, the attrition rate fell slightly to 30.1 percent.

The treatment effect on attrition is –1.3 percentage points, or a 4.3 percent reduction in attrition (Table 3 panel A column 1). This suggests teachers anticipating future incentives were more likely to stay in their jobs, but the difference is not statistically significant (the standard error is 1.4 percentage points).[30] Treated teachers were more likely to quit their teaching jobs (column 3), but this is offset by fewer switching to non-tested grades and subjects.[31]

Finally, perhaps treatment changed the causes of attrition, even if the levels of attrition were unchanged. As mentioned above, a teacher's choice to stay or leave might depend even more on her performance than it did before the new evaluation program. Again, the empirical evidence suggests this hypothesis cannot explain the

[30] Even if there was in fact differential selection, it is unlikely selection alone could explain the 0.047σ effect. Assume there was no treatment effect on performance, thus all of the 0.047σ effect came from retaining more teachers. Teachers who would have left without the new evaluation program, but who's value-added would have increased with or without the new program. The average improvement in value-added among those 4.3 percent extra retained teachers would need to be implausibly large: 1.09σ = 0.047 / 0.043. 1.09σ is roughly 5–10 standard deviations of improvement.

[31] To be precise, quitting here is defined as no longer teaching in Tennessee public schools, which is what I observe in the data. Some may have taken jobs in other states or in private schools.

0.047σ effect. Higher performing teachers are less likely to attrit: In the comparison years, a teacher whose value-added score in year $(e-1)$ is one standard deviation higher is 2.3 percentage points less likely to attrit in year e (Table 3 panel A column 2 row 2). First, however, that relationship does not change once the new evaluation program begins. The 2.3 estimate rises to 3.0, but the difference is far from statistically significant (difference –0.7, standard error 1.1).

Second, the threat to identification here is attrition correlated with value-added growth not levels. Prior studies of the returns to experience in teaching suggest a negative correlation between value-added levels and growth among early-career teachers (Kraft and Papay 2014, Atteberry, Loeb, and Wyckoff 2015). If the new program caused more teachers to stay in their jobs (Table 3 panel A columns 1–2 row 1), then marginal retained teachers likely had higher value-added in their $(e-1)$ year (column 2 row 2–3), and thus we would predict less growth from $(e-1)$ to e. That would imply the 0.047σ estimate is biased too small. Though, to reiterate, I find no statistically significant treatment effects on attrition.

4.3 Anticipating Future Incentives vs. New Performance Measures

Teachers anticipating tenure incentives in future years improved by 0.047σ. However, recall that two different treatments could contribute to that gain: (a) new performance measures and feedback, and (b) future incentives. Would the 0.047σ gain have occurred even without the anticipation of future incentives? For the reasons discussed in Section 1.2, (a) performance measures and feedback can reduce the costs of a teacher's investment in her skills. Reducing those costs could generate skill investments, and thus performance gains, even without a change in the returns. But (b) future incentives, like tenure, increase the returns on a teacher's investment in her skills. Those future returns alone could generate skill investments and performance gains. The conclusion of the paragraphs below is that the 0.047σ gain cannot be explained by new measures alone; half or more of the 0.047σ gain is a response to future incentives.

4.3.1 New Performance Measures

Teacher evaluation in Tennessee provides an opportunity to estimate the effect of (a) new performance measures and feedback alone, without any change in incentives. This estimate uses the same diff-in-diff strategy, and the same comparison group, but a different treated group.

Teachers hired in 2009 or earlier years already had tenure before 2012. When the new program began in 2012, these tenured teachers were scored on the new performance measures and received feedback, just like the pre-tenure teachers. But, as the statute made clear, these already-tenured teachers would never have any incentives or consequences linked to their evaluation scores. To estimate the effect of the (a) new measures treatment on these never-incentive teachers, I apply the diff-in-diff strategy described in Section 3, with $e \in \{4,5,6,7\}$. Again, the comparison group is teachers with the same years of experience teaching, but in years before the 2012 reforms. The dashed line in Figure 4 panel B shows the trend over time for this treated group.

Teachers receiving new performance measures and feedback, but without any anticipation of future incentives, improved by 0.024σ (Table 2 column 1 row 2). Both math and language teachers improved (columns 2–3). That 0.024σ gain is itself an important benefit of the new evaluation program. Prior studies have also found similar effects of measures and feedback in France, England, and elsewhere in the United States (Taylor and Tyler 2012, Papay et al. 2020, Burgess, Rawal, and Taylor 2021, Hanno 2022, Briole and Maurin in-press).

Again, empirically, there is little scope for attrition bias. Table 3 panel B reports the attrition effects relevant to this 0.024σ estimate. The attrition rate fell from 24.3 to 23 percent when the new performance measures began for these tenured teachers. The difference, –1.3 points, is not statistically significant, nor does the relationship between value-added and attrition change significantly (Table 3 panel B columns 1–2).

4.3.2 Tripple-Difference

Compare the performance effects for future-incentive and never-incentive teachers, 0.047σ and 0.024σ, respectively. The value-added gains for teachers anticipating future incentives are twice as large. That twice-as-large effect is evidence of anticipation effects—teachers change their current behavior in anticipation of tenure incentives in future years. There are two ways to reach that conclusion.

First, perhaps the two treatments—new measures and future incentives—have independent effects. Assume there are two treatments with linearly-additive effects. Under that assumption, the effect of anticipating future incentives is 0.023σ = 0.047σ – 0.024σ. Table 2 row 3 reports this triple-difference (diff-in-diff-in-diff) estimate.[32]

Second, alternatively, perhaps the two treatments are complements. In that case, the triple-difference is not an estimate of the incentive anticipation effect *per se*. Nevertheless, using Table 2 row 3 as test statistic, we can reject the null hypothesis: new measures combined with future incentives has the same effect as new measures alone. In short, if the two treatments have independent effects, then the triple-difference is explicit evidence of anticipation effects. If the effects are not independent, then the difference is implicit evidence of anticipation effects.

Attrition bias is unlikely to overturn the conclusion that the triple-difference, 0.023σ, is evidence of teachers anticipating future incentives. First,

[32] Strictly speaking, this triple-difference estimate also requires the assumption that the future incentive treatment effects and new measure treatment effects are independent of years of employment, e, at least over the range of $e \in [2,7]$.

More generally, one alternative hypothesis for the future-incentive vs. never-incentive difference is the following: The two teacher types are colinear with years of experience. Perhaps treatment effects are a decreasing function of experience, either because of how performance changes with experience or how turnover (selection out of teaching) changes with experience. The data are not consistent with this hypothesis. Appendix Table A2 shows effect estimates by years of experience. The relationship is not monotonic. However, the individual estimates, $\hat{\delta}_e$, are noisier than the estimates in Table 2.

there is little to no differential attrition in the triple-difference (Table 3 panel C column 1).

Second, if there is attrition bias in 0.023σ, it is unlikely that bias is positive. In other words, it is unlikely attrition bias leads us to incorrectly conclude that teacher value-added responds to future incentives when in fact it does not. Assume (i) that anticipating future tenure incentives has no effect on value-added performance. Under assumption (i), 0.047σ and 0.024σ are two estimates of the effect of new measures on value-added. The difference between the two could still be attrition bias, with differential attrition caused by the difference in future incentives (even if future incentives do not affect value-added).[33] If there was differential attrition, empirically (ii) never-incentive teachers were more likely to attrit than were future-incentive teachers (see Table 3 panel C row 1 columns 1–2).[34] Given (i) and (ii), positive attrition bias in 0.023σ would occur only if the marginal never-incentive attriters are teachers who would have experienced large positive effects from the new measures if they had stayed in their jobs. That seems unlikely. Improved performance would improve any teacher's career prospects. Further, never-incentive teachers could not be dismissed for poor performance let alone for improved performance.

4.4 Additional Considerations

In Appendix Section A4, I discuss several additional considerations about the magnitude of these effects and their mechanisms. For example, briefly, career concerns can motivate greater current effort even without current incentives (Fama 1980, Holmström 1999, Lazear and Oyer 2013). However, to explain the effect

[33] The reasoning in this paragraph is even clearer if we assume further, in addition to (i), that any effects of new performance measures on value-added are independent of teaching experience, e, at least over the range of $e \in [2,7]$. See the previous paragraph.

[34] In Table 3 panel C row 1 column 1, the triple-difference is zero (standard error 0.017). That zero is the estimate for the average teacher. In column 2, the estimate is –0.031 (standard error 0.013). That is estimate is for teachers with average value-added. Average value-added is zero by construction.

estimates in Table 2, the new evaluation program would need to create new career concerns channels. Also in the appendix, I discuss pay-for-performance programs during this period. The estimates in Table 2 remain essentially unchanged if the estimation sample excludes the small subset of teachers who have the potential to earn bonuses based on their performance measures (Appendix Table A1).

5. Effects When Performance Incentives Begin

The linking of performance measures to explicit performance incentives—earning tenure—begins in a teacher's fourth year. To earn tenure a teacher must score above a cutoff (LOE ≥ 4 or about the 33rd percentile) in both her fourth year and fifth year of employment. Scores from year 3 do not count for tenure (see Section 2). Given that new incentive in year 4, we might expect a meaningful improvement in performance between year 3 and 4.

However, I find little to no improvement in value-added when formal tenure incentives start, for the average teacher. The estimated effect is positive: 0.013σ (Table 4 panel A column 1). For that diff-in-diff estimate, the comparison group is teachers whose third and fourth years of teaching occurred before the new evaluation program began. Thus, the 0.013σ improvement is faster growth in value-added performance above the normal (counterfactual) growth between year 3 and 4.[35] Also, the 0.013σ estimate is unlikely to be biased by differential attrition. The new tenure rules increased attrition by only 0.4 percentage points between year 3 and 4 (standard error 1.3 percentage points, a 1.6 percent increase over the base of

[35] The estimates in Table 4 panel A apply the difference-in-differences strategy described in Section 3 equation (1) with $(e-1)=3$ and $e=4$. The treatment indicator $D_{je}=1$ when the teacher's performance will determine whether she earns tenure or not, and $D_{je}=0$ when there are no formal linked incentives. The treated group, $\{D_{j,e-1}=0, D_{j,e}=1\}$, was scored with the new performance measures in both years: $(e-1)=3$ and $e=4$. The comparison group, $\{D_{j,e-1}=0, D_{je}=0\}$, was not scored in either year.

21.8).[36] However, the 0.013σ estimate is relatively imprecisely estimated (standard error of 0.009) and thus I cannot reject zero effect on value-added.

That null result should not be interpreted as "no effect of performance incentives" in general. That interpretation would require assuming there are no effects of *anticipating* incentives; an assumption contrary to the evidence in Section 4. Any performance gains before year 4 are differenced out in equation (1). In other words, 0.013σ is a naïve estimate of the effects of performance incentives—naïve to anticipation effects.

If we ignore (or assume away) any anticipation effects, then we could use a conventional diff-in-diff design to estimate the effects of Tennessee's new tenure incentives. The first difference is teacher value-added in years e = 4–5 (post period) minus value-added in e = 1–3 (pre period). The second difference is between teachers hired in 2010 or later (treated group) and teachers hired before 2010 (comparison group). That estimate is 0.015σ (standard error 0.007).[37] That conventional estimate fails to detect the total effects of the new tenure incentives on teacher value-added, because those effects are (largely) anticipation effects.

Finally, the small average improvement, 0.013σ, may well mask heterogeneity correlated with job performance. Each teacher knows her own prior evaluation scores. At the start of year 4, some teachers (a) will expect to score above the tenure cutoff even if they make no change. Other teachers (b) will expect to miss the cutoff unless they increase their effort; the start of formal tenure incentives in year 4 is far more salient for group (b). As discussed in Appendix Section A5,

[36] Very similar to the pattern shown in Table 3 column 2, teachers with higher value-added scores in year 3 were less likely to attrit in year 4, but that relationship was not affected by the new tenure rules. Full results, like Table 3 column 2, are provided in Appendix Table A3.

[37] I obtain this more-conventional estimate from a single least-squares regression with the specification in (2). This is a two-way fixed effect specification with teacher fixed effects, α_j, and year of employment fixed effects, γ_e. The estimation sample is all observations in the data where teacher j has $e \in [1,5]$. The key indicator variable $D_{je} = 1$ for teachers hired in 2010 or later, and with $e \in [4,5]$.

the data suggest value-added did improve more in group (b) than in group (a). But there are important limitations to those estimates. Most notably, group (b) teachers were much more likely to quit or change jobs after year 3, creating substantial scope for attrition bias in estimates of effect heterogeneity.

6. Effects After Performance Incentives End

Formal performance incentives end after a teacher earns tenure, but newly-tenured teachers continue to perform at higher levels. Nearly two-thirds of teachers earn tenure after year 5 of employment, and their scores in year 6 no longer count for tenure (see Section 2). We might expect a decline in performance between year 5 and 6, but teacher value-added does not decline (and may improve further). The persistence of higher value-added is consistent with improvement in teacher's skills caused by the new evaluation program's tenure incentives.

To begin, consider all teachers subject to the new tenure rules, both those who earned tenure after year 5 and those who did not. Under the new rules, value-added increased 0.037σ between year 5 and 6 (Table 4 panel B column 1). Recall that the diff-in-diff comparison group here is teachers whose fifth and sixth years of teaching occurred before the new evaluation program began. Thus, the 0.037σ improvement is faster growth in value-added performance above the normal (counterfactual) growth between year 5 and 6.[38] Attrition rates after year 5 are essentially unchanged under the new program (column 2).

[38] The estimates in Table 4 panel B apply the difference-in-differences strategy described in Section 3 equation (1) with $(e-1)=5$ and $e=6$. Conceptually, we could keep the same definition of D_{je} as in Table 4 panel A, with $D_{je}=1$ when the teacher's performance is linked to formal incentives (earning tenure), and $D_{je}=0$ when there are no linked incentives. In Table 4 panel B that would make the treated group $\{D_{j,e-1}=1, D_{je}=0\}$ and the comparison $\{D_{j,e-1}=0, D_{je}=0\}$. However, in Table 4 panel B, I maintain the pattern of earlier tables where the treated group receives the treatment in year e. This choice only affects the sign of the point estimate.

However, that 0.037σ effect is an average of two effects: the effect for those who earned tenure after year 5 (two-thirds), and the effect for those who did not (one-third). Performance incentives continue in year 6 for the latter one-third who remain untenured. To sharply test what happens when incentives end, we need to study the two-thirds who did earn tenure.

6.1 Triple-Difference

We want to know whether performance incentives affect value-added after those incentives end. The causal relationship of interest is: The effect of turning off tenure incentives on teacher value-added. A convincing estimate of that causal effect requires extending the difference-in-differences identification strategy to a triple-difference.

Consider a naïve diff-in-diff estimate: 0.024σ (Table 5 panel A column 1 row 1). To obtain that estimate, I restrict the treated sample to teachers who successfully earned tenure under the new rules—on time, after year 5, by scoring LOE ≥ 4 in years 4 and 5. But I cannot restrict the comparison sample in the same way; the comparison is data from before the 2012 reforms and before LOE scores exist. Thus, the 0.024σ estimate means newly-tenured teachers' value-added grew 0.024σ faster than the average rate of growth, between year 5 and 6, in the time before Tennessee's new evaluation program began.

The naïve diff-in-diff estimate lacks a convincing counterfactual. First, the newly-tenured are selected because they have higher than average performance. It is possible that, even absent the new evaluation program, teachers who were relatively high performing at baseline might improve faster than the average teacher. Second, while tenure incentives ended after year 5, newly-tenured teachers continued to receive the new performance measures and feedback which comparison teachers did not receive. Even if tenure incentives had no effect, the 0.024σ gain could come from measures and feedback. Third, the newly-tenured are less likely to quit or change jobs, compared to the average teacher, potentially because of their higher performance (Table 5 panel A columns 2–3 row 1).

To construct the triple-difference, I use the never-incentive group. Recall that never-incentive teachers were hired before 2010 and earned tenure after year 3 under the old rules. Never-incentive teachers were scored using the new measures, but, by statute, they would never have any formal incentives or consequences linked to those scores.

Now consider the subset of never-incentive teachers who *would have* earned tenure under the new rules. The same naïve diff-in-diff estimate for that subset of teachers is –0.0004σ (Table 5 panel A column 1 row 2). For that estimate, I restrict the treated sample to teachers who *would have* successfully earned tenure under the new rules. I can make this restriction because the never-incentive teachers were still evaluated, and have LOE scores, even though those scores were not linked to tenure decisions.

The triple-difference estimate is 0.025σ (Table 5 panel A column 1 row 3). That is, the naïve estimate for (a) newly-tenured teachers, 0.024σ, minus the naïve estimate for (b) never-incentive but *would have* earned tenure teachers, –0.0004σ.

The triple-difference has two important benefits. First, it differences out the selection bias, described in the preceding paragraphs, which arises when the treated teachers are selected on prior performance (LOE scores).[39] The treatment groups in both (a) and (b) are constructed in the same way. Second, the triple-difference sharpens the contrast in treatment features. Both (a) and (b) are treated by the new performance measures and feedback in years 5 and 6. But only (a) had a change in

[39] Mean reversion is one potential source of bias from selecting on prior performance. I discuss mean reversion further in Appendix Section A7. In short, the triple-difference avoids bias from mean reversion. Additionally, for newly-tenured teachers, mean reversion would make the estimated effect too small, since newly-tenured teachers are selected partly on positive error draws. The potential bias is opposite signed for teachers who fail to earn tenure on time.

performance incentives: from active tenure incentives in year 5 to no incentives in year 6.

There are two ways to interpret the difference (a) minus (b). Interpretation one: Assume the two treatments—(i) performance measures and feedback and (ii) performance incentives—have linearly-separable effects. Then 0.025σ is a triple-difference estimate of the treatment effect of (ii) performance incentives on value-added growth from year 5 to 6. Interpretation two: Assume (i) and (ii) are complements in producing teacher performance. Then the difference is not the treatment effect *per se.* Nevertheless, Table 5 panel A row 3 is still a test statistic for the null hypothesis: (i) measures and (ii) incentives together have the same treatment effect as (ii) measures alone.

6.2 Newly-Tenured Teachers

Teachers continued to perform at higher levels, even after they earned tenure and the formal performance incentives ended. The average newly-tenured teacher's value-added in year 6 continued at her year 5 level. Importantly, that year 5 level was itself higher than it would have been absent Tennessee's evaluation reforms; the year 5 level includes treatment effect gains accumulated over years 1–5 during which teachers were anticipating or experiencing tenure incentives. The persistence of higher value-added performance is consistent with skill growth caused by the new evaluation program's tenure incentives.

Newly-tenured teachers—who had been subject to the new tenure rules during their early career—improved a further 0.025σ between year 5 and 6 (Table 5 panel A column 1 row 3). That 0.025σ gain was caused by tenure incentives, even though tenure incentives ended after year 5. The 0.025σ gain is on top of the counterfactual growth—growth between year 5 and 6 among teachers not subject to the new tenure rules. However, the 0.025σ effect estimate is somewhat noisy and the 95 percent confidence interval includes zero (standard error 0.014). In other words, I cannot rule out the conclusion that, after year 5, these newly-tenured teachers stopped improving faster than normal.

Still, while the newly-tenured may not have improved further, their value-added did not decline. I can rule out declines larger than -0.002σ (95 percent confidence interval), which is roughly 1–2 percent of a teacher standard deviation in value-added performance.

The conventional view of performance incentives would predict a decline in value-added after year 5 when tenure incentives end. Assume that any boost in value-added performance in year 5 (or year 4) was only caused by the conventional mechanism—higher current effort in response to current incentives—and not caused by skill improvements. Under that assumption we would predict a negative treatment effect when incentives turn off. If year 5 value-added was boosted by tenure incentives, then to return to the counterfactual level of performance in year 6 would require a decline in value-added between years 5 and 6. That decline did not occur.[40]

There is little scope for attrition bias in the 0.025σ triple-difference estimate. After year 5, newly-tenured teachers were 1.3 percentage points less likely to attrit than the already-tenured teachers who *would have* earned tenure under the new rules (Table 5 panel A column 2 row 3). That –1.3 difference is far from statistically significant (standard error 2.3 points). Still, perhaps the newly-tenured teachers were, in fact, less likely to attrit. Assume tenure incentives did increase retention, after year 5, but had no effect on value-added. For differential attrition to generate the 0.025σ estimate, the marginal retained teacher would have to be someone who's value-added grew substantially faster, between year 5 and 6, than the inframarginal

[40] This reasoning also requires that the performance of treated teachers—those subject to the new tenure rules—had not fallen below the counterfactual trend in some earlier year before year 4 or 5. The prior results are consistent with the opposite: treated performance was higher than the counterfactual even in those early years of employment.

teacher. I discuss the potential threat from attrition bias further in Appendix Section A6, including possible reasons why the marginal retained teacher might have faster growth.[41]

6.3 Not-Yet-Tenured Teachers

What about the teachers who did not earn tenure? Roughly one-third of teachers did not score high enough to earn tenure on time, at the end of their fifth year of employment (Figure 2). They were allowed to continue working as teachers in year 6 and beyond, but still under a probationary contract, and would only earn tenure after scoring above the cutoff in two consecutive years (see Section 2).

These not-yet-tenured teachers did continue teaching in year 6. More precisely, the new tenure rules did not increase (or decrease) attrition rates. Teachers subject to the new rules, but who did not earn tenure after year 5, were 0.8 percentage points less likely to attrit after year 5 (standard error 4.4 points, Table 5 panel B column 2 row 3). That is 0.8 points less than never-incentive teachers who *would not have* earned tenure under the new rules. Both the not-yet-tenured and would-not-have groups—all performing in the bottom one-third of the teacher distribution—were more likely to attrit compared to the *average* teacher in the years before Tennessee's reforms in 2012 (4.6 and 5.4 points, respectively, Table 5 panel C column 2 rows 1–2), but the tenure incentives did not affect attrition. In that regard, the new tenure rules were not successful as a selection policy, at least not in the short run of year 6.[42]

[41] Table 5 column 2 reports on attrition between year 5 and 6, but attrition before year 5 may also be relevant to interpreting value-added effects in Table 5 column 1. Appendix Section A6 discusses cumulative attrition between year 1 and 6. I find no (statistically significant) treatment effect on cumulative attrition or on the relationship between performance and cumulative attrition.

[42] There is some evidence that not-yet-tenured teachers were more strongly selected on performance than the would-not-have earned tenure teachers, even if attrition rates were similar on average for the two groups. Among not-yet-tenured teachers, the estimated relationship between value-added and attrition tripled in strength from –0.023 to –0.073, but fell very slightly for would-not-have teachers (Appendix Table A4 panel D column 2). However, all of these estimates are noisy.

Teachers who continued teaching presumably wanted to earn tenure, and we might expect an increase in effort and performance in year 6. I find little evidence that value-added improved (or declined) among teachers still subject to the tenure incentives in year 6. The triple-difference point estimate is –0.016σ (Table 5 panel B column 1 row 3), but I cannot rule out declines of –0.07σ or gains of 0.04σ (95 percent confidence interval). These teachers' value-added may still have been higher than it would have been without the new tenure incentives, even if that performance boost was not sufficient to earn them tenure on time.[43]

7. Conclusion

This paper documents teachers' (employees') responses to performance measures and performance incentives in a new evaluation program. Early-career teachers' value-added improved faster because of Tennessee's new tenure incentives. Those gains include improvements in the years before teachers' scores counted for tenure—*anticipation effects.* Value-added improved 0.023σ faster among second- and third-year teachers who knew earning tenure would require high scores in the future, in their fourth and fifth year, but who had no current performance incentives. The gains also include improvements sustained after scores no longer counted for tenure—*persistent effects*. Most teachers earned tenure on time at the end of their fifth year, and thus tenure incentives ended, but their value-added remained higher in the sixth year (and may have improved further).

Ignoring these anticipatory and persistent effects substantially understates the benefits of Tennessee's performance incentives. Consider a research design which assumes incentives only affect performance by raising effort when scores are

[43] The naïve diff-in-diff estimates (Table 5 panel B column 1 rows 1–2) are large but not convincing estimates of the causal effect of interest, for the reasons discussed in Section 6.1. In particular, mean reversion (very likely) inflates these naïve estimates. I discuss mean reversion in Appendix Section A7.

linked to explicit incentives. The average teacher's value-added improves just 0.015σ when her scores count for earning tenure, that is, in her fourth and fifth years of teaching (standard error 0.007). Contrast that naïve estimate with the cumulative effect. Between her first and sixth year of teaching, the average teacher's value-added improved 0.079σ faster under the new tenure rules, compared to the average teacher tenured under the old rules (standard error 0.023, Appendix Table A5).[44]

This pattern of effects—especially the anticipation effects and persistent effects—is consistent with teachers investing in human capital (equivalently, improving their skills) as a response to the evaluation program's performance incentives. Combining the familiar features of agency theory and human capital investment models predicts both anticipation and persistence effects.

An alternative argument, sometimes raised in education policy discussions, is that evaluation can improve teacher performance without any extrinsic incentives, because teachers are motivated agents (Dixit 2002) who will use the individualized feedback from evaluation to improve. In the human capital investment framework, think of performance measures and feedback as new information which reduces the employee's costs of skill investments. The evidence presented here does not necessarily contradict that hypothesis. The new performance measures alone improved value-added by 0.024σ for already-tenured but still early-career teachers. These teachers had earned tenure under the old rules, after their third year, and, by statute, did not have any current or future incentives linked to their scores. Still, the gains were twice as large, 0.047σ, for pre-tenure teachers who were anticipating future incentives. If "motivated agents" plus "feedback" alone were sufficient for evaluation-induced skill improvements, then both groups would have experienced similar value-added gains.

[44] This cumulative effect estimate uses the same difference-in-differences strategy as the rest of the paper. All references to $(e-1)$ in equation (1) and elsewhere become $(e-5)$, thus the first difference is the change in value-added between $e=1$ and $e=6$ (see Appendix Section A6).

One limitation of this paper is that I do not have data measuring effort or skills directly. The estimated improvements in teacher performance, measured by value-added contributions to student test scores, are consistent with teachers putting effort into improving their skills. But evidence for or against skill investments would be clearer with direct measures of skill and effort inputs to complement measures of performance outputs. For example, the discussion in Section 1 differentiates between effort for current production and effort for learning skills, but I cannot measures these types of effort separately. Some skills improve through "learning by doing," that is, an employee can become more efficient at completing some task simply through repeating the task over and over in the normal course of work. Teachers may have learned new skills simply by working harder day to day.

Additionally, because of the lack of data on skills, in this paper I have not differentiated among different types or features of teaching skills. Define skill as an individual's efficiency in producing units of output, for example, the number of units produced in a given time interval or with a given amount of effort (as in Autor and Handel 2013). Skills can improve in a variety of ways: gaining greater understanding of the production process, increasing a capacity like physical or mental stamina, developing productive work habits, etc. While I cannot differentiate among these features of skill, all require effort to develop. Skill may also depend on innate endowments which, by definition, do not change over time, and would be differenced out in my identification strategy.

The most direct application of these results is in understanding the effects of similar teacher evaluation programs. Many states, like Tennessee, have policies which link teacher employment security to a "multi-measure" evaluation score. Popularized over the past decade, the multi-measure score typically combines both input measures, often rubric-scored classroom observations, with output measures, derived from student test scores. The estimates in this paper show that such programs can improve early-career teacher performance, making students better off. Additionally, Tennessee's new tenure rules had little effect on the patterns of teacher quits and job changes, despite the selection pressure suggested by only two-thirds of teachers performing above the new tenure cutoff. The benefits to teachers and students arose without the selection and turnover contemplated in Staiger and Rockoff (2010), Rothstein (2015), Dinerstein and Opper (2023), and elsewhere.

The results also have important practical implications for managers and policymakers designing performance measurement and incentive programs. First, the intended benefits of such programs—improved employee performance—can occur before or after the period when rewards are actively linked to performance. Reward costs occur when the rewards are active. Thus, the traditional focus on benefits which occur when rewards are active will understate the cost-effectiveness of the program. Second, the paper's results also raise design questions about the frequency of evaluation. The Tennessee program in this study, for example, evaluates each teacher annually, while programs in Cincinnati and France only evaluate teachers every five years or so (Taylor and Tyler 2012, Briole and Maurin in-press). If evaluation incentives cause skill development, and thus persistently higher performance, then annual evaluation may not optimize the cost-benefit tradeoff. However, these possibilities turn on the extent to which between-employee differences in performance are the result of differences in skills. The effects found here for teachers may or may not occur, for example, in the repair technicians case studied in Lazear (2000).

References

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

Aucejo, E. M., & Romano, T. F. (2016). "Assessing the effect of school days and absences on test score performance." *Economics of Education Review, 55*, 70-87.

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*(3), 417-430.

Autor, D. H., & Handel, M. J. (2013). "Putting tasks to the test: Human capital, job tasks, and wages." *Journal of Labor Economics, 31*(S1), S59-S96.

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.

Bacher-Hicks, A., & Koedel, C. (2023). "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, Volume 6,* 93-134. Elsevier.

Baker, G. (2002). "Distortion and risk in optimal incentive contracts." *Journal of Human Resources, 4*(4), 728-751.

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

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

Bell, C., James, J., Taylor, E. S., & Wyckoff, J. (in-press). "Measuring returns to experience using supervisor ratings of observed performance: The case of classroom teachers." *Journal of Policy Analysis and Management.*

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.

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. (2023). "Inducing positive sorting through performance pay: Experimental evidence from Pakistani schools." RISE Working Paper 23/123.

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.

Cisternas, G. (2018). "Career concerns and the nature of skills." *American Economic Journal: Microeconomics, 10*(2), 152-189.

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: Evaluating bias in teacher value-added estimates." *American Economic Review, 104*(9), 2633-2679.

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

de Chaisemartin, C. & D'Haultfoeuille, X. (2020) "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review, 110*(9), 2964-2996.

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

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. (2022). "Human capital depreciation and returns to experience." *American Economic Review, 112*(11), 3725-3762.

Dinerstein, M., & Opper, I. (2023). "Screening with multitasking." NBER No. 30310.

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

Duflo, E., Hanna, R., & Ryan, S. P. (2012). "Incentives work: Getting teachers to come to school." *American Economic Review, 102*(4), 1241-1278.

Fama, E. F. (1980). "Agency problems and the theory of the firm." *Journal of Political Economy, 88*(2), 288-307.

Fitzpatrick, M. D., Grissmer, D., & Hastedt, S. (2011). "What a difference a day makes: Estimating daily learning gains during kindergarten and first grade using a natural experiment." *Economics of Education Review, 30*(2), 269-279.

Gibbons, R. (1998). "Incentives in organizations." *Journal of Economic Perspectives, 12*(4), 115-132.

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

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, D.C.: Brookings Institution.

Griffith, R., & Neely, A. (2009). "Performance pay and managerial experience in multitask teams: evidence from within a firm." *Journal of Labor Economics, 27*(1), 49-82.

Hanno, E. C. (2022). "Immediate changes, trade-offs, and fade-out in high-quality teacher practices during coaching." *Educational Researcher, 51*(3), 173-185.

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

Hanushek, E. A., & Rivkin, S. G. (2010). "Generalizations about using value-added measures of teacher quality." *American Economic Review, 100*(2), 267-271.

Holmström, B. (1979). "Moral hazard and observability." *The Bell Journal of Economics, 10*(1), 74-91.

Holmström, B. (1999). "Managerial incentive problems: A dynamic perspective." *Review of Economic Studies, 66*(1), 169-182.

Holmström, B., & Milgrom, P. (1987). "Aggregation and linearity in the provision of intertemporal incentives." *Econometrica, 55*(2), 303-328.

Holmström, 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.

Hunter, S. B. (2018). *History of TEAM Teacher Evaluation Policy.* Tennessee Education Research Alliance, Vanderbilt University.

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

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., & Staiger, D. O. (2008). “Estimating teacher impacts on student achievement: An experimental evaluation.” NBER No. 14607.

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.

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. (2000). “Performance pay and productivity.” *American Economic Review*, *90*(5), 1346-1361.

Lazear, E. P., & 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.

Levin, J. (2003). “Relational incentive contracts.” *American Economic Review, 93*(3), 835-857.

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.

Mincer, J. (1962). “On-the-job training: Costs, returns, and some implications.” *Journal of Political Economy, 70*(5, Part 2), 50-79.

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

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, Volume 4*, 495-550. 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. March 30, 2013. “Curious grade for teachers: Nearly all pass.”

Ng, K. (2022). “The effects of teacher tenure on productivity and selection.” Working paper.

Oyer, P., & Schaefer, S. (2011). “Personnel economics: hiring and incentives.” In Ashenfelter, O., & Card, D. (eds), *Handbook of Labor Economics, Volume 4, Part B,* 1769-1823. 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.

Prendergast, C. (1993). “The role of promotion in inducing specific human capital acquisition.” *Quarterly Journal of Economics, 108*(2), 523-534.

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., 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.

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.

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.

SAS Institute. (2021). *SAS® EVAAS for K-12: Statistical Models*. SAS Institute White Paper. Retrieved from https://www.sas.com/en_us/software/evaas.html.

Springer, M. G., Hamilton, L., McCaffrey, D. F., Ballou, D., Le, V., Pepper, M., Lockwood, J. R., & Stecher, B. M. (2012). *Final report: Experimental evidence from the Project on Incentives in Teaching (POINT).* 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.

Taylor, E. (2023). "Teacher evaluation and training." In Hanushek, E. A., Machin, S., & Woessmann, L. (eds.), *Handbook of the Economics of Education, Volume 7,* 61-141. Elsevier.

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

Tennessee Department of Education. (2014). *New Tenure Law: Frequently Asked Questions.* https://team-tn.org/wp-content/uploads/2013/10/New-Tenure-Law-FAQs.pdf (last accessed April 22, 2024)

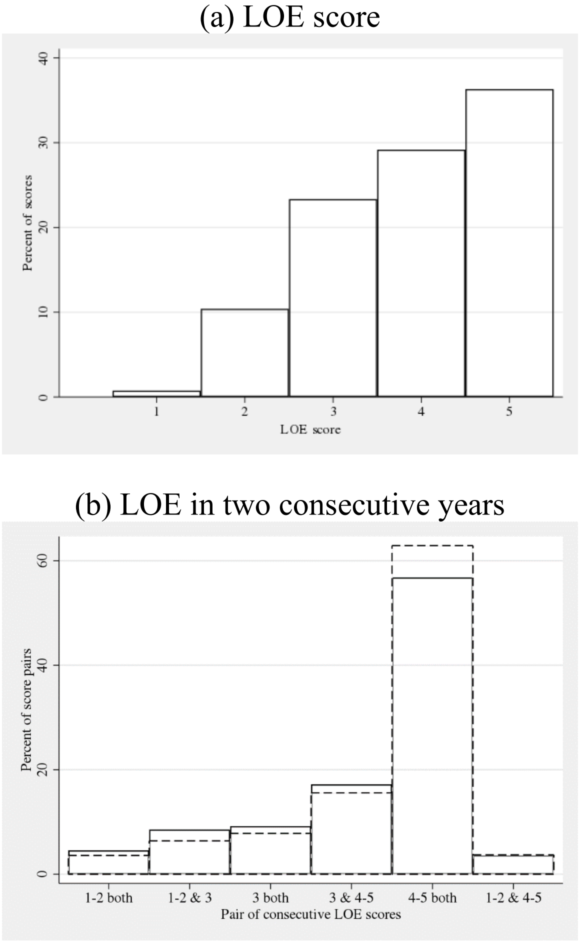
Todd, P. E., & Wolpin, K. I. (2007). "The production of cognitive achievement in children: Home, school, and racial test score gaps." *Journal of Human Capital, 1*(1), 91-136.

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*. The New Teacher Project.

Instruction: Questioning		
Significantly Below Expectations (1)	**At Expectations (3)**	**Significantly Above Expectations (5)**
Teacher questions are inconsistent in quality and include few question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions are random and lack coherence. • A low frequency of questions is asked. • Questions are rarely sequenced with attention to the instructional goals. • Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is inconsistently provided. • The teacher mostly calls on volunteers and high-ability students.	Teacher questions are varied and high quality providing for some, but not all, question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions are usually purposeful and coherent. • A moderate frequency of questions asked. • Questions are sometimes sequenced with attention to the instructional goals. • Questions sometimes require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is sometimes provided. • The teacher calls on volunteers and nonvolunteers, and a balance of students based on ability and sex.	Teacher questions are varied and high quality, providing a balanced mix of question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions are consistently purposeful and coherent. • A high frequency of questions is asked. • Questions are consistently sequenced with attention to the instructional goals. • Questions regularly require active responses (e.g., whole class signaling, choral responses, written and shared responses, or group and individual answers). • Wait time (3-5 seconds) is consistently provided. • The teacher calls on volunteers and nonvolunteers, and a balance of students based on ability and sex. • Students generate questions that lead to further inquiry and self-directed learning.

Figure 1—Classroom observation rubric example

Note: Reproduced from TEAM Educator Rubric 2012.


(a) LOE score

(b) LOE in two consecutive years

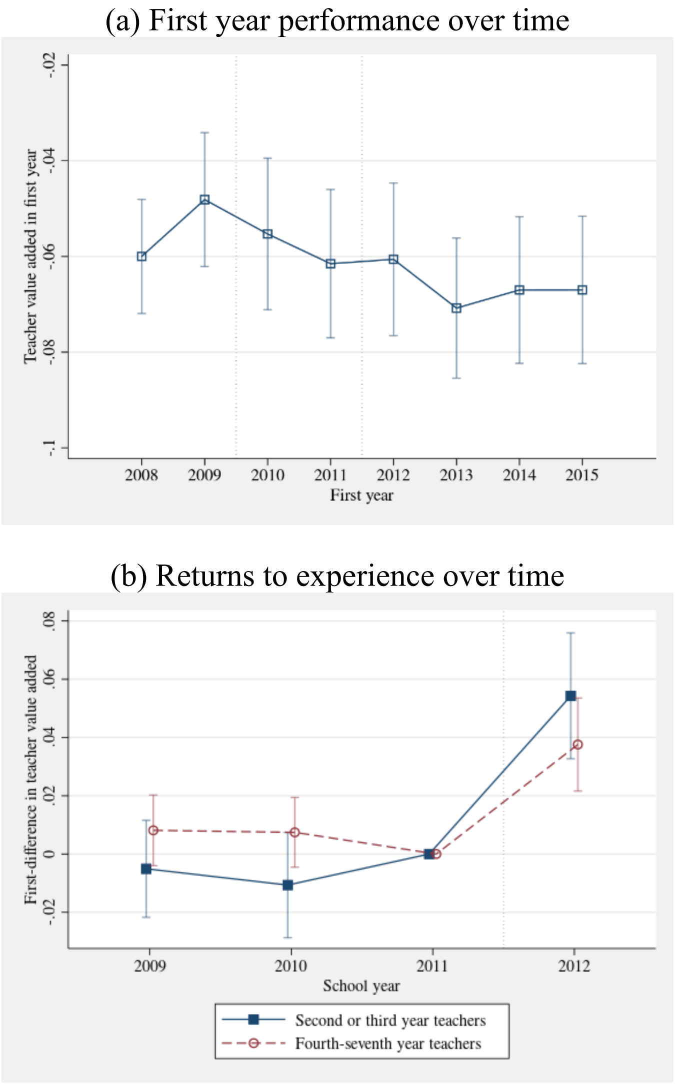
Figure 2—Distribution of final LOE scores

Note: LOE scores from 2012-2015 for teachers in this paper's analysis sample: teaching grades 4-8, math and English language arts; and in years 1-7 of employment. 18,974 teacher-by-year observations. Panel (a) annual LOE score. Panel (b) LOE scores in two consecutive years. Full sample shown with solid line bars. Dashed line bars show LOE scores specifically in year four and five of employment.

Year of employment	*Year hired*		
	≥ 2010		*< 2010*
1-3	no incentives		no incentives
4-5	must score LOE “4” or “5” in both years 4 and 5 to receive tenure cutoff for “4” ≅ 33rd percentile		
6+	*tenured*	*not tenured*	
	if rated LOE “1” or “2” two consecutive years tenure revoked cutoff for “2” ≅ 10th percentile	must score LOE “4” or “5” two consecutive years to receive tenure	

Figure 3—Performance incentives

Note: Author’s summary. See main text for a detailed description.


(a) First year performance over time

(b) Returns to experience over time

Figure 4—Trends in teacher performance and the returns to experience

Note: Panel (a): Each marker is a point estimate from a single least-squares regression. Vertical lines mark 95 percent cluster-corrected confidence intervals, with teacher clusters. The dependent variable is math or English language arts test score, standardized (mean 0, standard deviation 1), for student i taught subject s by teacher j in year t. The specification includes a flexible function of prior year test score, year fixed effects, and several other observable student characteristics. The x-axis = 2008 point in the graph is the estimated coefficient on an indicator = 1 if teacher j is in her first year teaching, $e = 1$, in year t and $t = 2008$. And similarly for 2009-2015. The omitted group is teachers in year $e \geq 7$ in year t.

Panel (b): Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 2 with the following exceptions. Instead of estimating a series of $\hat{\delta}_e$ for each e, for this graph I first estimate $\hat{\delta}_{et}$ for each t-by-e combination, then take a weighted average across e for a given year to obtain $\hat{\delta}_t$. The $\hat{\delta}_t$ are plotted in panel (b). For the solid-square series $e \in \{2,3\}$, and for the dashed-circle series $e \in \{4,5,6,7\}$, matching Table 2.

Table 1—Characteristics of study teachers and their students

	Teaching grades 4–8, math and ELA		All
	Years 1–7	All	teachers
	(1)	(2)	(3)
(a) Teachers			
Year of employment			
1	0.13	0.08	0.08
2	0.18	0.07	0.06
3	0.17	0.06	0.06
4	0.15	0.06	0.05
5	0.14	0.05	0.05
6	0.14	0.05	0.05
7	0.10	0.04	0.04
8+	0.00	0.58	0.61
Final LOE score	3.90	3.91	3.90
	(1.04)	(1.04)	(1.00)
Observation score	3.84	3.91	3.85
	(0.55)	(0.58)	(0.58)
Total salary (1,000s)	39.57	45.01	47.40
	(6.81)	(9.52)	(13.49)
Observations (teacher-year)	36,831	110,642	621,720
(b) Students			
Prior year test score			
Math	0.03	0.06	
	(0.95)	(0.95)	
English language arts	0.05	0.07	
	(0.96)	(0.96)	
Grade level			
4	0.20	0.20	
5	0.20	0.20	
6	0.19	0.20	
7	0.22	0.20	
8	0.19	0.19	
Female	0.50	0.50	
Race/ethnicity			
White	0.66	0.68	
Black	0.24	0.23	
Other or more than one	0.09	0.09	
Free or reduced-price lunch	0.54	0.53	
English language learner	0.09	0.08	
Special education	0.09	0.10	
Observations (student-year-subject)	1,806,725	5,158,868	

Note: Means and standard deviations, in parentheses, for 2008–2015 school years. Year of employment = 1 is the teacher's first year working in Tennessee public schools. Year of employment increments up each school year even if the teacher took a leave of absence, following the paper's intent to treat approach. Student test scores are standardized (mean 0, standard deviation 1) within grade-by-year-by-subject cells; positive means for test scores reflect negative selection of students leaving Tennessee public schools.

Table 2—Effects of future incentives on value-added performance

	Pooled	Math	ELA
	(1)	(2)	(3)
(i) Future tenure incentives and new measures	0.047	0.065	0.028
	(0.009)	(0.015)	(0.009)
(ii) Never incentives and new measures	0.024	0.036	0.013
	(0.006)	(0.011)	(0.006)
Triple-difference: (i) – (ii)	0.023	0.029	0.015
	(0.011)	(0.018)	(0.011)
Teacher observations	6,998	4,291	5,406

Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. Each component regression is an iteration of the same least-squares specification, differing by the value of e. The dependent variable is math or English language arts (ELA) test score, standardized (mean 0, standard deviation 1), for student i taught subject s by teacher j in year t. The estimation sample for a given regression is limited to only observations, $ijst$, where teacher j is in year e or $(e-1)$ of her employment. The key estimate $\hat{\delta}_e$ from each regression is the coefficient on a treatment indicator D_{je}, and the estimation sample is further limited to only observations where either $\{D_{j,e-1}=0, D_{je}=1\}$ "treated" or $\{D_{j,e-1}=0, D_{je}=0\}$ "comparison" teachers. In this table $D_{je}=1$ if year e occurred in $t=2012$, the first year of the new program. The specification also includes an indicator for year e, teacher fixed effects, a flexible function of student i's prior year test score, and several other observable student and peer characteristics detailed in the text. The top row of the table is a weighted average of $\hat{\delta}_e$ across $e \in \{2,3\}$, where the weights are the number of treated teachers in the estimation sample for $\hat{\delta}_e$. The second row is the same weighted average across $e \in \{4,5,6,7\}$. The third row is the triple-difference: row 1 minus row 2. Columns 2–3 report estimates for subsamples described in the header.

Table 3—Effects of future incentives on turnover (attrition)

	Quit or changed jobs: no longer teaching tested grade/subject in year e		Quit: no longer teaching in year e	
	(1)	(2)	(3)	(4)
(a) Future tenure incentives and new measures				
(i) Future tenure incentives and new measures	-0.013	-0.017	0.019	0.017
	(0.014)	(0.010)	(0.009)	(0.008)
(ii) Value-added in year $e-1$		-0.023		-0.005
		(0.006)		(0.002)
(i) × (ii)		-0.007		-0.007
		(0.011)		(0.007)
Attrition rate in comparison group	0.314		0.073	
(b) Never incentives and new measures				
(i) Never incentives and new measures	-0.013	0.014	0.005	0.012
	(0.009)	(0.008)	(0.005)	(0.005)
(ii) Value-added in year $e-1$		-0.024		-0.005
		(0.004)		(0.002)
(i) × (ii)		-0.003		0.004
		(0.008)		(0.004)
Attrition rate in comparison group	0.243		0.042	
(c) Triple-difference				
(a)(i) – (b)(i)	0.000	-0.031	0.014	0.005
	(0.017)	(0.013)	(0.010)	(0.010)
(a)(ii) – (b)(ii)		0.001		0.001
		(0.007)		(0.003)
(a)(i) × (a)(ii) – (b)(i) × (b)(ii)		-0.004		-0.010
		(0.013)		(0.008)
Teacher observations	10,097	10,097	10,097	10,097

Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. Each component regression is an iteration of the same least-squares specification, differing by the value of e. The estimation sample for a given regression is limited to teachers, j, observed working as teachers of a tested grade/subject in their $(e-1)$th year of employment. In columns 1–2, the dependent variable is an indicator $=1$ if teacher j had attrited in year e: (i) teacher j was no longer working as a teacher in Tennessee public schools in year e, or (ii) j was teaching but no longer teaching of a tested grade/subject in year e. In columns 3–4, the dependent variable $=1$ only in case (i). The estimates in columns 1 and 3 are analogous to the estimates in Table 2. Columns 2 and 4 add regressors for the teacher's value-added score from her $(e-1)$th year of employment and the interaction with the treatment indicator. Panel A is a weighted average of $\hat{\delta}_e$ across $e \in \{2,3\}$, where the weights are the number of treated teachers in the estimation sample for $\hat{\delta}_e$. Panel B is the same weighted average across $e \in \{4,5,6,7\}$. Panel C is the triple-difference.

Table 4—Effects when incentives begin and after incentives end

	Pooled			Math	ELA
	Value-added	Quit or changed jobs	Quit	Value-added	Value-added
	(1)	(2)	(3)	(4)	(5)
(a) When tenure incentives begin, year 4					
Tenure incentives begin, year 4	0.013	0.004	0.004	0.019	0.006
	(0.009)	(0.013)	(0.007)	(0.014)	(0.009)
Attrition rate in comparison group		0.257	0.054		
Teacher observations	3,849	5,295	5,295	2,304	2,665
(b) After tenure incentives end (or are scheduled to end), year 6					
(i) Tenure incentives end or are scheduled to end, year 6	0.037	-0.005	-0.006	0.039	0.036
	(0.012)	(0.021)	(0.008)	(0.020)	(0.013)
(ii) Never incentives, year 6	0.021	-0.005	0.002	0.033	0.009
	(0.009)	(0.015)	(0.006)	(0.015)	(0.010)
Triple-difference: (i) – (ii)	0.016	-0.001	-0.007	0.007	0.027
	(0.013)	(0.021)	(0.009)	(0.021)	(0.014)
Attrition rate in comparison group		0.232	0.036		
Teacher observations	3,426	4,557	4,557	2,011	2,402

Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 2 for columns 1 and 4–5, and the same as in Table 3 for columns 2–3, with the following exceptions: Panel A reports only estimates for the case $(e-1)=3$ and $e=4$. The treated group is only teachers who were subject to the new tenure rules, i.e., earning tenure required scoring above a cutoff in both years $e=4$ and 5. The comparison group is limited to teachers who reached $e=4$ in $t \leq 2011$, before the new program began. Panel B reports only estimates for the case $(e-1)=5$ and $e=6$. The treated group in row (i) is only teachers subject to the new rules. The treated group in row (ii) is teachers who received new measures but were not subject to the new tenure rules. The comparison group is limited to teachers who reached $e=6$ in $t \leq 2011$.

Table 5—Effects after incentives end

	Pooled			Math	ELA
	Value-added	Quit or changed jobs	Quit	Value-added	Value-added
	(1)	(2)	(3)	(4)	(5)
(a) Teachers who (would have) successfully earned tenure on time, under the new tenure rules					
(i) Incentives end, newly-tenured	0.024	-0.057	-0.012	0.039	0.006
	(0.013)	(0.022)	(0.008)	(0.020)	(0.015)
(ii) Never incentives but would have earned tenure	-0.000	-0.044	-0.006	0.010	-0.013
	(0.010)	(0.016)	(0.007)	(0.015)	(0.011)
Triple-difference: (i) – (ii)	0.025	-0.013	-0.006	0.029	0.019
	(0.014)	(0.023)	(0.008)	(0.022)	(0.015)
(b) Teachers who (would have) failed to earn tenure on time, under the new tenure rules					
(i) Incentives continue, not-yet-tenured	0.069	0.046	0.009	0.030	0.092
	(0.025)	(0.038)	(0.021)	(0.056)	(0.022)
(ii) Never incentives, but would not have earned tenure	0.086	0.054	0.003	0.134	0.055
	(0.016)	(0.026)	(0.010)	(0.031)	(0.015)
Triple-difference: (i) – (ii)	-0.016	-0.008	0.007	-0.104	0.036
	(0.028)	(0.044)	(0.022)	(0.061)	(0.025)
Attrition rate in comparison group		0.232	0.036		
Teacher observations	3,426	4,557	4,557	2,011	2,402

Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 2 for columns 1 and 4–5, and the same as in Table 3 for columns 2–3, with the following exceptions: Table 5 reports only estimates for the case $(e-1)=5$ and $e=6$. The treated groups in rows (a)(i) and (b)(i) were subject to the new tenure rules; the treated groups in rows (a)(ii) and (b)(ii) were not. The comparison group, in all rows, is limited to teachers who reached $e=6$ in $t \leq 2011$.

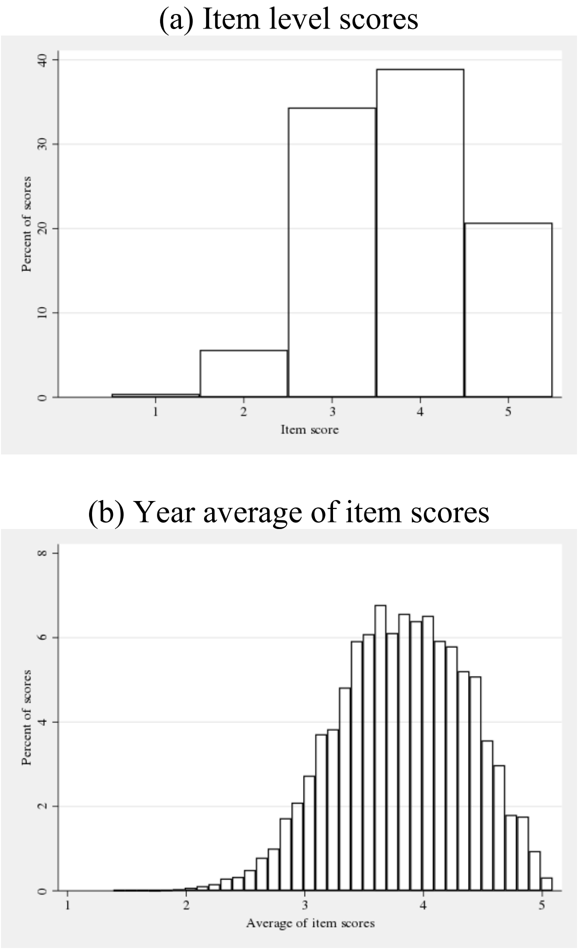
Employee evaluation and skill investments: Evidence from public school teachers

Appendix

Eric S. Taylor
Harvard University and NBER

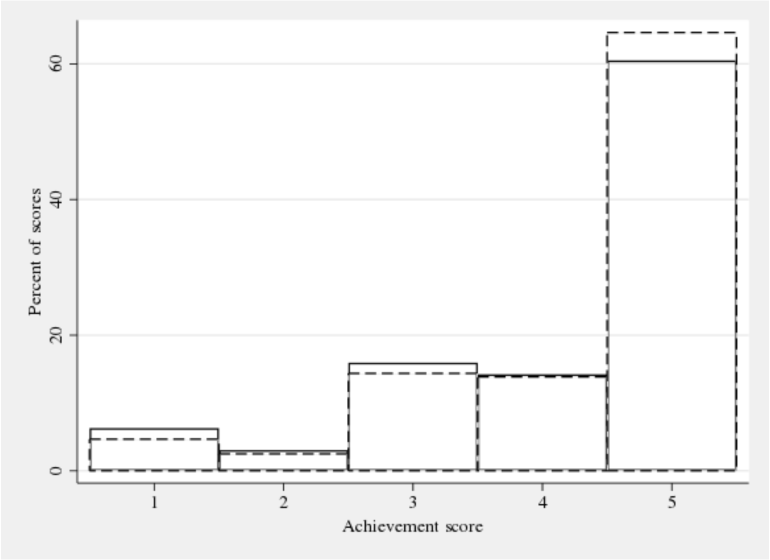
May 2024

A1. Additional Figures and Tables



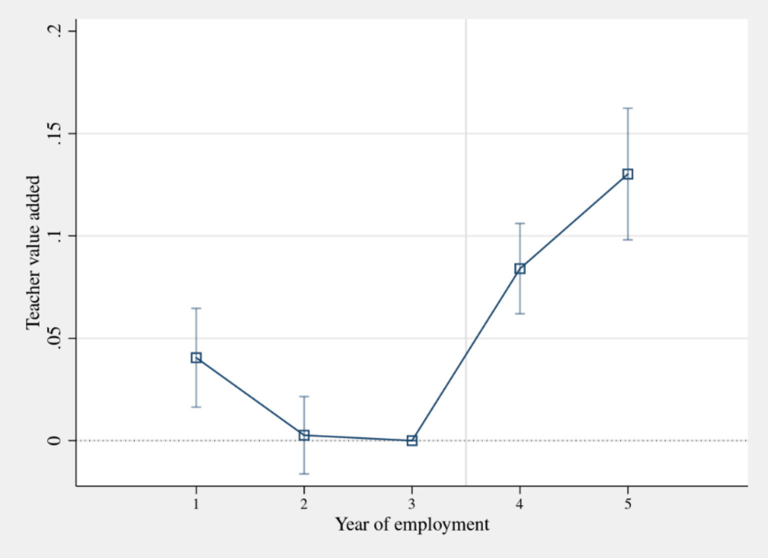
Appendix Figure A1—Distribution of classroom observation scores

Note: Classroom observation scores, TEAM rubric, from 2012–2015 for teachers in this paper's analysis sample: teaching grades 4–8, math and English language arts; and in years 1–6 of employment. Panel (a) shows item level scores—one score for each time a task was scored. 565,885 item score observations. Panel (b) shows a teacher's annual average of item scores. 15,169 teacher-by-year observations.



Appendix Figure A2—Distribution of "achievement score" component of LOE

Note: Achievement scores from 2012–2015 for teachers in this paper's analysis sample: teaching grades 4–8, math and English language arts; and in years 1–6 of employment. 19,172 teacher-by-year observations. Full sample shown with solid line bars. Dashed line bars show the subsample of districts that adopted the "value-added override" rule that the student growth (TVAAS) score replaces the achievement score when student growth score is 3 or higher.



Appendix Figure A3—Difference in performance between teachers who scored below the tenure cutoff in year 3 and those who scored above

Note: Each marker is a point estimate from a single least-squares regression. Vertical lines mark 95 percent cluster-corrected confidence intervals, with teacher clusters. The dependent variable is math or English language arts test score, standardized (mean 0, standard deviation 1), for student i taught subject s by teacher j in year t. The specification includes teacher and year fixed effects, a flexible function of prior year test score, and several other observable student and peer characteristics. The key independent variables are (i) a series of indicators for teacher j's year of employment, with $e = 3$ the omitted category, (ii) an indicator $= 1$ if teacher j scored LOE ≤ 3 in year $e = 3$, and (iii) the interaction of (i) and (ii). The plotted estimates are the point estimates on the interaction terms.

Appendix Table A1—Effects on value-added performance, randomization inference *p*-values and excluding pay-for-performance schools

	Main estimates	Rand. inf. *p*-values	Excluding P4P schools
	(1)	(2)	(3)
(a) Table 2			
(i) Future incentives and new measures	0.047	0.038	0.046
	(0.009)		(0.010)
(ii) Never incentives and new measures	0.024	0.142	0.027
	(0.006)		(0.007)
Triple-difference: (i) – (ii)	0.023	0.094	0.019
	(0.011)		(0.012)
Teacher observations	6,998		6,016
(b) Table 4			
Tenure incentives begin, year 4	0.013	0.110	0.017
	(0.009)		(0.010)
Teacher observations	3,849		3,279
(i) Tenure incentives end	0.037	0.056	0.044
or are scheduled to end, year 6	(0.012)		(0.016)
(ii) Never incentives, year 6	0.021	0.154	0.020
	(0.009)		(0.010)
Triple-difference: (i) – (ii)	0.016	0.250	0.023
	(0.013)		(0.016)
Teacher observations	3,426		2,965
(c) Table 5			
(i) Incentives end, newly-tenured	0.024	0.030	0.025
	(0.013)		(0.018)
(ii) Never incentives	-0.000	0.424	0.000
but would have earned tenure	(0.010)		(0.011)
Triple-difference: (i) – (ii)	0.025	0.046	0.025
	(0.014)		(0.018)
(i) Incentives continue, not-yet-tenured	0.069	0.168	0.096
	(0.025)		(0.034)
(ii) Never incentives,	0.086	0.072	0.084
but would not have earned tenure	(0.016)		(0.018)
Triple-difference: (i) – (ii)	-0.016	0.530	0.012
	(0.028)		(0.037)
Teacher observations	3,426		2,965

Note: This table is an extension of Tables 2, 4, and 5 in the main text. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. Column 1 in this table repeats the estimates from column 1 in Tables 2, 4, and 5. Column 2 reports randomization inference *p*-values for the estimates in column 1, where hire cohorts (cluster) are randomly assigned to treatments to construct a null distribution (see Appendix Section A3). The only difference between column 1 and 3 in this table is that in column 3 the estimation sample excludes districts and schools that had pay for performance (P4P) programs (see Appendix Section A2).

Appendix Table A2— Effects of future incentives on value-added performance, additional estimates

	Pooled	Math	ELA
	(1)	(2)	(3)
(a) Future incentives and new measures			
$e \in \{2,3\}$, Table 2 row 1	0.047	0.065	0.028
	(0.009)	(0.015)	(0.009)
$e = 2$	0.057	0.083	0.029
	(0.012)	(0.020)	(0.013)
$e = 3$	0.035	0.042	0.028
	(0.013)	(0.022)	(0.013)
(b) Never incentives and new measures			
$e \in \{4,5,6,7\}$, Table 2 row 2	0.024	0.036	0.013
	(0.006)	(0.011)	(0.006)
$e = 4$	0.014	0.032	-0.003
	(0.012)	(0.021)	(0.012)
$e = 5$	0.029	0.044	0.014
	(0.012)	(0.021)	(0.012)
$e = 6$	0.023	0.036	0.010
	(0.012)	(0.021)	(0.011)
$e = 7$	0.033	0.031	0.035
	(0.013)	(0.023)	(0.013)
Teacher observations	6,998	4,291	5,406

Note: This table is an extension of Table 2 in the main text. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 2. Panel A row 1 and panel B row 1 repeat estimates from Table 2. The remaining rows report individual $\hat{\delta}_e$ from the component regressions.

Appendix Table A3— Effects of future incentives on turnover (attrition), additional details

	Quit or changed jobs: no longer teaching tested grade/subject in year *e*		Quit: no longer teaching in year *e*	
	(1)	(2)	(3)	(4)
(a) Table 4, panel a				
(i) Tenure incentives begin, year 4	0.004	0.032	0.004	0.011
	(0.013)	(0.012)	(0.007)	(0.007)
(ii) Value-added in year 3		-0.029		-0.010
		(0.008)		(0.004)
(i) × (ii)		-0.004		-0.002
		(0.011)		(0.006)
Attrition rate in comparison group	0.257		0.054	
Teacher observations	5,295	5,295	5,295	5,295
(b) Table 4, panel b				
(i) Tenure incentives end or are scheduled to end, year 6	-0.005	0.020	-0.006	0.001
	(0.021)	(0.020)	(0.008)	(0.009)
(ii) Value-added in year 5		-0.023		-0.002
		(0.009)		(0.003)
(i) × (ii)		-0.021		-0.012
		(0.019)		(0.010)
Attrition rate in comparison group	0.232		0.036	
Teacher observations	4,557	4,557	4,557	4,557

Note: This table is an extension of Table 4 in the main text. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. Columns 1 and 3 in this table repeat the estimates from columns 2–3 in Table 4. Columns 2 and 4 in this table add regressors for the teacher's value-added score from year 3 of employment and the interaction with the treatment indicator.

Appendix Table A4—Effects after incentives end on turnover (attrition), additional details

	Quit or changed jobs: no longer teaching tested grade/subject in year 6		Quit: no longer teaching in year 6	
	(1)	(2)	(3)	(4)
(a) Teachers who successfully earned tenure on time, under the new tenure rules				
(i) Incentives end, newly-tenured	-0.057	-0.015	-0.012	-0.004
	(0.022)	(0.024)	(0.008)	(0.009)
(ii) Value-added in year 5		-0.023		-0.002
		(0.009)		(0.003)
(i) × (ii)		-0.004		0.000
		(0.019)		(0.005)
(b) Teachers who would have successfully earned tenure on time, under the new tenure rules				
(i) Never incentives	-0.044	-0.003	-0.006	0.002
but would have earned tenure	(0.016)	(0.016)	(0.007)	(0.006)
(ii) Value-added in year 5		-0.023		-0.002
		(0.009)		(0.003)
(i) × (ii)		0.005		0.002
		(0.013)		(0.004)
(c) Triple-difference				
(a)(i) – (b)(i)	-0.013	-0.012	-0.006	-0.006
	(0.023)	(0.025)	(0.008)	(0.010)
(a)(i) × (a)(ii) – (b)(i) × (b)(ii)		-0.009		-0.002
		(0.019)		(0.006)

Table A4 continues on the next page.

Appendix Table A4 (cont.)—Effects after incentives end on turnover (attrition), additional details

	Quit or changed jobs: no longer teaching tested grade/subject in year 6		Quit: no longer teaching in year 6	
	(1)	(2)	(3)	(4)
(d) Teachers who failed to earn tenure on time, under the new tenure rules				
(i) Incentives continue, not-yet-tenured	0.046	0.006	0.009	-0.019
	(0.038)	(0.035)	(0.021)	(0.016)
(ii) Value-added in year 5		-0.023		-0.002
		(0.009)		(0.003)
(i) × (ii)		-0.050		-0.043
		(0.040)		(0.029)
(e) Teachers who would have failed to earn tenure on time, under the new tenure rules				
(i) Never incentives,	0.054	0.077	0.003	0.017
but would not have earned tenure	(0.026)	(0.030)	(0.010)	(0.012)
(ii) Value-added in year 5		-0.023		-0.002
		(0.009)		(0.003)
(i) × (ii)		0.002		0.008
		(0.031)		(0.010)
(f) Triple-difference				
(d)(i) – (e)(i)	-0.008	-0.071	0.007	-0.036
	(0.044)	(0.044)	(0.022)	(0.019)
(d)(i) × (e)(ii) – (d)(i) × (e)(ii)		-0.052		-0.052
		(0.049)		(0.031)
Attrition rate in comparison group	0.232		0.036	
Teacher observations	4,557	4,557	4,557	4,557

Note: This table is an extension of Table 5 in the main text. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. Columns 1 and 3 in this table repeat the estimates from columns 2–3 in Table 5. Columns 2 and 4 in this table add regressors for the teacher's value-added score from year 3 of employment and the interaction with the treatment indicator. In panels C and F there is no triple-difference estimate for the main effect of value-added in year 5. This estimate does not change across panels because the comparison group is constant across panels.

Appendix Table A5—Cumulative effects of incentives, from year 1 to 6

	Pooled					Math	ELA
	Value-added	Quit or changed jobs		Quit		Value-added	Value-added
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Triple-difference	0.079	0.026	0.023	-0.001	-0.004	0.106	0.051
	(0.023)	(0.020)	(0.018)	(0.011)	(0.010)	(0.038)	(0.021)
			-0.056		0.000		
			(0.013)		(0.007)		
			0.033		-0.003		
			(0.022)		(0.012)		
Teacher observations	961	2,700	2,700	2,700	2,700	530	619

Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 2 for columns 1 and 6–7, and the same as in Table 3 for columns 2–5, with the following exceptions: In this table the first difference is between $(e - 5) = 1$ and $e = 6$; in equation (1) $(e - 1)$ is replaced with $(e - 5)$. The treated group is only teachers who were subject to the new tenure rules. The comparison group is teachers who were not subject to the new tenure rules. See Appendix Section A6 for a discussion of how the estimates in this table are triple-difference estimates.

Appendix Table A6—Effects when incentives begin,
additional details

	Value-added	Quit or changed jobs: no longer teaching tested grade/subject in year *e*		Quit: no longer teaching in year *e*	
	(1)	(2)	(3)	(4)	(5)
(a) Below tenure cutoff, LOE < 4, in year 3					
(i) Tenure incentives begin, year 4	0.077	0.062	0.052	0.021	0.014
	(0.012)	(0.020)	(0.021)	(0.011)	(0.013)
(ii) Value-added in year 3			-0.029		-0.010
			(0.008)		(0.004)
(i) × (ii)			-0.009		-0.008
			(0.022)		(0.015)
(b) Above tenure cutoff, LOE ≥ 4, in year 3					
(i) Tenure incentives begin, year 4	-0.018	-0.064	-0.013	-0.011	-0.001
	(0.009)	(0.014)	(0.013)	(0.007)	(0.008)
(ii) Value-added in year 3			-0.029		-0.010
			(0.008)		(0.004)
(i) × (ii)			0.005		0.007
			(0.011)		(0.006)
(c) Classroom observation rating < 4, in year 3					
(i) Tenure incentives begin, year 4	0.018	0.001	0.025	0.014	0.019
	(0.010)	(0.015)	(0.014)	(0.009)	(0.009)
(ii) Value-added in year 3			-0.029		-0.010
			(0.008)		(0.004)
(i) × (ii)			-0.010		-0.006
			(0.015)		(0.008)
(d) Classroom observation rating ≥ 4, in year 3					
(i) Tenure incentives begin, year 4	0.005	-0.053	-0.001	-0.010	0.003
	(0.012)	(0.017)	(0.017)	(0.009)	(0.010)
(ii) Value-added in year 3			-0.029		-0.010
			(0.008)		(0.004)
(i) × (ii)			0.008		0.007
			(0.014)		(0.007)
		0.257		0.054	
Teacher observations	3,849	5,295	5,295	5,295	5,295

Note: This table is an extension of Table 4 panel A in the main text, and Appendix Table A3 panel A. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in Table 4 panel A and Appendix Table A3 panel A, except that in this table the treated group is limited as described in the panel headers in this table. The comparison group is unchanged from Table 4 panel A and constant across all panels in this table.

Appendix Table A7—Effects after incentives end,
by prior observation rating

	Pooled			Math	ELA
	Value-added	Quit or changed jobs	Quit	Value-added	Value-added
	(1)	(2)	(3)	(4)	(5)
(a) Classroom observation rating ≥ 4, in year 5					
(i) Subject to new tenure rules	0.029	-0.078	-0.017	0.051	0.006
	(0.016)	(0.026)	(0.009)	(0.026)	(0.018)
(ii) Not subject to new tenure rules	0.000	-0.036	-0.003	0.004	-0.005
	(0.013)	(0.020)	(0.008)	(0.021)	(0.013)
Triple-difference: (i) – (ii)	0.028	-0.042	-0.015	0.047	0.010
	(0.019)	(0.029)	(0.011)	(0.031)	(0.020)
(b) Classroom observation rating < 4, in year 5					
(i) Subject to new tenure rules	0.036	0.044	0.006	0.019	0.057
	(0.018)	(0.031)	(0.014)	(0.026)	(0.021)
(ii) Not subject to new tenure rules	0.024	-0.027	0.002	0.038	0.011
	(0.011)	(0.018)	(0.009)	(0.019)	(0.012)
Triple-difference: (i) – (ii)	0.012	0.071	0.004	-0.019	0.046
	(0.019)	(0.033)	(0.015)	(0.030)	(0.022)
Attrition rate in comparison group		0.232	0.036		
Teacher observations	3,426	4,557	4,557	2,011	2,402

Note: This table is closely related to Table 5 in the main text. Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are identical to Table 5, except that here the two groups (the two panels in the table) are defined by classroom observation ratings in year 5. The comparison group is unchanged from Table 5 and constant across panels in this table.

A2. Setting and Data, Additional Details

Sampling constraints—The 2008-2015 period is defined by the following constraints. First, the available data begin in 2007, and thus 2008 is the first year for which I observe lagged test scores. Second, the period after 2015 includes challenges with test administration and resulting changes to teacher evaluation rules. In 2016 a new online testing system failed and students in grades 3-8 were not tested. Given the importance of lagged test scores in my empirical approach, the lack of 2016 scores also excludes 2017. In 2018 various events complicated and delayed testing. As a result of subsequent legislation, 2018 student test scores could not determine any adverse consequences for teachers, like dismissal or tenure denial. Further each teacher's final 1-5 evaluation rating was calculated with and without 2018 student scores, and the teacher was given the higher of the two. Last a teacher could choose to void their entire evaluation score for 2018. In 2020 testing was cancelled because of the pandemic.

Number of classroom observations per year—Untenured and low-performing teachers are scored three times on instruction-related tasks and twice on classroom environment and planning tasks. Tenured and high-performing teachers are scored as little as once per task.

Differences between TEAM and systems—This paragraph describes details of the TEAM system which applies to more than 80 percent of teachers in Tennessee. And the results in this paper are robust to limiting the analysis sample to just TEAM school districts. The other systems are their key differences are: (i) TEM. 10 percent of teachers. Used in Shelby County. TEM uses a different rubric, which groups tasks into a smaller number of scored items. Though the state requires that all rubrics cover the same basic teaching tasks. Teachers rate themselves in addition to the observer's scores. (ii) COACH. 6 percent of teachers. Used in Hamilton County and a few nearby districts. COACH uses a different rubric, where many more distinct tasks are scored. Visits are shorter but more frequent. At the end of the year the school principal rates each task at her discretion, informed by the results of the year's observations but not a mechanical function of them. (iii) TIGER. 2-3 percent of teachers. TIGER uses the same rubric as TEAM. Observations are conduced by coaches. At the end of the year the school principal chooses the overall observation rating of 1-5 at her discretion, informed by observation results and other information.

Additional details of LOE scoring—First, the state allowed districts to adopt (or not) two "value added override" rules: (a) achievement = max(achievement , value-added) if value-added is 3 or higher, and (b) LOE = max(LOE , value-added) if value-added is 4 or higher. Rule (b) was possible only from 2013 forward. In my sample, about half of teachers were in districts that adopted these rules, ranging from 37 to 73 percent depending on the year and rule. Note, however, that these rules change the final LOE score only after the school year is over. Uncertainty in predicting one's own value-added score would make ignoring observation scores a risky strategy. And sampling error alone can generate substantial uncertainty in an individual teacher's value-added score. Second, starting in 2014 school districts could choose to add a fourth measure based on student surveys. Approximately 5 percent of the treated teachers in my sample have an LOE based partly on student surveys. For those teachers the 0.50 weight to observations is divided into 0.45 for observations and 0.05 for student surveys.

Additional details on tenure rules—Years teaching outside of Tennessee do not count toward the requirements, both for these new rules and the old tenure rules discussed below. Two additional details, which are not first-order in practice: First, to earn tenure the teacher must hold a "Professional" certification level, as opposed to the entry-level "Practitioner" (equivalently "Apprentice") certification. Earning the Professional certificate requires completing a state-approved teacher preparation program and scoring LOE ≥ 2 for three consecutive years. Second, assume a teacher has met the LOE requirement in years t and $(t-1)$. The local school district can still choose to fire a teacher immediately after year t, but retaining the teacher into year $(t + 1)$ grants the teacher tenure.

Pay for Performance—Also beginning in the 2012 school year, 10 percent of Tennessee school districts (14) began paying some teachers based partly on evaluation scores. This pay-for-performance treatment is confounded with the evaluation treatment, however, later in Appendix Section A4, I show that the paper's results are robust to excluding those districts entirely. Plan details differed considerably across districts, but the basic features were similar. Teachers could receive cash bonuses based on individual, team, or school performance, though three-quarters of bonuses paid were based on individual evaluation scores. Roughly half of bonuses were based on teachers' annual LOE score, with another one-quarter based on test-score value-added measures.

In 2015, the last school year in my data, one-third of Tennessee districts (48) began new or revised pay-for-performance programs. Again, the paper's results are robust to excluding these observations from the estimation sample. In the 2015 plans, three-quarters of bonuses were paid based on teachers' annual LOE score, with another one-eighth based on test-score value-added.

Pay for Performance, Additional Details—The paragraph's above describe pay for performance programs during the period 2012-2015. I draw mainly on Ballou et al. (2016) for the plans starting in 2012, and Tennessee Department of Education (n.d.) for the plans starting in 2015.

The motivation and funding for the 2012 programs came from federal grants: the Race to the Top and Teacher Incentive Fund. For a complete description of the programs and evaluations see Canon et al. (2012), Ballou et al. (2015, 2016). School-level bonuses accounted for 22 percent, with another 3 percent based on grade-level or department performance. Teachers also still received raises based on experience and earned degrees, but those increases were reduced. Teachers already working in the district in 2012 could opt out of the LOE-based salary raises schedule. The typical bonus earned was under \$2,000, adding about 3-5 percent to the average teacher's salary. Across districts the maximum possible bonus ranged between \$2,000-7,000 (5th-95th percentile). Additionally, in a small subset of these pay-for-performance districts (1 percent of my sample), teachers' annual salary increases were based in part on their LOE scores. A teacher would receive no raise if he scored LOE < 3. Then raises of 1-3 percent were scaled to LOE scores ≥ 3.

At the end of the 2013 school year Tennessee offered a one-time retention bonus to teachers who worked in "priority" schools (the 5 percent lowest performing schools in the state), and who had scored LOE = 5 (Springer, Swain, and Rodriguez 2016). The bonus was unlikely to affect performance: it was announced in May after LOE scores were largely determined for 2013 and there was no promise of repeating the bonuses in the future. Additionally, one-third of priority schools chose not to participate in the bonus program.

The well-known POINT pay-for-performance experiment in Tennessee occurred in 2007-2009. The POINT sample was grade 5-8 math teachers in Metro Nashville Public Schools, with about 150 treated teachers. Among this paper's estimation sample, POINT treatment teachers represent less than 0.5 percent in the pre-2012 period. Moreover, the experiment found almost no effects on teacher performance (Springer et al. 2012).

The Common Core in Tennessee—In 2012, the same year the new evaluation program began, Tennessee also began implementation of new state standards consistent with the Common Core initiative. However, in 2012 the new standards were only used in kindergarten through grade 2, not in grades 4-8 which contribute to this paper's estimates. The new math and English language arts standards for grades 3-8 were used by some districts in 2013 and all districts by 2014.

A3. Identification Strategy, Additional Details

A3.1 Comparison to Two-Way FE Estimator

The diff-in-diff estimator (1), “DID_M,” was proposed by de Chaisemartin and D’Haultfœuille (2020) to address potential bias in the two-way fixed effects diff-in-diff estimator (“two-way FE”).[1] Indeed, if I fit (2) without sample restrictions it would be a two-way FE estimate. A brief summary of the differences: First, my DID_m-style estimator weights each $\hat{\delta}_e$ simply by N_e, while two-way FE weights by a function of $N_e + M_e$ and $var(D_{je})$. The two-way FE weights are precision-maximizing if the δ_e are homogeneous but introduce bias in $\hat{\delta}$ if δ_e are heterogenous. Second, the two-way FE estimator is also biased when treatment effects are heterogeneous over time within units, because previously treated units are used in the comparison group for later treated units (sometimes called the “negative weights” problem). Thus DID_M (i) uses only variation most proximate to the change in treatment status, i.e., $(e-1)$ and e in the current case; and (ii) includes only untreated units in the comparison group, i.e., teachers for whom $\{D_{j,e-1} = 0, D_{je} = 0\}$. Third, as Goodman-Bacon (2021, Section IV) shows, when additional controls are included, two-way FE uses control coefficients estimated using the full sample, which again can introduce bias. By re-estimating (2) for each $\hat{\delta}_e$ separately my approach avoids this problem.

A3.2 Randomization Inference

Throughout the paper I report cluster-corrected standard errors with teacher clusters. For comparison, Appendix Table A1 reports randomization inference *p*-values (column 2) for the paper’s main estimates (repeated in column 1). Matching the statistical inference reported in Tables 2 and 5, anticipation and persistence

[1] Even if μ_{je} were observed, notice that (1) could itself be carried out using a system of least squares regressions, with appropriately defined sample constraints and weights.

estimates are statistically significant at conventional levels with randomization inference.

In the paper I estimate the effects of several different treatments experienced by teachers: announcement of new tenure rules, new performance measures, the start and end of performance incentives, etc. Treatment status—which teachers are treated, with which specific treatment(s), when they are treated—is determined by the teacher's years of employment, e, and hire cohort, c. Hire cohort is synonymous with the school year, t, when the teacher, j, was in her first year of employment; thus, $c = c(j) = t(j|e = 1)$. All teachers in a given hire cohort experience the same treatment(s) at the same point in time. Each hire cohort's actual treatments (or lack of treatments) were determined by Tennessee's evaluation policies.

To construct a null distribution, I iteratively randomly assign hire cohorts to different placebo treatment conditions. All other features of the quasi-experiment are kept constant across iterations: the number of treated and comparison hire cohorts in each effect estimate, each teacher's years of employment, etc. I repeat the procedure 500 times, and use the resulting *t*-statistics to form a null distribution for each effect estimate. The *p*-values reported in Table A1 are the proportion of *t*-statistics, in absolute value, which are equal to or greater than the actual estimate using actual treatment.

There is one cavate to the "all other features…constant" component. In Table 5, and Table A1 panel C, I split the sample based on who did and did not (and would have or would not have) earned tenure on time under the new rules. This split requires knowing a teacher's LOE score in $e = 5$. But the early hire cohorts in my data were in year $e \geq 6$ when the new evaluation measures began in 2012. For the randomization inference *p*-values in Table A1 panel C column 2, I split the sample based on each teacher's first observed LOE score. The "actual estimate" used to calculate the *p*-value follows the same procedure for splitting the sample.

A4. Effects of Future Performance Incentives, Additional Considerations

If a teacher wanted to improve her skills, in anticipation of future incentives, her efforts to improve would only succeed if she knew where to direct those efforts. To guide their improvement efforts, beginning in 2012, Tennessee's teachers had a new more-detailed classroom observation rubric. As described in Section 2, each teacher was observed multiple times per school year, and rated on 19 different teaching tasks. The rubric itself provided detailed descriptions of what a teacher should do in practice to score higher, for example, see Figure 1 on how to ask students questions. Prior (quasi-)experiments have also shown teacher value-added gains as a result of similar rubric scores and feedback (Taylor and Tyler 2012, Papay et al. 2020, Burgess, Rawal, and Taylor 2021, Hanno 2022, Briole and Maurin in-press).

The value-added gains, in Table 2, come in the first year of the new program, which may seem too quick for skill investments to improve performance. But the anticipation mechanisms have an entire school year to play out. Existing estimates suggest two additional weeks of class time could add 0.05σ or more to student achievement, without any change in teacher skills (Fitzpatrick, Grissmer, and Hastedt 2011, Aucejo and Romano 2016). Several related (quasi-)experiments also show teaching improvements in the first year (see citations in the previous paragraph).

What else might explain the value-added improvements shown in Table 2? First, for a small subsample of teachers, at most 15 percent, the local school district linked new pay-for-performance bonus incentives to the new evaluation scores in 2012 (see Appendix Section A2). In Appendix Table A1 column 3 I exclude these districts entirely and re-estimate. The treatment effect estimates remain essentially unchanged.

Second, a career concerns explanation is a possible alternative to the skill investment explanation. Career concerns can motivate greater current effort even without explicit current incentives (Fama 1980, Holmström 1999, Lazear and Oyer 2013). However, many career concerns considerations would be part of the counterfactual and differenced out in the diff-in-diff estimates in row 1 and 2 of Table 2 (and also in the triple-diff estimates in row 3). To explain the effect estimates, the new evaluation program would need to create new career concerns channels. One possibility is the following: A teacher may expect that her classroom observation ratings will be conducted by the same school principal year after year, and she may believe that her ratings in years 1–3 will affect her ratings in years 4–5 when they do count for earning tenure. Empirical evidence partly supports and partly contradicts that second belief (Ho and Kane 2013). However, the example applies to classroom observation ratings, which occur only a few days per year, and the effect estimates are measured in teachers' contributions to student test scores which accumulate over the entire year.

Third, the teacher's (employee's) problem described in Section 1 requires understanding the relationship between effort and output. Perhaps teachers increased their current production effort in an attempt to learn about the relationship between effort and their evaluation scores. That would still be an anticipation effect, and also be a kind of learning about the education production process, but not the kind of skill investment which would necessarily persist after the tenure incentives end.

A5. Effects When Performance Incentives Begin, Heterogeneous Effects

The small average improvement, 0.013σ, may well mask heterogeneity correlated with job performance. At the start of year 4, each teacher will have received a relevant signal: her evaluation score in year 3 and prior years. Some teachers will expect to score above the tenure cutoff even if they make no change between year 3 and 4. Other teachers will expect to miss the cutoff unless they increase their effort in year 4; the start of formal incentives is far more salient for this second type.

Among teachers who scored below the tenure cutoff in year 3 (LOE < 4), value-added increased by 0.077σ with the start of tenure incentives in year 4 (Appendix Table A6 panel A column 1). By contrast, value-added changed much less among teachers who scored above the cutoff in year 3 (LOE ≥ 4) and may have declined between year 3 and 4 (panel B column 1). The contrast, 0.077σ vs –0.018σ, is consistent with the differences between fourth-year teachers in the salience of the new tenure rules. However, there are alternative explanations that may explain the difference in estimates.

The first potential alternative explanation is mean reversion. The tenure cutoff is based on a teacher's Level of Effectiveness (LOE) score, which itself is a weighted average of several performance measures, including value-added scores (weighted 0.35, see Section 2). Consider a teacher whose *measured* value-added is below her own *true* value-added in year 3, because of measurement error. She will also be more likely to have LOE < 4 in year 3. And her *measured* value-added will be more likely to increase from year 3 to 4 by mean reversion. This contributes upward bias to the 0.077σ estimate, and the opposite scenario contributes downward bias to the –0.018σ estimate.[2] Two additional analyses provide further information about the mean reversion explanation.

First, Appendix Figure A3 provides some evidence against the mean reversion explanation. Figure A3 is an event study of the difference in value-added between the two groups—value-added for those scored LOE < 4 in year 3 minus value-added for those who scored LOE ≥ 4. The difference in value-added is stable

[2] As described in Section 2, the value-added scores in Tennessee's LOE calculation come from the SAS Institute. I do not use the SAS value-added scores in this paper's analysis, but the SAS methods and the specification in (2) will produce highly correlated measures of teacher value-added.

for two years leading up to the start of tenure incentives (that is, parallel pre-trends over $e = 2$ and $e = 3$). But that trend changes sharply when the tenure incentives begin in $e = 4$. Value-added improves sharply for teachers who had missed the tenure cutoff in $e = 3$, suggesting those teachers gave greater effort in $e = 4$ compared to $e = 3$.

Second, to avoid mean reversion in value-added, I repeat the same heterogeneity test but with a different year 3 measure. Table A6 panels A and B split the sample by year 3 LOE $<$ 4 and ≥ 4, respectively. Panels C and D split the sample by year 3 classroom observation rating $<$ 4 and ≥ 4, respectively. A teacher's rubric-based classroom observation rating is the most heavily weighted component of her LOE score (weighted 0.50). Thus, receiving an observation rating below the tenure cutoff ($<$ 4) in year 3 is a strong signal to the teacher that she should give more effort in year 4 if she wants to earn tenure. Still, earning tenure only requires LOE ≥ 4, which could be achieved with an observation rating $<$ 4. Observation ratings are a strong signal, but not as strong or direct as the LOE signal.

Among teachers whose observation ratings were below the cutoff ($<$ 4) in year 3, value-added increased by 0.018σ (Table A6 panel C column 1). Value-added increased by 0.005σ (panel D column 1) for those with observation ratings above the cutoff in year 3 (≥ 4). Neither of these point estimates is statistically significantly different from zero. However, the estimated gain for teachers below the cutoff is more than 3.5 times as large as for those above the cutoff. That is similar to the magnitude difference between panels A and B for LOE.

One solution which may seem available in this setting is not available. Recall the already-tenured never-incentive teachers hired in 2009 or earlier. After the new evaluation program began, the never-incentive teachers were scored with new performance measures but already had tenure (and it could not be taken away based on performance measures). Assume that I had a sample of third-year never-incentive teachers and their year 3 LOE scores. With that sample I could construct a triple-difference estimate. For example, (i) minus (ii), where (i) is the 0.077σ estimate for teachers who scored LOE < 4 in year 3 and who were subject to the new tenure rules (Table A6 panel A); and (ii) would be the corresponding $\hat{\delta}$ for teachers who scored LOE < 4 in year 3 but who were *not* subject to the new tenure rules. This triple-difference would difference out the average mean reversion because both (i) and (ii) are selected on LOE < 4. However, I cannot estimate (ii). Teachers hired in 2009 were in their fourth year ($e = 4$) in 2012 when the new performance measures began. Thus, I do not observe LOE in $e = 3$ for any teachers who are not subject to the new tenure rules.[3]

A second potential alternative explanation is the lack of a heterogeneous counterfactual. Both point estimates—0.077σ and –0.018σ (Table A6 panels A and B)—use the same comparison group and counterfactual estimate. I cannot split the comparison group based on LOE scores because those scores did not exist prior to the 2012 reforms. Thus, 0.077σ means that teachers who scored LOE < 4 in year 3 improved 0.077σ more between year 3 and 4 than the *average* teacher improvement between year 3 and 4 prior to the new evaluation program. And –0.018σ is also relative to the average comparison teacher. It seems quite plausible that—even absent the new evaluation program and absent mean reversion—teachers who were relatively low (high) performing in year 3 might improve faster (slower) than the average teacher improves between year 3 and 4 (Kraft and Papay 2014, Atteberry, Loeb, and Wyckoff 2015).

The third potential alternative explanation is attrition bias. In the years prior to Tennessee's new evaluation rules, 5.4 percent of teachers quit teaching after their

[3] Note that, while this triple-difference strategy for testing heterogeneity is not available for the "incentives begin" estimates in Table A6 where $(e - 1) = 3$ and $e = 4$, the same strategy is available for the "incentives end" estimates in Table 5 where $(e - 1) = 5$ and $e = 6$. I do observe never-incentive teachers whose LOE scores in $e = 5$ were below the tenure cutoff.

third year and another 20.3 percent switched jobs to a non-tested grade or subject, for a total attrition rate of 25.7 percent. As discussed in Section 5, that rate increased slightly to 26.1 percent among teachers subject to the new tenure rules (Table 4 panel A column 2 row 1).

Attrition rates were quite different between teachers who scored above and below the tenure cutoff in year 3. Among those who scored LOE < 4 in year 3, attrition was 31.9 percent or 6.2 points higher than the average in years before the new evaluation program (Table A6 panel A column 2). Among those who scored LOE ≥ 4 in year 3, attrition was 19.3 percent or 6.4 points lower (panel B column 2). The swing from +6.2 to –6.4 is quite large. If that 12.6 point difference was the heterogeneous effect of treatment on attrition, then it would create substantial scope for attrition bias to explain the difference between 0.077σ and –0.018σ. However, a causal interpretation of the difference in attrition rates is threatened by the same lack of a heterogeneous counterfactual as threatens the value-added differences.

Most intuitively, teachers who expect they will not improve enough to meet the tenure cutoff may be more likely to quit or switch jobs. Scoring LOE < 4 in year 3 may simply be correlated with a teacher's expectations about her performance trajectory, even if scoring LOE < 4 does not causally change her expectations or causally change her performance. If lower-growth teachers select out, then 0.077σ will be biased too large. Notably, most of the difference in attrition between the LOE < 4 and ≥ 4 groups is switching to a non-tested grade and subject, not a difference in quitting teaching. Switching to a non-tested job eliminates value-added from a teacher's LOE calculation, which suggests selection on value-added performance specifically.

However, that most intuitive selection pattern may not be the case. First, as discussed in Section 4.2, teachers with lower value-added scores in year $(e - 1)$ are more likely to attrit in year e. At the end of year 3, a teacher whose value-added score in year $(e - 1)$ is one standard deviation lower is 2.9 percentage points more likely to attrit in year e (Table A6 panel A or B column 3 row 2). But there is no change in that relationship for either group—LOE < 4 and LOE ≥ 4—when the new tenure rules begin, at least no statistically significant change (column 3 row 3). Further, as discussed in Section 4.2, value-added levels are likely negatively correlated with value-added growth.

Second, the effects of the negative signal—missing the tenure cutoff in year 3—may be non-monotonic. Some teachers may quit. Other teachers who expect they will not improve enough may remain in their jobs in year 4 (or years 4 and 5) even though they know they (may) have to leave eventually. This second type of teacher may reduce their effort and performance, contributing negative growth to the 0.077σ estimate average. Bell et al. (in-press) shows empirical evidence from Tennessee that, among early-career teachers who eventually quit at the end of year t, performance falls between $(t - 1)$ and t.

A6. Effects After Performance Incentives End, Attrition

A6.1 Attrition After Year 5

There is little scope for attrition bias in the 0.025σ triple-difference estimate. After year 5, newly-tenured teachers were 1.3 percentage points less likely to attrit than the already-tenured teachers who *would have* earned tenure under the new rules (Table 5 panel A column 2 row 3). That –1.3 difference is far from statistically significant (standard error 2.3 points). Still, perhaps the newly-tenured teachers were, in fact, less likely to attrit. Assume tenure incentives did increase retention, after year 5, but had no effect on value-added. For differential attrition to generate the 0.025σ estimate, the marginal retained teacher would have to be someone who's value-added grew substantially faster, between year 5 and 6, than the inframarginal teacher.

Why might a high performing, or fast improving, teacher choose to quit in the counterfactual, but choose to stay in her job because of the tenure incentives?

One possibility is that high-performing or fast-improving individuals prefer to work alongside other high-performing or fast-improving individuals. The new tenure rules were (purportedly) designed to select for higher performing teachers. A high-performing teacher might stay because the new tenure rules changed the composition of her coworkers for the better. However, empirically, teachers who fell below the new tenure cutoff where not more likely to leave. Consider the sample of teachers who did not earn tenure under the new rules or *would not have* earned tenure under the new rules. The bottom one-third of the performance distribution in Tennessee. Among these low-performing teachers, the triple-difference estimate for attrition is –0.8 percentage points (standard error 4.4, Table 5 panel B column 2 row 3). In other words, there was little change in one's coworkers that might boost the retention of high-performers.

A second possibility is an improvement in career concerns. High-performing or fast-improving teachers might be more optimistic about their future careers as teachers, and be more likely to stay in their jobs, because the new performance measures (more credibly) reveal they are high-performing or fast improving. However, value-added scores for individual teachers had been reported in Tennessee since the 1990s, many years before the new evaluation reforms in 2012. Additionally, in the triple-difference comparison, never-incentive teachers also were scored with the new performance measures and would have captured the same improvement in career concerns.

To fully explain the 0.025σ estimate, the marginally retained teacher would have to be improving very quickly over one school year (from year 5 to 6). Assume there was no effect of tenure incentives on value-added; all of the 0.025σ came from greater retention of teachers who's value-added would have increased anyway, but who were induced to remain in teaching by the new tenure rules. The –1.3 point difference is a reduction of 6.9 percent over the attrition rate of 18.8 points among would-have-earned teachers. For those 6.9 percent alone to generate the 0.025σ would require that average value-added growth, between year 5 and 6, be 0.362σ (= 0.025σ / 0.069) among the marginally retained. That is an improvement of roughly 1.5–3.5 standard deviations in the teacher value-added distribution. For comparison, the average Tennessee teacher improves 0.10–0.15σ over the first 10 years of her career (Bell et al. in-press).

Finally, perhaps treatment changed the causes of attrition, even if the levels of attrition were unchanged. The available data suggests this hypothesis cannot explain the 0.025σ effect. Higher performing teachers are less likely to attrit. In the years before the evaluation reforms in 2012, a teacher whose value-added score in year 5 was one standard deviation higher was 2.3 percentage points less likely to attrit after year 5 (Appendix Table A4 panel A column 2 row 2). That relationship is slightly stronger among teachers who successfully earned tenure under the new rules—on time, after year 5, by scoring LOE ≥ 4 in years 4 and 5. The coefficient increases from 2.3 percentage points to 2.7 percentage points, but the difference of 0.4 points is not statistically significant (column 2 row 3). The triple-difference is a little stronger still, with a difference of 0.9 points, though still far from statistically significant (Table A4 panel C column 2 row 2).

Perhaps the new tenure incentives did, in fact, strengthen the relationship between value-added levels in year 5 and the probability of attiring after year 5. But the threat to identification here is attrition correlated with value-added growth not levels. Prior studies of the returns to experience in teaching, though imperfect, suggest a negative correlation between value-added levels and growth among early-career teachers (Kraft and Papay 2014, Atteberry, Loeb, and Wyckoff 2015). Assume the new tenure incentives did cause more teachers to stay in their jobs in year 6. Stayers likely had higher value-added levels in year 5 than did leavers. Given the negative correlation between levels and growth, we would predict less growth in value-added among those induced to stay. In the end, the 0.025σ estimate may well be biased too small.

Both newly-tenured and would-have-earned teachers were more likely to stay in their jobs compared to the average teacher working in the years before Tennessee's evaluation reforms (5.7 and 4.4 percentage points, respectively, Table 5 panel A column 2 rows 1–2). Those higher retention rates could partly be the result of the new performance measures and feedback. But the differences could simply be explained by selection, without any effect of the new evaluation program.

Further evidence that the naïve diff-in-diff estimates are biased by selection: Appendix Table A4 panels A and B column 1 row 1 shows the naïve estimates for attrition: –5.7 and –4.4 (these are the same estimates as in Table 5 column 2). Table A4 column 2 shows the estimates controlling for value-added in year 5. Notice the main effects become much smaller, for example, –5.7 vs. –1.5 in panel A. The panel A column 2 estimate of –1.5 is the effect on attrition for teachers who just earned tenure under the new rules *and* who had average value-added scores in year 5. By construction value-added has mean zero. The panel B column 2 estimate of –0.3 is the effect for the would-have-earned teachers who had average value-added. Focusing on teachers with average value-added partly undoes the selection based on LOE scores.

A6.2 Cumulative Attrition Between Year 1 and 6

The estimates in Tables 5 and A4, and the discussion in Section A6.1, focus on attrition between year 5 and 6. Even if there is no differential attrition after year 5, the new tenure rules may have affected attrition before year 5. Imagine two teachers who are identical potential outcomes (potential performance), but the first is subject to the new tenure rules (and the second is not). Both teachers would earn tenure on time, after year 5, under the new tenure rules. Still, perhaps the first teacher is more likely to quit (or change jobs) sometime before year 5. For example, because anticipating the new tenure expectations, or the delay in tenure, create additional uncertainty which reduce the job's compensation (relative to the second teacher's compensation).

To create attrition bias, however, such differences in attrition must be correlated with a teacher's potential growth in value-added from year 5 to 6. Consider the 0.025σ estimated effect on value-added growth from year 5 to 6 for newly-tenured teachers (Table 5 panel A column 1 row 3). To create positive bias—to make 0.025σ too large—the teachers who leave prior to year 5 would have to be teachers whose year 5 to 6 growth would have been relatively small among the set of teachers who also (would have) scored above the new tenure cutoff.

The available evidence suggests the opposite pattern, though that evidence has limits. Roughly half of first-year teachers are still teaching—specifically, still teaching a tested grade and subject—five years later. The cumulative attrition rate, from year 1 to 6, is 60.3 percent among never-incentive teachers. The same cumulative rate is slightly higher, at 62.9 percent, among teachers subject to the new tenure rules. The difference of 2.6 points is not statistically significantly different from zero (standard error 2.0, Appendix Table A6 column 2).

The 2.6 percentage point estimate, and all others in Table A6, are triple-difference estimates, as the triple-difference estimates are defined in this paper.[4] I use the same difference-in-differences estimation strategy in equation (1). The only change required is that all references to $(e-1)$ in equation (1) and elsewhere become $(e-5)$. The estimation sample is all teachers, j, observed in the data

[4] Given the years covered by the available data, there is no comparison group, as there is elsewhere in the paper. In this case, the comparison group would include be teachers who (i) reached year $e=6$ in 2011 or earlier, before the reforms in 2012; but who also (ii) are observed in the data in $e=1$. The data required to estimate value-added begin in 2008, and teachers for whom $e=1$ reach $e=6$ in 2013.

Nevertheless, 0.024 and the other estimates in Appendix Table A6 are triple-difference estimates, as triple-difference is defined in this paper. Let A_C be the unknown cumulative attrition rate in the comparison group (if the data went back further in time A_C could be estimated). Let A_I be the rate for teachers who were subject to the new tenure incentives, and A_N the rate for never-incentive teachers. Then the triple-difference is $(A_I-A_C)-(A_N-A_C)=(A_I-A_N)$. Though without A_C the final triple-difference estimate is perhaps less precisely estimated.

during their first year teaching, $e = 1$. The outcome is an indicator $= 1$ if teacher j is no longer working in a tested grade and subject in $e = 6$.

Note that the 2.6 point estimate is the difference in cumulative attrition for all teachers, not just the newly-tenured. Here I cannot split the sample into the two groups—(i) newly-tenured or would have earned tenure, and (ii) not-yet-tenured or would not have earned tenure—as I do in Section 6 and Table 5. That split can only be done with teachers who continue teaching through year 5. Teachers who quit do not have LOE scores. Teachers who change jobs, to non-tested grades and subjects, have an LOE score; but that LOE score is substantively different, with different components and weights. Moreover, I cannot split the sample by (or examine heterogeneity related to) any other measure of performance, if that measure occurred after year 1 of a teacher's career.

I can examine how attrition by year 6 is related to value-added in year 1. Performance in year 1 is a long way from performance in year 4 and 5, when it counts for tenure. Nevertheless, year 1 performance should be a strong predictor of teachers who (would have) earned tenure under the new rules.

Teachers with higher value-added in year 1 are more likely to still be teaching in year 6 (less likely to attrit). This relationship is similar to the value-added-attrition relationship reported elsewhere in the paper. Here, over a longer time horizon, the relationship is stronger. Among the never-incentive sample, a teacher who is one standard deviation higher in the value-added distribution is about 5.6 percentage points less likely to attrit (Table A6 column 3). Among teachers subject to the new tenure rules, the estimate is 2.3 percentage points less likely to attrit. The estimated effect of the new tenure rules is thus 3.3 points—a weakening of relationship between value-added and attrition. In other words, after the new tenure incentives began, the marginal retained teacher was a lower-performing teacher. However, the estimated effect of 3.3 points is not statistically significant (standard error 2.2 points, Table A6 column 3 row 3).

A7. Effects After Performance Incentives End, Mean Reversion

The estimates in Table 5 split teachers into groups based partly on prior value-added. This raises questions about mean reversion as a threat to the substantive inferences discussed in Section 6. However, mean reversion is unlikely to threaten the triple-difference estimates, and would bias against finding effects for the newly-tenured teachers.

First consider the naïve diff-in-diff estimate of 0.069σ (Table 5 panel B column 1 row 1). That 0.069σ is the difference in value-added growth, from year 5 to 6, between two groups: (i) teachers who were subject to the new rules, but who did not earn tenure after year 5, and (ii) the average teacher in the years before Tennessee's evaluation reforms. Mean reversion very likely biases that naïve estimate of 0.069σ, making it too large.[5]

Group (i) is defined by having LOE scores below the new tenure cutoff. LOE score is a weighted average of several performance measures, including, notably, value-added scores (weighted 0.35, see Section 2).[6] The value-added component opens the door to mean reversion. Consider a teacher whose *measured* value-added is below her own *true* value-added in year 5, because of measurement error. She will also be more likely to have LOE < 4 in year 5 and fail to earn tenure on time. And her *measured* value-added will be more likely to increase from year 5 to 6 by mean reversion.[7]

However, mean reversion (likely) does not bias the triple-difference estimate, –0.016σ (Table 5 panel B column 1 row 3). That –0.016σ estimate is the

[5] Additionally, for reasons discussed earlier, this naïve diff-in-diff lacks a convincing counterfactual even if there were no mean reversion concern.

[6] As described in Section 2, the value-added scores in Tennessee's LOE calculation come from the SAS Institute. I do not use the SAS value-added scores in this paper's analysis, but the SAS methods and the specification in (2) will produce highly correlated measures of teacher value-added.

[7] This potential threat from mean reversion is attenuated somewhat because LOE is only partly determined by value-added, with a weight of 0.35.

difference between: (i) teachers who were subject to the new rules, but who did not earn tenure after year 5, and (iii) teachers who were *not* subject to the new rules, but who *would not have* earned tenure after year 5. Both (i) and (iii) are selected in the same way based on prior LOE performance scores. The naïve 0.069σ estimate for group (i) is likely biased by mean reversion. But the same bias also inflates the naïve 0.086σ estimate for group (iii) (Table 5 panel B column 1 row 2).

As a robustness test, I repeat the paper's estimation strategy with one change. Instead of splitting teachers into groups based on their prior LOE scores, I group teachers based on their classroom observation ratings in year 5. LOE is a weighted average of both value-added (weighted 0.35) and observation ratings (0.50) plus other measures (0.15). Using observation ratings alone avoids the mean reversion threat that arises from value-added being both the outcome and a contributor to LOE.

Appendix Table A7 panel B reports estimates for teachers whose year 5 classroom observation rating was < 4. As we might expect with mean reversion, the naïve diff-in-diff estimates shrink: from 0.069σ to 0.036σ and from 0.086σ to 0.024σ (compare Table 5 panel B to Table A7 panel B). However, estimates using observation ratings might be smaller even without bias from mean reversion. First, having "an observation rating < 4" is correlated with "(would have) failed to earn tenure under the new rules," but the correlation is certainly not perfect. Some teachers with an observation rating < 4 may feel little pressure to improve because their value-added scores bring their LOE average well above 4. Second, observation ratings do have measurement error too, often just as much as value-added (Taylor 2023).

The triple-difference estimate remains quite similar in magnitude. The sign changes, –0.016σ compared to 0.012σ, but both have similar precision, and I cannot reject the null that they are equal. In other words, the triple-difference is more robust to the threat from mean reversion, as we would expect given its construction.

Finally, consider (potential) mean reversion in the estimates for newly-tenured teachers. First, and most importantly, mean reversion would likely bias against finding gains for the newly-tenured teachers; that is, the 0.025σ estimate would be biased too small. Newly-tenured teachers are positively selected on prior outcomes. They are more likely to have had a positive measurement error shock in their baseline value-added and LOE scores (opposite of the negative shock for teachers who failed to earn tenure), and mean reversion would reduce value-added from year 5 to 6. Second, Table A7 panel A reports the robustness test where teachers are selected by having observation ratings ≥ 4 in year 5. The alternative triple-difference estimate is quite similar, 0.028σ, compared to the main estimate, 0.025σ (compare Table A7 panel A column 1 row 3 to Table 5 panel A column 1 row 3).

References

Ballou, D., Canon, K., Ehlert, M., Wu, W. W., Doan, S., Taylor, L., & Springer, M. (2016). *Final Evaluation Report Tennessee's Strategic Compensation Programs: Findings on Implementation and Impact 2010-2016.* Peabody College, Vanderbilt University. Retrieved on July 31, 2021 from https://peabody.vanderbilt.edu/TERA/teachers-and-leaders-publications.php

Ballou, D., Barcy, K., Canon, K., Ehlert, M., Gronberg, T., Gurwit, M., Jansen, D., Lewis, J., Li, J., Palmer, S., Parsons, E., Stahlheber, S., & Taylor, L. (2015). *Evaluation of Tennessee's Strategic Compensation Programs: Interim Findings on Design, Implementation, and Impact in Year 2 (2012-13).* Peabody College, Vanderbilt University. Retrieved on July 31, 2021 from https://peabody.vanderbilt.edu/TERA/teachers-and-leaders-publications.php

Canon, K., Greenslate, C., Lewis, J., Merchant, K., & Springer, M. (2012). *Evaluation Report Tennessee's Strategic Compensation Programs: Interim Findings on Development, Design, and Implementation.* Peabody College, Vanderbilt University. Retrieved on July 31, 2021 from https://peabody.vanderbilt.edu/TERA/teachers-and-leaders-publications.php

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

Ho, A. D., & Kane, T. J. (2013). *The Reliability of Classroom Observations by School Personnel.* Seattle, WA: Bill & Melinda Gates Foundation.

Springer, M. G., Swain, W. A., & Rodriguez, L. A. (2016). "Effective teacher retention bonuses: Evidence from Tennessee." *Educational Evaluation and Policy Analysis, 38* (2), 199-221.

Tennessee Department of Education. (n.d.). *2014-15 Differentiated Pay Plan Summary.* Downloaded September 10, 2022 from https://www.tn.gov/content/dam/tn/education/educators/diff_pay/diff_pay_summary_report.pdf.

All other references as listed in the main text.