ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-686

Employee evaluation and skill investments: Evidence from public school teachers

Eric S. Taylor
Harvard University and NBER

When an employee expects repeated evaluation and performance incentives over time, the potential future rewards create an incentive to invest in building relevant skills. Because new skills benefit job performance, the effects of an evaluation program can persist after the rewards end or even anticipate the start of rewards. I test for persistence and anticipation effects, along with more conventional predictions, using a quasi-experiment in Tennessee schools. Performance improves with new evaluation measures, but gains are larger when the teacher expects future rewards linked to future scores. Performance rises further when incentives start and remains higher even after incentives end.

VERSION: December 2022

Suggested citation: Taylor, Eric S.. (2022). 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

November 2022

When an employee expects repeated evaluation and performance incentives over time, the potential future rewards create an incentive to invest in building relevant skills. Because new skills benefit job performance, the effects of an evaluation program can persist after the rewards end or even anticipate the start of rewards. I test for persistence and anticipation effects, along with more conventional predictions, using a quasi-experiment in Tennessee schools. Performance improves with new evaluation measures, but gains are larger when the teacher expects future rewards linked to future scores. Performance rises further when incentives start and remains higher even after 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. Seminar participants at AEFP, Arizona State, Brown, CESifo, Harvard, and the Northeast Economics of Education Workshop provided feedback on earlier drafts. The Spencer Foundation provided 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 utility to their employer's success, thus inducing more effort at work or a better allocation of effort across tasks. In the literature this agency-theory view of employee evaluation and incentives dates to Holmstrom (1979) among others, and today we have empirical tests from many occupations and sectors.[1] However, the literature largely ignores potential intertemporal features of the problem.

This paper considers one intertemporal feature: When the employee expects repeated evaluation and performance incentives over time, the potential future rewards create an incentive to invest in human capital in the present. In many ways this is just a special case of the familiar human capital investment models which began with Becker (1962). However, unlike, say, the returns on earning a college degree, evaluation rewards can end abruptly. Even if the evaluation rewards end, the employee retains the newly developed skills and the higher job performance those skills produce. In other words, the effects of evaluation on performance can persist even after incentives end. Effects can also anticipate a future evaluation. The paper begins with predictions from a simple framework combining familiar features from agency theory and human capital investment.

To test the predictions, I study public school teachers in Tennessee. The state of Tennessee adopted new teacher evaluation rules, beginning in 2012, which required new performance measures and new tenure incentives linked to those measures.[2] I estimate the effect of various policy features—the measures, the start of incentives, the end of incentives—on teacher contributions to student achievement test scores.

[1] For general reviews see Oyer and Schaefer (2011) and Gibbons and Roberts (2013). For teachers specifically see Neal (2011).

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

The potential skill development effects of evaluation are of particular interest in the case of school teachers. First, prior empirical evidence from teachers has been inconsistent with conventional theoretical predictions. A basic prediction of agency models is that changes in employee effort, and thus performance, should coincide with changes in incentives. Contrary to that prediction, studies of teachers in the United States and France find improvements in job performance which persist years after the evaluation and incentives end (Taylor and Tyler 2012, Briole and Maurin in-press). Both papers argue that teachers are motivated agents (Dixit 2002), thus disposed to use evaluation feedback to improve their teaching skills, and that improved skills explain the persistent higher performance. By contrast, the framework presented in this paper does not rely on a motivated-agents assumption; any employee could be motivated to invest in skills by the promise of future evaluation rewards.

More generally, 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.[3] 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). Teacher pay-for-performance schemes have had little success in practice, at least in the United States (for a review see Neal 2011).

[3] 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) and Jackson, Rockoff, and Staiger (2014).

Tennessee's new evaluation program improved teacher performance, and the pattern of effects is consistent with teachers making investments in their own skills in response to the program. This paper has four main results. First, performance improved broadly when the new performance measures were introduced. In the program's first year, 2012, the average early-career teacher's contribution to student achievement ("teacher value added") increased by 0.032σ (σ = student test-score standard deviations) more than it would have absent the new evaluation program.

To estimate this effect, and the others in this paper, I use a difference-in-differences strategy. I apply an estimator proposed in de Chaisemartin and D'Haultfœuille (2020) which, among other things, focuses on observations most proximate in time to a change in treatment status. The first difference is the change in a teacher's performance between year $(e-1)$ and year e of employment: improvement between the first and second year of teaching, second and third, third and fourth, etc. The second difference is between groups of teachers who experience different evaluation "treatments" because they began their careers at different times. For the 0.032σ estimate specifically, the comparison group is teachers who had already reached e years of experience before the new program began in 2012, and the treated group is teachers who were at $(e-1)$ in 2011 and e in 2012. I focus on early-career teachers, $e \in \{1,2,\ldots,7\}$, because the new evaluation incentives were mainly about earning tenure, as I explain further shortly. In short, early-career performance grew 0.032σ faster year over year than it had before the program.

Second, performance gains were twice as large for teachers expecting future incentives linked to their future evaluation scores—anticipation effects. In the program's first year all teachers were scored with the new classroom observation rubric and other performance measures, but there were no incentives linked to those measures for any teacher. However, there would soon be incentives for some

teachers. For those hired in 2010 or later, earning tenure would require scoring above a cutoff, about the 33rd percentile of teachers, in both their fourth and fifth years of employment. The 0.032σ overall effect averages across both these future-incentive teachers, 0.047σ, and already tenured teachers who would never have incentives attached to their scores, 0.024σ. The 0.024σ effect for never-incentive teachers is consistent with the motivated-agents hypothesis; they improved faster with the new performance measures than without. But the even larger 0.047σ effect for future-incentive teachers is consistent with teachers developing skills in anticipation of the rewards in future school years.

Third, performance rises further when the tenure incentives begin in a teacher's fourth year of employment. This is consistent with the conventional prediction: changes in employee effort, and thus performance, should coincide with changes in incentives. The anticipation effects, described in the previous paragraph, do not completely crowd out conventional effects of performance incentives. Moreover, teachers who scored below the tenure cutoff in year 3 seem to respond more strongly to the start of incentives in year 4, consistent with the signal they received in year 3.

Fourth, teacher performance continues at higher levels after the performance incentives end. These persistent effects are consistent with skill improvements caused by the evaluation program. The key comparison is between two groups of teachers: (a) Teachers who scored above the cutoff in year 5, earned tenure, and no longer had any current or future incentive from the evaluation program. (b) Teachers who already had tenure before the new program, never had incentives linked to their evaluation scores, but who also scored above the new tenure cutoff in year 5. Between years 5 and 6 group (a)'s performance grew 0.024σ faster than group (b)'s, even though group (a)'s performance incentives ended after year 5. Moreover, group (b)'s performance was not different from the average teacher in the pre-program comparison group. If feedback and motivated agents

alone were sufficient for evaluation-induced skill improvements, then groups (a) and (b) would have similar performance trajectories, and similarly outperformed the comparison group. Additionally, if teachers simply increased their effort when it counted for earning tenure, we would predict a decline in performance between year 5 and 6, but I can reject declines larger than -0.001σ.

Claims about causation require a parallel trends style assumption: Absent the new evaluation program, teacher performance would have improved with experience at the same rate observed in cohorts prior to the new program. Returns to experience are a first-order feature of teacher performance (Rockoff 2004, Jackson, Staiger, and Rockoff 2014), but growth with experience is not a threat to identification. The assumption only requires that the rate of return to experience is not changing over cohorts of teachers. Consistent with that assumption, I show evidence that cohorts of new hires in Tennessee are not getting better or worse over time, as measured by their first-year value-added scores. I also show the time series over cohorts of estimated returns to experience, which is flat in the years before the new program.

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), Duflo, Dupas, and Kremer (2011), Dee and Wyckoff (2015), Deming et al. (2016), Macartney (2016), Aucejo, Romano, and Taylor (2022), among others. In contemporaneous work, Dinerstein and Opper (2022) and Ng (2022) study teachers' response to performance-based tenure rules. The literature spans a variety of performance measures and incentives, including monetary bonuses and dismissal threats (see reviews by Neal 2011, Jackson, Rockoff, and Staiger 2014). Though, as with other sectors and occupations, the empirical tests focus on the conventional prediction: changes in performance should coincide temporally with changes in incentives. This paper adds tests for anticipation effects and persistence effects.

First, the anticipation test is, to my knowledge, a novel contribution. Teacher effort anticipating future incentives is rational in a framework that combines familiar agency-theory and human capital investment models. And the potential for anticipation effects suggests some existing estimates may understate the total effect of performance measures and linked incentives.

Second, empirical tests for persistent effects are also still rare in the literature. The notable examples are the Taylor and Tyler (2012) and Briole and Maurin (in-press) papers mentioned already. The novel feature of the current paper is variation in incentives between teachers, which allows me to test the motivated-agent hypothesis raised by these two prior papers. The persistence prediction here does not rely on motivated agents, and thus may have broader applicability across occupations and sectors. Griffith and Neely (2009) find similar persistent effects, though noisily estimated, among employees in retail sales.

While this paper and others find performance improvements as a result of incentives, Rockoff et al. (2012) and Burgess, Rawal, and Taylor (2021) are examples where teacher performance improved without formal incentives attached to evaluation scores. This paper also adds to this small "without incentives" evidence; recall that the already-tenured never-incentive teachers also improved when the new program began. The Tennessee results leave open the motivated-agents mechanism, even if that mechanism cannot fully explain the results presented here.

To begin, in Section 1, I describe a simple model combining familiar agency theory and human capital investment features, which generates the persistence and anticipation predictions. Section 2 details the empirical setting, including the Tennessee evaluation program's performance measures and linked incentives. The identification strategy and estimated effects of evaluation are in Sections 3 and 4 respectively. Section 5 concludes.

1. Theory and Predictions

When employees expect repeated evaluation and performance incentives over time, the repetition can create an incentive to invest in human capital. The return on that investment is (partly) the rewards attached to future evaluation scores. In many ways this is just a special case of more general human capital investment models—well-established ideas dating back to Becker (1962), Mincer (1962), and Ben-Porath (1967). What distinguishes the employee evaluation case is, first, the returns are more likely to change discontinuously over time. Performance incentive systems can start or end while the employee remains in the same job. For example, contrast a specific pay for performance program with the, much more stable, labor market returns to a college education. Second, the returns are attached to specific performance measures, but any particular measure may be weakly related to the employee's general labor market value.

In this section I describe a simple model to highlight the interaction between employee evaluation and human capital investment. I combine features of human capital investment models, especially the intertemporal features, with the features of agency models typical in the study of employee evaluation and performance incentives.[4] The agency models also have a long history in the study of employee evaluation, dating to Holmstrom (1979) among others. More recent examples include Lazear (2000), Baker (2002), and, specific to teachers, Barlevy and Neal (2012).

Consider an employee whose work in period t contributes $q_t = Q(e_t, s_t)$ to the firm's value or other organizational objectives. That contribution depends on

[4] There are temporal features in Holmstrom and Milgrom (1987), but the implications are about how frequently to measure and provide rewards. In practice there will be some periodicity, e.g., evaluation and incentives once per year, and that repetition is the temporal feature I focus on here.

both the employee's level of skill, s_t, and on the employee's chosen level of effort, e_t. Assume the employee's job-related utility can be described by:

$$u_t = w_t + U_t[Q(e_t, s_t)] - C(e_t).$$

Here utility depends on base salary, w_t, and the employee's cost of effort, $C(e_t)$. Utility may also depend on the employee's contribution to the organization's objectives, $U_t[Q(e_t, s_t)]$. The function U can include explicit performance incentives offered by the firm, career concerns, the employee's own intrinsic value of their work, or some combination. Given the paper's empirical setting, I simplify the theoretical discussion in this section by focusing on explicit performance incentives (or rewards) offered by the employer. The intrinsic value of work is an important consideration when studying school teachers and other motivated-agent occupations (Dixit 2002). However, including or excluding "intrinsic value" from U does not change the basic predictions below, as long as intrinsic and extrinsic rewards are additively separable in U.[5]

The employee's problem is to choose work effort over time, $\mathbf{e} = (e_0, e_1, \dots, e_T)'$, to maximize utility, $\max_{\mathbf{e}} \sum_{t=0}^{T} u_t \frac{1}{(1+r)^t}$. So far, however, this problem does not clearly suggest any intertemporal tradeoffs.

One potential tradeoff arises through the employee's skills. Assume that skills, s_t, can be increased by investing effort, f_t, in developing one's own skills (synonymously, investing in human capital production):

$$s_t = S(f_{t-1}, s_{t-1}).$$

Then the employee's problem becomes

$$\max_{\mathbf{e},\mathbf{f}} \sum_{t=0}^{T} \{w_t + U_t[Q(e_t, S(f_{t-1}, s_{t-1}))] - C(e_t + f_t)\} \frac{1}{(1+r)^t}. \qquad (1)$$

[5] Moreover, the use of extrinsic rewards by the firm, or any positive effects on job performance empirically, do not require that the employee has zero intrinsic motivation. The cost of effort, C, is (likely) increasing and convex, $C', C'' > 0$. Thus the employee will dislike work and prefer leisure at some margin. Extrinsic rewards may move that margin further out (Lazear and Oyer 2013).

The new features in (1) are familiar features of human capital investment models. First, the costs of skill investments, $C(f_t)$, are paid in the present (period t), but the benefits, $U_{t+1}[q_{t+1}]$, are only realized in the future (period $t+1$ and beyond). Second, effort given to skill improvement, f_t, competes with effort given to current production, e_t. Even if there is no binding constraint on total effort, $(e_t + f_t)$, causing a mechanical tradeoff, the cost of effort is (likely) convex, creating a tradeoff between e_t and f_t at the margin.[6,7]

The optimal allocation of effort across current production, e_t, and skill improvement, f_t, will depend on how the employee expects the rewards, U_t, to change over time. First, consider the two-period case where (1) simplifies to

$$\max_{e_0, f_0, e_1} \{w_0 + U_0[Q(e_0, s_0)] - C(e_0 + f_0)\} + \{w_1 + U_1[Q(e_1, S(f_0, s_0))] - C(e_1)\}d_1 \quad (2)$$

where $d_t = \frac{1}{(1+r)^t}$. Then the optimal ratio of skill-development effort to current-production effort in period $t = 0$ is:

$$\left(\frac{f_0}{e_0}\right)^* = \frac{\frac{\partial U_1}{\partial f_0} d_1}{\frac{\partial U_0}{\partial e_0}} \quad (3)$$

which sets the ratio of efforts equal to marginal rate of substitution of utility in present value terms. Further, (3) implies that $\left(\frac{f_0}{e_0}\right)^*$ is increasing in $\frac{\partial U_1}{\partial Q} / \frac{\partial U_0}{\partial Q}$, other things constant. For example, relevant to this paper's empirical setting, if the

[6] For some specific skill improvements there may be little or no tradeoff. For example, an employee may become more efficient at completing some task simply through repeating the task over and over in the normal course of work. However, even other common examples labeled "learning by doing" involve effort by the employee that does not contribute to current production, Q, e.g., reflecting on past choices and outcomes, and planning for what to try differently in the future.

[7] Evaluation rewards incentivize investment in skills relevant to succeeding in the evaluation. Those skills may be general or firm-specific in the Becker (1962) sense. If firm-specific, the promised future evaluation rewards are similar to promised future wage increases, and the potential for the firm to back out of the promise should weaken the skill-investment incentives to some degree.

employer announces a new reward for performance will begin in period 1, the employee will give (relatively) more effort to skill development in period 0.

The skill-investment incentive grows when the returns will continue for multiple future periods. Return to the multi-period problem in (1). Assume the employee expects the firm's evaluation program and its performance incentives will continue until period $V \leq T$. Then expression (3) becomes

$$\left(\frac{f_t}{e_t}\right)^* = \frac{\sum_{v=t+1}^{V} \frac{\partial U_v}{\partial f_t} d_v}{\frac{\partial U_t}{\partial e_t}} \tag{4}$$

The numerator is now a stream of benefits over time, i.e., over $n = [V - (t+1)]$ periods. As n grows so does the incentive to invest (relatively more) effort in skill development at time t.[8]

While expression (4) shows the basic tradeoff between current-production effort, e_t, and skill-development effort, f_t, there are some additional considerations worth mentioning. First, skills can decay or depreciate. Thus, the marginal utility of a skill investment in period t may shrink somewhat over time; in (4) this would show up as $\frac{\partial U_v}{\partial f_t}$ decreasing in v even if the performance rewards offered by the firm were not changing. Second, new skills build on existing skills. Thus, current skill investments likely increase the returns to future skill investments; over three periods $\frac{\partial^2 U_{t+2}}{\partial f_{t+1} \partial f_t} > 0$. Third, skills and production effort are (typically) complements in the production functions represented by $Q(e_t, s_t)$. Effort given to skill development now, f_t, increases the marginal utility of effort given to production in

[8] Looking at expression (4), one might imagine examples where the expected stream of returns on skill investments is so large that little to no effort would be given to current production. In a sense, this is one of the main points in Ben-Porath (1967) about investing in skills generally. However, in practice, firms are likely to construct U_t such that some minimum level of current production is required to remain employed, as in Lazear (2000).

the future, e_{t+1}, $\frac{\partial^2 U_{t+1}}{\partial e_{t+1} \partial f_t} > 0$. Thus, as time approaches V—the expected end of the evaluation program—the value of skill investments shrinks for two reasons: less time to capture any returns, and effort given to current production becomes more valuable because of prior skill investments.[9]

The presentation above assumes U_t is a differentiable function. Some performance incentive programs fit that assumption, like piece rate bonuses, but others do not.[10] Nevertheless, even with more-complex U_t the basic tradeoff and solution remain: choose the ratio of skill-development effort and current-production effort to match the ratio of returns on those efforts. In this paper's empirical setting, the new evaluation rewards introduced by Tennessee create a discontinuity in U_{t^*}. As detailed in section 2, teachers who score above a cutoff in year t^* earn tenure, with job security and thus an increase in expected future earnings. Teachers do have an incentive to give more current-production effort, e_t, in year t^*; and, because of the discontinuity, that incentive is increasing in their uncertainty about passing the cutoff. Still, in the years before t^*, teachers also have an incentive to give more effort to skill development, f_t. The skill-investment incentive is also increasing in uncertainty about passing, but that uncertainty is likely larger when t^* is further out in the future.

In the remainder of the paper, I test three predictions from this framework. The first is a conventional prediction about evaluation incentives, not unique to this

[9] Additionally, in equation (1) time periods are linearly separable. This excludes production processes where output in one period, q_t, depends on output in some prior period, $q_{t'<t}$. Macartney (2016) documents an example where schools (thus teams of teachers) shift output across school years in response to school evaluation. Allowing for intertemporal production relationships would make the denominator in (4) a function of future periods as well, but the tradeoff between production and skill investment remains. In this paper's empirical analysis the main outcome is teacher value added, a teacher's contribution to student achievement conditional on the contributions of prior teachers among other things.

[10] For a reviews of performance incentives for teachers see Neal (2011) and Jackson, Rockoff, and Staiger (2014); and more generally Oyer and Schaefer (2011).

framework but still an important feature. *Prediction 1: Conventional Effects*—An employee's current performance responds to current incentives. Consider an evaluation program where the rewards for performance in period t are greater than in period $t-1$. In expression (4)'s terms $\frac{\partial U_t}{\partial e_t} > \frac{\partial U_{t-1}}{\partial e_{t-1}}$; and thus, other things equal, the employee should choose $e_t > e_{t-1}$ producing $q_t > q_{t-1}$.

Prediction 2: Persistent Effects—Improvements in employee performance will (partially) persist after any evaluation performance incentives end, because skills persist. A corollary of the conventional prediction is that when evaluation rewards end, so too will any direct effect on employee performance. In other words, if $\frac{\partial U_t}{\partial Q} = 0$ (or falls sharply) after some point in time, then the employee will no longer have an evaluation incentive for greater current-production effort, e_t, or skill-investment effort, f_t. That conventional prediction remains true in this framework. However, the persistent-effects prediction is not about current effort but about past effort to build skills. Because skills persist over time (to some degree), an evaluation program's past positive effects on skills will increase future performance even after the evaluation rewards end.

Consider a three-period version of the problem in (1)

$$\begin{aligned}\max_{e_0,f_0,e_1,f_1,e_2} \quad & \{w_0 + U_0[Q(e_0, s_0)] - C(e_0 + f_0)\} \\ & + \{w_1 + U_1[Q(e_1, S(f_0, s_0))] - C(e_1 + f_1)\}d_1 \\ & + \{w_2 + U_2[Q(e_2, S(f_1, S(f_0, s_0)))] - C(e_2)\}d_2\end{aligned} \tag{5}$$

Assume the employee knows there will be no evaluation rewards in period $t = 2$. The solution in period $t = 0$ is the same as shown in (3) above. The employee invests some effort in skills, f_0, motivated by the expected return in $t = 1$, $\frac{\partial U_1}{\partial f_0} d_1$. When period $t = 2$ arrives, the employees' skills are higher than they would have been without the evaluation rewards in $t = 1$. In period $t = 2$ there is no evaluation

reward linked to job performance, $\frac{\partial U_2}{\partial Q} < \frac{\partial U_1}{\partial Q} = \frac{\partial U_0}{\partial Q}$, and thus no incentive to give greater current-production effort, e_2. Nevertheless, the employee's performance in $t = 2$ is higher because of the evaluation program; the evaluation program increased f_0 and thus $Q\big(e_2, S(f_0, s_0)\big) > Q(e_2, s_0)$.[11]

This prediction does not require that the performance improvements will persist fully or forever. First, skills can depreciate over time. That potential depreciation has been left implicit in the function S. In the three-period example, perhaps the skills gained in $t = 0$ have decayed somewhat by $t = 2$, thus weakening the evaluation program's effects on performance in $t = 2$. As skill depreciation accumulates over time, evaluation effects will weaken more and more. Dinerstein, Megalokonomou, and Yannelis (in-press) 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. Second, the employee may reduce their current-production effort, e_t, after evaluation ends, and not just because the rewards have ended. With the benefit of improved skills, the employee's effort will be more productive, $\frac{\partial^2 Q}{\partial e_t \partial s_t} > 0$. Thus, compared to the counterfactual where the evaluation program never occurred, a rational employee could reduce current-production effort and still have higher performance. But potentially lower performance than if the evaluation program had continued.

Prediction 3: Anticipation Effects—A new evaluation program can improve performance before its performance incentive rewards begin, because employee skill investments can anticipate the future rewards. This prediction is likely less empirically relevant than the first two; it requires some amount of time lag between

[11] This result does not require the assumption that the employee anticipates no returns in $t = 2$, though that assumption is effectively true for the empirical setting in this paper. If the employee believes the evaluation rewards will continue, then they will invest more in skill development in $t = 0$ by expression (4) and also invest in skills in $t = 1$ expecting, incorrectly, a return in $t = 2$.

the announcement of the program and the start of rewards. Still, there are examples from teacher evaluation including the setting for this paper, as well as Taylor and Tyler (2012) and Briole and Maurin (in-press).[12]

Return to the three-period case shown in (5). But now assume the employee knows evaluation rewards will not begin until period $t = 2$. In period $t = 0$ the employee invests some effort in skills, f_0, motivated by the expected return in $t = 2$, $\frac{\partial U_2}{\partial f_0} d_2$.[13] Then in period $t = 1$ performance is higher because of the increase in skills caused by evaluation program; the evaluation program increased f_0 and thus $Q\big(e_1, S(f_0, s_0)\big) > Q(e_1, s_0)$. Performance is higher even though there is no evaluation reward for performance in $t = 1$.[14]

The discussion to this point has focused on how performance incentives can increase the return on investments in skills. But performance measures can also reduce the cost of investments in skills, by reducing the effort required of the teacher herself, f_t. Evaluation measures create new information: feedback about an individual's current performance, often with comparison to coworkers' performance. The cost of creating that new information is borne largely by the employer; absent an evaluation program, the employee would be left to self-

[12] An evaluation program might also affect average employee performance through selection into (out of) the workforce, and such selection effects might also lead the start of rewards. Selection effects are not the focus of this paper. In the empirical analysis, however, my identification strategy addresses potential selection.

[13] The employee will make skill investments in both period $t = 0$ and $t = 1$. In addition to balancing the ratio of current effort and skill effort, $\left(\frac{f_0}{e_0}\right)^* = \frac{\frac{\partial U_2}{\partial f_0} d_2}{\frac{\partial U_0}{\partial e_0}}$, the employee should also choose a balance of skill effort over time, $\left(\frac{f_0}{f_1}\right)^* = \frac{\frac{\partial U_2}{\partial f_0}\frac{\partial C_0}{\partial f_0}}{\frac{\partial U_2}{\partial f_1}\frac{\partial C_1}{\partial f_1}} \cong \frac{\frac{\partial U_2}{\partial f_0}}{\frac{\partial U_2}{\partial f_1}}$. The incentive to invest in period $t = 0$ is stronger if we make the plausible assumption that skill investments are complementary: $\frac{\partial^2 S_2}{\partial f_1 \partial f_0} > 0$.

[14] Anticipatory skill investments could also plausibly affect performance in period $t = 0$. If each t period is long enough, skill investments early in $t = 0$ might affect performance in $t = 0$. In this paper's empirical setting, time periods are school years, and teachers knew about the new program at the very start of the $t = 0$ school year.

assessment and data gathering on coworkers. Additionally, the new information can make skill investments more efficient by directing the employee's effort toward specific skills. Reducing the costs of skill investments reinforces the anticipation and persistence predictions.

Finally, the reduction in costs may (partly) explain why introducing new performance measures can improve teacher performance even without linking performance incentives to those measures. Assume teachers are motivated agents who intrinsically value their contribution the success of their students, and thus a teacher's utility is increasing in her job performance, $\frac{\partial U}{\partial Q} > 0$ (Dixit 2002). For motivated agents, intrinsic rewards provide a return on investment in skills. Even if the potential returns—intrinsic or extrinsic—remain constant, a reduction in costs should generate new investments in skills.

2. Setting and Data

To test these predictions, 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.[15]

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, and that feature is important for identifying a teacher's contribution to student achievement. Second, teachers who are in the early years of their career,

[15] Additional details on the setting and data are provided in appendix B.

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 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 Evaluation 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 Rate 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.[16]

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."

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

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

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. Over the school year, the teacher is scored 1-3 times on each task, 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.[17]

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; after probation the state only required evaluation every five years. The pre-2012 process also used a rubric which covered several different items (or teaching tasks), and for each item described three levels of performance. While these basic features of the rubric were similar to the new rubric, the types and specificity of tasks covered were different. For example, figure 1

[17] 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 B.

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 other tasks on lesson pacing, communication, etc. into one single scored item for "Teaching Strategies."[18]

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.[19] 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 Koedel, Mihaly, and Rockoff 2015). When describing TVAAS to teachers, Tennessee emphasizes the growth characteristic and that students are compared to peers who scored similarly in prior years.[20] 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.

Achievement Score—The third performance measure is known as the "student achievement score." This measure is also based on student test scores, but often focuses on the level of student achievement as opposed to growth. Each

[18] Appendix C includes both rubrics, as well as a side-by-side comparison of the two. Given the length of appendix C it is available online at https://scholar.harvard.edu/erictaylor.
[19] What is now SAS EVAAS began with William Sanders and colleagues' work in Tennessee in the 1990s (Sanders and Horn 1998).
[20] See for example https://team-tn.org/tvaas/ and https://www.tn.gov/education/data/tvaas.html. Copies of these websites as of August 17, 2022 are available from the author.

teacher defines this measure 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, 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).

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.[21] 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.

[21] 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. The LOE score intervals are [1,2), [2,2.75), [3,3.5), [3.5,4.25), and [4.25,5]. Additional details of LOE scoring are provided in appendix B.

2.2 Evaluation Performance Incentives

Along with the new performance measures in 2012, Tennessee also adopted new rules linking teacher tenure to those measures. But the new rules did not apply to teachers who had earned tenure before July 2011. Tenure incentives are the primary incentives of the new evaluation program. Additionally, for a small number of teachers—15 percent of my sample—compensation was linked to evaluation scores.

New Tenure Rules—Since 2012, annual LOE scores determine who earns tenure. 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.[22]

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.

[22] Tennessee Code § 49-5-504.

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.

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 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.[23]

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 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.

3. Identification Strategy

The paper's main estimates are difference-in-differences style estimates. Each estimate compares teachers who are at the same point in their career—their first year teaching, or second, or third, etc.—but who experience different

[23] These paragraphs describe the period 2012-2015. Additional details are provided in appendix B, including a discussion of 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.

evaluation "treatments." Treatments differ between cohorts, because they began their careers at different times. In each case, my objective is to estimate the causal effect of some evaluation treatment bundle—measures or performance incentives or a combination—on teacher job performance. My outcome measure is the teacher's contribution to student achievement test scores, often called "teacher value added" in the literature.

The basic diff-in-diff features are as follows: Let μ_{je} be teacher j's contribution to (equivalently, casual effect on) student achievement during her e-th year working as a teacher. The first difference is the change in teacher job performance between year $(e-1)$ and e, $E\left[\mu_{j,e} - \mu_{j,e-1}\right]$. The second difference is across groups of teachers. The comparison group is always teachers who had no evaluation treatment in either $(e-1)$ or e, because their e-th year was before the 2012 school year. The treated group experienced a sharp change in some evaluation treatment between $(e-1)$ and e, though the nature of that change differs from estimate to estimate. The key treatment changes are: (i) no evaluation in $(e-1)$ but new evaluation measures in e, because e was in 2012; or (ii) evaluation measures in both $(e-1)$ and e, but a change between $(e-1)$ and e in the performance incentives attached to those measures.[24]

[24] I set $e = 1$ for j's first year working as a teacher in Tennessee, and then 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 several more years which allows me to identify a teacher's first year in Tennessee with confidence.

I apply the diff-in-diff estimator proposed by de Chaisemartin and D'Haultfœuille (2020). If we observed μ_{je} then the estimated effect, $\hat{\delta}$, of a given treatment would be:

$$\hat{\delta} = \sum_e \frac{N_e}{N} \hat{\delta}_e$$

$$\hat{\delta}_e = \left[\frac{1}{N_e} \sum_{\substack{j: D_{j,e-1}=0, \\ D_{j,e}=1}} \left(\mu_{j,e} - \mu_{j,e-1} \right) \right] - \left[\frac{1}{M_e} \sum_{\substack{j: D_{j,e-1}=0, \\ D_{j,e}=0}} \left(\mu_{j,e} - \mu_{j,e-1} \right) \right] \tag{6}$$

where D_{je} is an indicator for treatment status, N_e is the number of treated group teachers, and M_e comparison teachers.

However, because μ_{je} is not directly observable, I use student test score data to fit a regression-based version of (6). The basic specification is

$$A_{ist} = \delta D_{je} + \alpha_j + \gamma_e + f\left(A_{is,t-1}\right) + X_{it}\beta + \varepsilon_{ist}. \tag{7}$$

where A_{ijst} is the test score for student i taught by teacher j in subject s and year t. I fit (7) repeatedly, once for each $\hat{\delta}_e$. In each case the estimation sample is limited to only teachers who are in year e or $(e-1)$ of their teaching career, and for whom either $\{D_{j,e-1}=0, D_{j,e}=1\}$ or $\{D_{j,e-1}=0, D_{j,e}=0\}$. These sample constraints reproduce the key features of (6). 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 $=1$ for year e. For $\hat{\delta}$, just as above, I take the weighted average $\hat{\delta} = \sum_e \frac{N_e}{N} \hat{\delta}_e$, though in some cases an individual $\hat{\delta}_e$ will be of interest. Finally, 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.[25]

[25] 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.

Other features of (7) are more typical of the literature. Student test scores, A_{ijst}, 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, $f(A_{is,t-1})$, where the parameters are allowed to vary by grade and subject, and several other student and peer characteristics in the vector X_{it}.[26] This "lagged test 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 $E[\varepsilon_{ist}] = E[\varepsilon_{ist}|j]$.[27]

The first example of $\hat{\delta}$ is 0.032σ (σ = student standard deviations, standard error 0.005) in table 2 column 1 row 1. For this estimate $D_{je} = 1$ if the teacher's e-th year is 2012. Thus, $\hat{\delta}$ is the average effect of the new evaluation program during its first year 2012, among teachers early in their careers. The 0.032σ estimate is a weighted average of the individual $\hat{\delta}_e$ estimates for $e \in \{2,3,\ldots,7\}$, which are shown in figure 4 by the six unfilled markers.

An alternative to this student-level regression approach would be to, first, estimate $\hat{\mu}_{je}$ itself using methods common in the literature (see Jackson, Rockoff, and Staiger 2014 for a review). Then, second, apply (6) to those estimates. The results are quite similar to those shown in the paper.

In Appendix B I discuss the comparison between this estimator and a two-way fixed effects estimator.

[26] 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 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.

[27] 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).

A causal interpretation of the 0.032σ estimate, and others presented in the next section, requires a parallel trends style assumption: Absent the new evaluation program, teacher performance would have improved with experience (from $(e-1)$ to e) at the same rate observed in cohorts prior to the new program. Econometrically this assumption is clear in (6). 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). Thus, the importance of a counterfactual estimate which includes the typical 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 makes the returns to experience steeper or shallower as each cohort begins their career.[28]

Figure 5 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. There is little evidence that Tennessee's new teacher cohorts are systematically improving or declining over this period. The series is noisy, but we cannot reject a flat trend line. The bottom panel summarizes the returns to experience over time. The estimation procedure follows (6) and (7) 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, averaged over $e \in \{2,3,\dots,7\}$. There is a clear trend break in 2012

[28] 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.

when the new evaluation program begins. Note that the 0.032σ estimate uses only the 2009-2012 years from this graph.[29]

4. Results

Tennessee's new evaluation program improved teacher performance, and the pattern of effects is consistent with teachers making skill investments in response to the program. Performance improves broadly when the new evaluation measures are introduced, but the gains are larger among teachers who have incentives to score well in the future. Performance remains higher after the incentives end.

4.1 The Beginning of Evaluation and Anticipation Effects

In its first year, the new evaluation program improved performance by 0.032σ, on average, among the early-career teachers who comprise the analysis sample. The counterfactual here is not zero improvement, but instead the improvement expected for a teacher moving from $e-1$ to e years of experience absent the new evaluation. Teacher performance grew 0.032σ faster, year over year, than it had grown before the program. In other words, in the classes of early-career teachers, student math and ELA achievement was 0.032σ higher than it would have been without the new evaluation program. This 0.032σ estimate is shown in table 2 column 1 row 1, alongside other estimates discussed in this section.

What role did or could incentives play in generating the 0.032σ improvement? In that first program year there were no explicit incentives linked to

[29] Regarding attrition or composition threats, it is possible that some teachers left teaching as a result of the new evaluation program, e.g., after failing to meet the new tenure cutoff. However, there is little empirical evidence of treatment effects on turnover, as shown in Appendix Figure A3. Recall from section 2 that if a teacher missed the tenure requirement at the first opportunity in year 5, then he could continue working on a probationary contract until he met the requirement. Moreover, this paper's identification strategy avoids composition effects directly by limiting the estimation sample to teachers observed in both $(e-1)$ and e.

evaluation scores. No teacher was being evaluated for tenure in 2012.[30] Still, some teachers were anticipating future incentives. In 2012, those teachers already knew that earning tenure would depend directly on their evaluation scores in future school years. The 0.032σ estimate is the average effect, averaging across these future-incentive teachers and the never-incentive teachers. The latter group already had tenure by 2012 and thus would never have incentives or consequences linked to their evaluation scores.

Comparing effects between future-incentive and never-incentive teachers provides a test of anticipation effects. Both groups improved under the new evaluation program. The performance of never-incentive teachers grew 0.024σ faster than it would have without the new evaluation program. But performance growth was twice as large, 0.047σ, among teachers who were anticipating future incentives. The difference is statistically significant, $p < 0.01$.

This pattern of results is consistent with teachers investing effort to improve their skills, because they anticipate returns in future evaluation rewards. But the results do not rule out a role for teachers' intrinsic motivation to improve. After all, the new evaluation improved the performance of never-incentive teachers. Consider potential mechanisms for the 0.024σ effect among never-incentive teachers. It is plausible, for example, that the key mechanism for that 0.024σ is the personalized feedback generated by the evaluation rubric, feedback which reduced the effort costs of skill investments. This same "motivated agents plus reduced costs" mechanism may also partly explain the 0.047σ effect among future-incentive teachers. Still, the future-incentive effect is twice as large as the never-incentive,

[30] Section 2.2 describes the program incentives in detail. For a small subsample of teachers, at most 15 percent, there were new pay-for-performance incentives linked to evaluation scores in 2012. As shown in table 2 column 2 the estimates are quite robust to excluding these school districts entirely.

strongly suggesting the role of some other mechanism like the anticipation effects prediction.[31]

The general pattern holds for both math and English language arts (columns 3 and 4 of table 2). In both cases teachers with future incentives improve as a result of the new evaluation. In both cases effects for future-incentive teachers are larger than for their never-incentive colleagues. However, the estimates for math are larger than for ELA.

These effects come in the first year of the new program, which may seem too early for a skill investment interpretation. But the 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). And related (quasi-)experiments have also shown teacher performance improvements in the first year of a new program (Taylor and Tyler 2012, Jackson and Makarin 2018, Papay et al. 2020, Burgess, Rawal, and Taylor 2021, Briole and Maurin in-press, Hanno 2022).

A career concerns explanation is one potential alternative to the skill investment explanation. Career concerns can motivate greater current effort even without explicit current incentives (Fama 1980, Holmstrom 1999, Lazear and Oyer 2013). Many career concerns considerations would be part of the counterfactual and differenced out. To explain the effect estimates discussed above, the new evaluation program would need to create new career concerns channels. One

[31] One alternative hypothesis for the difference is the following: The two teacher types are colinear with years of experience, and perhaps treatment effects are a decreasing function of experience. The data are not consistent with this hypothesis. Figure 4 shows effect estimates by years of experience (equivalently, by cohort). The relationship is not monotonic. Additionally, effects are larger for second year teachers than for third, but among the never-incentive group we cannot reject homogeneity. The alternative hypothesis is also not consistent with prior (quasi-)experimental evidence (Taylor and Tyler 2012, Papay et al. 2020, Burgess, Rawal, and Taylor 2021, Briole and Maurin in-press).

example 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 count for earning tenure. Empirical evidence partly supports and partly contradicts that second belief (Ho and Kane 2013). Additionally, the example applies to classroom observation ratings, which are infrequent, and the effect estimates are measured in teachers' contributions to student test scores.

4.2 The Beginning of Incentives and Conventional Effects

Teacher performance improved further when the new program's incentives began in a teacher's fourth year of employment. That improvement is consistent with the conventional prediction that an employee's current effort responds to current incentives.

To test that conventional prediction, focus on the change in performance between year 3 and year 4 of employment. That is, apply the difference-in-differences estimator in (6) to the case $e - 1 = 3$ and $e = 4$. Recall from section 2 that, under the new evaluation program, to earn tenure a teacher had to score LOE ≥ 4 in both year 4 and year 5. Scores from year 3 did not count for tenure or any other reward.

Between year 3 and 4 average teacher performance improved a further 0.013σ under the new program. That estimate is not statistically significantly different from zero (its standard error is 0.009). However, there are two important considerations when interpreting that estimate.

First, the 0.013σ excludes any skill improvements made before year 4. The 0.013σ gain is the effect of the onset of tenure incentives, plus any further skill investments in year 4. In the difference-in-differences mechanics for this estimate year 3 is the "untreated" period, subtracted off of year 4; but, of course, year 3 is treated by any anticipation mechanisms.

Second, the 0.013σ estimate likely masks heterogeneity across teachers. At the start of year 4, each teacher had a relevant signal from their evaluation score in year 3 and earlier. The turning on of tenure consequences in year 4 was thus likely more salient for teachers who had previously scored below the tenure cutoff. And indeed, the effects are larger for lower-performing teachers. Among teachers who scored below the tenure cutoff in year 3, performance improved by an extra 0.077σ between year 3 and 4. Among those who had met the cutoff in year 3, the effect estimate is -0.018σ. That difference—0.077σ versus -0.018σ—is stark but it could simply reflect persistent differences in performance levels between the two groups. Both point estimates are relative to the same comparison group of teachers; I do not observe LOE evaluation scores for the comparison group because their early-career years occurred before the new evaluation program began in 2012.[32]

To sharpen the comparison of these two groups—teachers who scored above the tenure requirement in year 3 versus those who did not—figure 6 plots the difference in their value-added performance from year to year. Figure 6 is an event study style plot, where the difference between the two groups is set to zero in $e = 3$. The difference in performance is stable for two years leading up to the beginning 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 $e = 3$. That change in effort would be consistent with the conventional prediction.[33]

[32] Additionally, in contrast to table 4, for table 3 I do not observe LOE scores for the never-incentive teachers, because they were not scored in year 3. Recall this group had earned tenure under the old rules, and thus by definition had completed year 3 before the new evaluation program and scoring began.

[33] The change in figure 6 is a change in the difference between the two groups, and thus could also partly reflect a decline in effort among the teachers who scored above the tenure cutoff in $e = 3$. In figure 6, the estimate for $e = 1$ is positive, suggesting teachers who missed the cutoff in $e = 3$ may have started their careers slightly higher performing before a decline. Nevertheless, the point estimate for $e = 4$ is significantly different from $e = 1$.

4.3 The End of Incentives and Persistent Effects

For most teachers the evaluation program's incentives ended after year 5, when they earned tenure, and yet their performance remained higher. To test the persistent effects prediction in table 4, I compare performance in year 5 and year 6, applying the estimator in (6) to $e - 1 = 5$ and $e = 6$. Recall that, if a teacher scored above the cutoff in years 4 and 5, she earned tenure and her year 6 and future scores no longer had any incentive attached. In the analysis sample nearly two-thirds of teachers earned tenure after year 5 (figure 2).

Among all teachers subject to the new tenure requirements, performance in year 6 was 0.037σ higher as a result of the evaluation program. Again, as with the other estimates, the 0.037σ gain is in addition to the typical (counterfactual) improvement in teacher performance between year 5 and 6.

That 0.037σ estimate is evidence of persistence after incentives end. Perhaps stronger evidence than it first seems. For a moment, assume—contrary to the skill investment predictions—that any boost in performance in year 5 (or earlier) was only the result of the conventional mechanism: higher current effort in response to current incentives. Under that assumption we would predict a *negative* effect estimate in year 6 when incentives end. If year 5 performance was boosted by current incentives, then to return to the counterfactual level of performance in year 6 would require a decline between 5 and 6.[34]

As before, I can test for heterogeneity of effects by prior evaluation score. That analysis is sharper for year 6 effects because I observe year 5 scores for both the teachers subject to the new tenure requirements, and for the already-tenured never-incentive teachers. Moreover, this heterogeneity is first order for testing

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

persistence predictions, because the incentives only end if the teacher earns tenure in year 5. Teachers who miss the tenure cutoff are still (presumably) trying to earn tenure in year 6. The results are shown in table 4 and summarized in figure 7.

Teacher performance grows 0.024σ (standard error 0.013) faster between year 5 and 6 among teachers who scored above the cutoff in year 5, earned tenure, and no longer had any incentive from the evaluation program in year 6. Under the conventional prediction, these newly-tenured teachers' performance should fall in year 6, but the 95 percent confidence interval excludes declines larger than -0.001σ. Moreover, contrast that 0.024σ gain with the analogous estimate of -0.000σ for teachers who were already tenured, had no incentive in year 5, but nevertheless scored above the cutoff in year 5. All together, these results are consistent with the persistence of performance gains after the evaluation incentives end.

5. Conclusion

This paper documents teachers' (employees') responses to performance measures and performance incentives in a new evaluation program. Consistent with conventional predictions, teacher performance improved when the program provided an explicit incentive to score high. To earn tenure teachers had to meet a score cutoff in year 4 and 5 of their employment; between year 3 and 4 average performance improved faster than expected based on prior cohorts of teachers.

But there is also evidence of anticipation effects before performance incentives begin, and persistent effects after the incentives end. In the new evaluation program's first year, performance improved 0.047σ faster among year 2 and 3 teachers who knew they would have future tenure incentives attached to future evaluation scores. Most teachers subsequently earned tenure at the end of year 5, and thus the performance incentives ended, but those teachers' performance was 0.024σ higher than expected in year 6.

This pattern of effects—especially the anticipation and persistence effects—is consistent with teachers investing in human capital (equivalently, improving their skills) as a response to the evaluation program's performance incentives. Indeed, combining familiar features of agency theory and human capital investment models yields predictions of 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. The evidence presented here does not necessarily contradict that hypothesis. In the new evaluation program's first year performance also improved by 0.024σ for teachers who already had tenure and did not have any future incentives. Still, the gains were twice as large, 0.047σ, for teachers who were anticipating future incentives. Additionally, among teachers whose scores met the new tenure requirement in year 5, performance in year 6 remained higher for teachers who had just earned tenure; while among the already-tenured teachers, for whom meeting the requirement was meaningless, performance was similar to pre-program comparison teachers. If "motivated agents plus feedback" alone were sufficient for evaluation-induced skill improvements, then both groups would have similar performance trajectories, and similarly outperformed the comparison group.

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 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. Additionally, and partly 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 estimated 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). 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 1-2 weeks of additional class time to the school year (Fitzpatrick, Grissmer, and Hastedt 2011, Aucejo and Romano 2016). 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.

These results have important practical implications for managers and policymakers designing performance measurement and incentive programs. First, the intended benefits of such programs—improved teacher (employee) performance—can occur before or after the program is actively linking performance and rewards. While the costs typically occur during the active period. Thus the traditional focus on benefits only during the active program period will understate the cost-effectiveness of the program. Moreover, if research efforts to estimate the benefits compare the active period to a before or after period, then those estimates may further understate the true benefits. Second, the results also suggest 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 how much of between-employee differences in performance are the result of differences in skills. The intertemporal effects found here for teachers may or may not occur, for example, in the repair technicians case studied in Lazear (2000).

The most direct application of these results is in understanding the effects of Tennessee's program and similar programs in other states. That is, 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 teacher performance and student achievement. But the estimates focused on specific predictions may make the larger story hard to see. Figure 8 shows the returns to experience trajectory for teachers hired in 2009 and 2010, alongside teachers hired before 2009.[35] The 2009-2010 teachers were already working in Tennessee schools when the new evaluation program began in 2012, but were subject to the new tenure requirements. Figure 8 shows the broader story of improvements both before performance incentives begin in year 4 and persisting after incentives end in year 5, though here some of the persistence is teachers who remain subject to the tenure incentives.

[35] The estimates in figure 8 come from a within-teacher approach common in the literature on teachers (Rockoff 2004, Papay and Kraft 2015). In the teacher fixed effects regression, the outcome is student test scores, and controls include prior scores and other controls as in (7). The key right-hand-side variables are indicators for years of employment. For figure 8 I allow the coefficients on experience to differ for the two groups of teachers defined by hire year. Standard errors are clustered at the teacher level.

References

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.

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.

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.*

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.

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. (in-press). "Human capital depreciation and returns to experience." *American Economic Review.*

Dinerstein, M., & Opper, I. (2022). "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., Dupas, P., & Kremer, M. (2011). "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya." *American Economic Review*, *101*(5), 1739-74.

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., & 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.

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

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

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

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

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

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, K., & Makarin, A. (2018). “Can online off-the-shelf lessons improve student outcomes? Evidence from a field experiment.” *American Economic Journal: Economic Policy*, *10*(3), 226-54.

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.

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

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.

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.

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

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

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

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

New York Times. 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.* 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.

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. & Tyler, J. (2012). "The Effect of Evaluation on Teacher Performance." *American Economic Review, 102*(7), 3628-3651.

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.

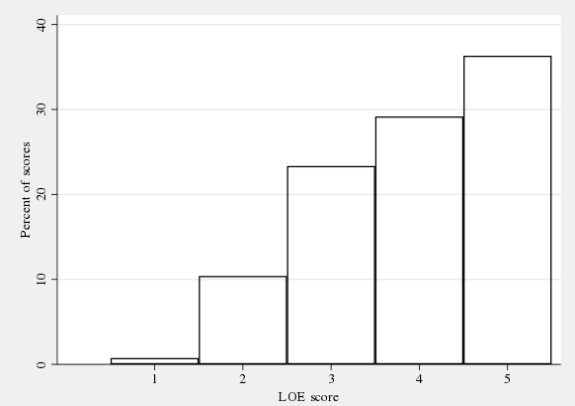
Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. The New Teacher Project.

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

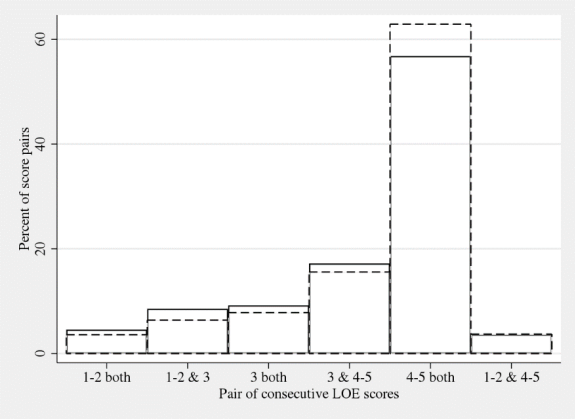


Figure 2—Distribution of final LOE scores

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 in year 4 and 5 specifically.

<table>
<tr><td rowspan="2">Year of employment</td><td colspan="2">Year hired</td></tr>
<tr><td colspan="2">≥ 2010</td><td>< 2010</td></tr>
<tr><td>1-3</td><td colspan="2">no incentives</td><td rowspan="4">no incentives</td></tr>
<tr><td>4-5</td><td colspan="2">must score LOE “4” or “5”
in both years 4 and 5
to receive tenure

cutoff for “4” ≅ 33rd percentile</td></tr>
<tr><td rowspan="2">6+</td><td>tenured</td><td>not tenured</td></tr>
<tr><td>if rated LOE “1” or “2”
two consecutive years
tenure revoked

cutoff for “2”
≅ 10th percentile</td><td>must score
LOE “4” or “5”
two consecutive years
to receive tenure</td></tr>
</table>

Figure 3—Performance incentives

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

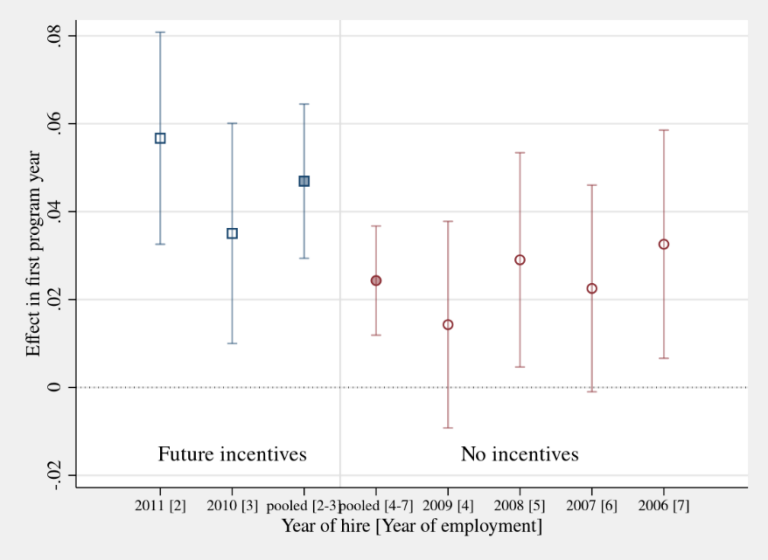
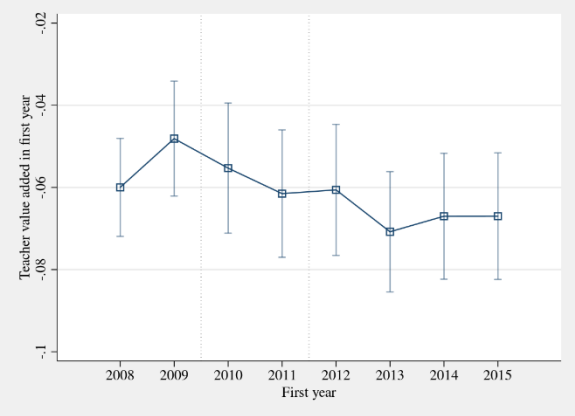


Figure 4—Effects of new performance measures in first program year

Note: This figure plots diff-in-diff point estimates from table 2 column 1 and appendix table A1 column 1. See the table 2 note for details. Vertical lines mark 95 percent cluster-corrected confidence intervals, with teacher clusters.

(a) First year performance over time



(b) Returns to experience over time

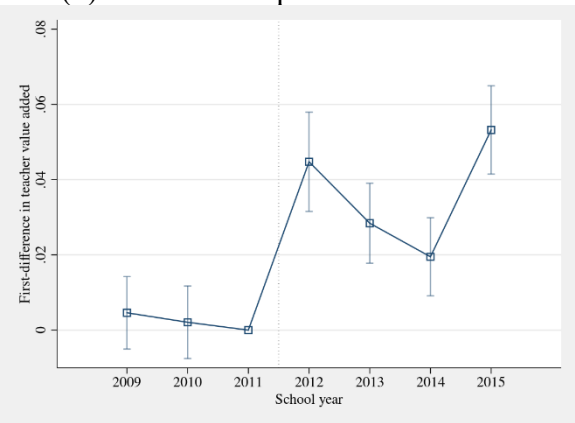


Figure 5—Trends in teacher performance and the returns to experience

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).

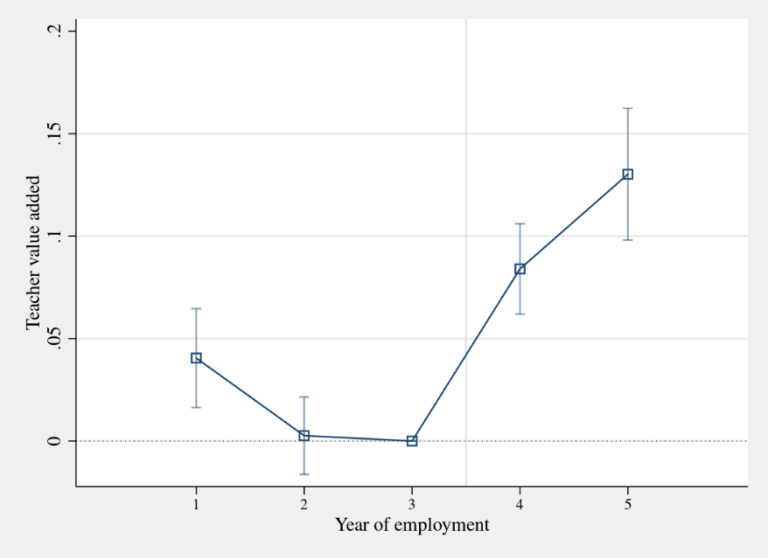


Figure 6—Difference in performance between teachers who scored below the tenure cutoff in year 3 and those who scored above

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.

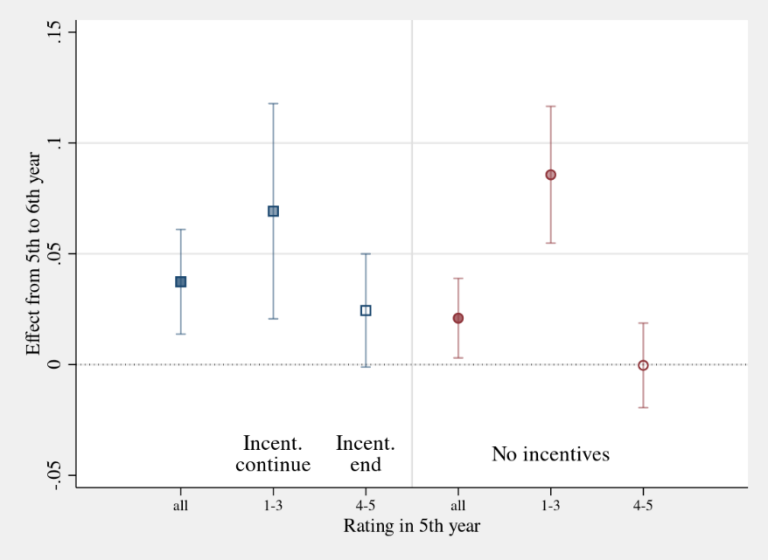


Figure 7—Effects from year 5 to year 6 of employment

Note: This figure plots diff-in-diff point estimates from table 4 column 1. See the table 4 note for details. Vertical lines mark 95 percent cluster-corrected confidence intervals, with teacher clusters.

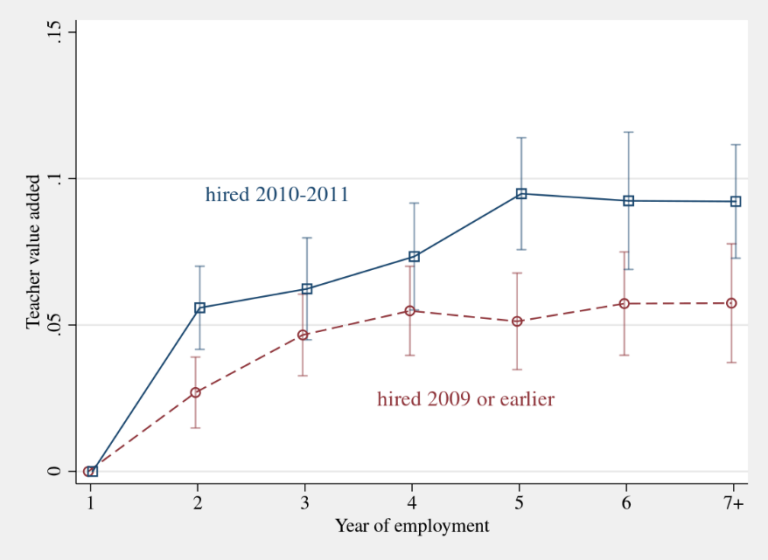


Figure 8—Returns to experience and the new evaluation program

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 a series of indicators for teacher j's year of employment, e, in year t, with $e \geq 7$ a single top-coded indicator. The coefficient on each e indicator is allowed to differ for teachers whose first year, $e = 1$, was in 2010 or 2011 (solid line, boxes) versus $\leq$ 2009 (dashed line, circles).

Table 1—Characteristics of study teachers and their students

	Teaching grades 4-8, math and ELA		All teachers
	Years 1-7	All	
	(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. The positive means in column 3 reflect negative selection of students leaving Tennessee public schools.

Table 2—Performance effects in first program year

	Full sample	Excluding P4P districts	Math	ELA
	(1)	(2)	(3)	(4)
All teachers	0.032	0.033	0.046	0.018
	(0.005)	(0.006)	(0.009)	(0.005)
Performance incentive in the future	0.047	0.046	0.065	0.028
	(0.009)	(0.010)	(0.015)	(0.009)
No incentive, already tenured	0.024	0.027	0.036	0.013
	(0.006)	(0.007)	(0.011)	(0.006)
Teacher observations	6,998	6,016	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 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 $D_{j,e}$, and the estimation sample is further limited to only observations where either $\{D_{j,e-1}=0, D_{j,e}=1\}$ "treated" or $\{D_{j,e-1}=0, D_{j,e}=0\}$ "comparison" teachers. In this table $D_{j,e}=1$ 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 $e \in \{2,3,\ldots 7\}$, 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 $e \in \{2,3\}$ where treated teachers had a future incentive, and the third $e \in \{4,5,6,7\}$ where there was no future incentive. Columns 2-4 report estimates for subsamples described in the header. P4P = pay for performance (see section 2.2).

Table 3—Performance effects when incentives begin
(year 3 to 4 effects)

	Full sample	Excluding P4P districts	Math	ELA
	(1)	(2)	(3)	(4)
Tenure incentives	0.013	0.017	0.019	0.006
	(0.009)	(0.010)	(0.014)	(0.009)
Evaluation score in year 3				
Above tenure cut	-0.018	-0.016	-0.015	-0.022
	(0.009)	(0.010)	(0.014)	(0.010)
Below tenure cut	0.077	0.094	0.109	0.053
	(0.012)	(0.014)	(0.021)	(0.013)
Teacher observations	3,849	3,279	2,304	2,665

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 with the following exceptions. Table 3 reports only estimates for the case $(e-1)=3$ and $e=4$. In the top row, $D_{j,e}=1$ in year $e=4$ if teacher j was 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. Row 2 further limits the treated group to teachers who scored LOE = 4-5 in $e=3$. Similarly, row 3 limits to LOE = 1-3 in $e=3$. Columns 2-4 report estimates for subsamples described in the header. P4P = pay for performance (see section 2.2).

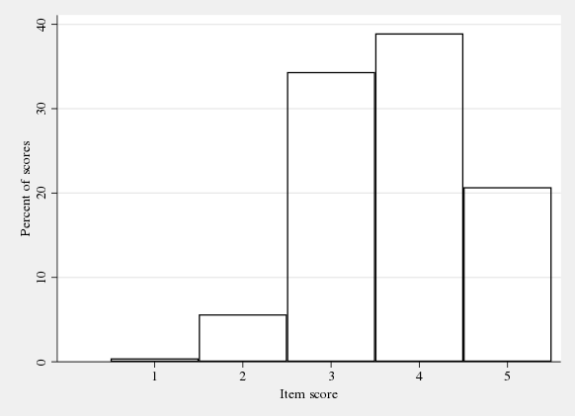
Table 4—Performance effects when incentives end
(year 5 to 6 effects)

	Full sample	Excluding P4P districts	Math	ELA
	(1)	(2)	(3)	(4)
Tenure incentives	0.037	0.044	0.039	0.036
	(0.012)	(0.016)	(0.020)	(0.013)
Evaluation score in year 5				
Above tenure cut, incentives end	0.024	0.025	0.039	0.006
	(0.013)	(0.018)	(0.020)	(0.015)
Below tenure cut, incentives continue	0.069	0.096	0.030	0.092
	(0.025)	(0.034)	(0.056)	(0.022)
No incentive, already tenured	0.021	0.020	0.033	0.009
	(0.009)	(0.010)	(0.015)	(0.010)
Evaluation score in year 5				
Above tenure cut	-0.000	0.000	0.010	-0.013
	(0.010)	(0.011)	(0.015)	(0.011)
Below tenure cut	0.086	0.084	0.134	0.055
	(0.016)	(0.018)	(0.031)	(0.015)
Teacher observations	3,426	2,965	2,011	2,402

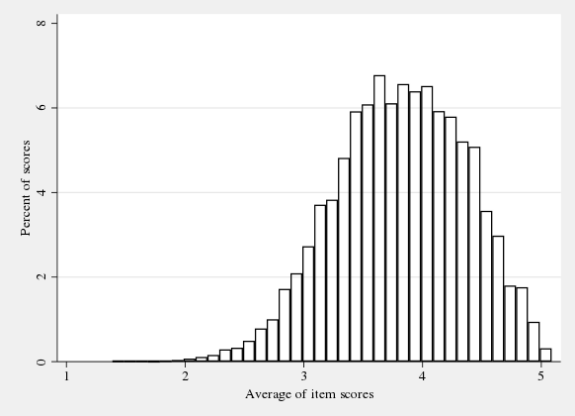
Note: Difference-in-differences estimates from a system of seemingly unrelated regressions, with standard errors in parentheses corrected for clusters (teacher) across equations. The details of estimation are the same as in table 2 with the following exceptions. Table 4 reports only estimates for the case $(e-1)=5$ and $e=6$. In the top row, $D_{j,e}=1$ in year $e=6$ if teacher j was 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 $e=6$ in $t\leq 2011$, before the new program began. Row 2 further limits the treated group to teachers who scored LOE = 4-5 in $e=5$, teachers for who earned tenure and for whom the incentives end. Similarly, row 3 limits to LOE = 1-3 in $e=5$, teachers for whom the incentives continue in $e=6$. Rows 4-6 are constructed just as rows 1-3, except that the "treated" group pf teachers is different. For rows 4-6 the "treated" group is teachers for whom $(e-1)=5$ and $e=6$ both occurred in $t\geq 2012$, during the new evaluation program years; teachers who already had tenure before 2012 and thus had no incentive attached to their LOE or other evaluation scores. Columns 2-4 report estimates for subsamples described in the header. P4P = pay for performance (see section 2.2).

Appendix A: Additional figures and tables

(a) Item level scores

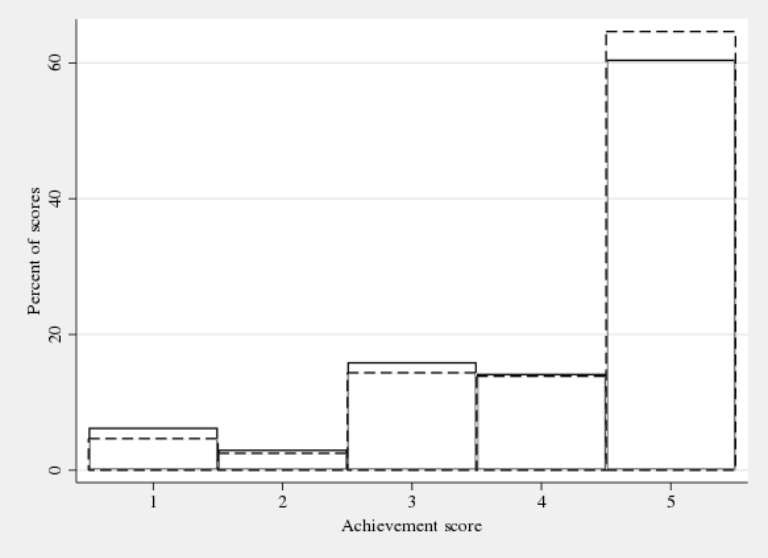


(b) Year average of item scores



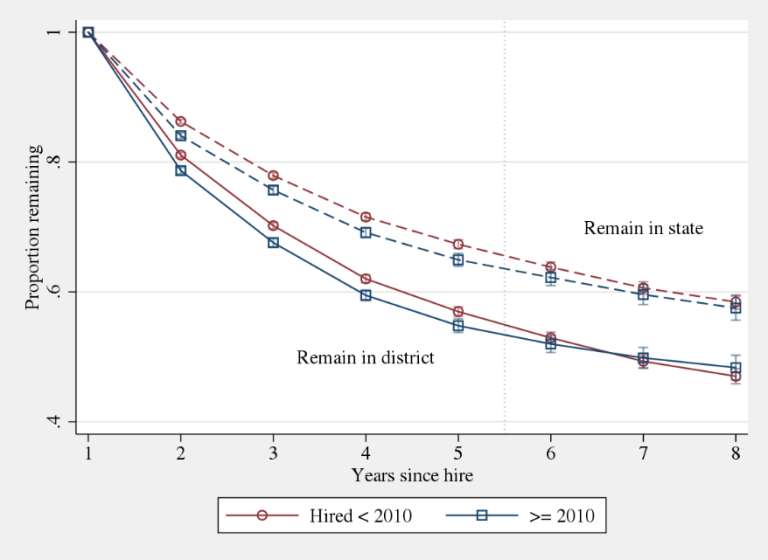
Appendix Figure A1—Distribution of classroom observation scores

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 scores

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—Proportion of teachers still employed by district or state over early career

Note: Solid lines track the proportion of teachers (no regression adjustments) who are still teaching in the same school district where they first worked as a teacher in Tennessee. Dashed lines track the proportion still teaching anywhere in the state's public schools. The series with square markers is teachers who began teaching in 2010 or later, and the circle markers teachers who began before 2010. Vertical lines mark 95 percent cluster-corrected confidence intervals, with teacher clusters.

Appendix Table A1—Performance effects in first program year, additional estimates

	Full sample	Excluding P4P districts	Math	ELA
	(1)	(2)	(3)	(4)
Performance incentive in the future				
Year 2	0.057	0.060	0.083	0.029
	(0.012)	(0.014)	(0.020)	(0.013)
Year 3	0.035	0.029	0.042	0.028
	(0.013)	(0.014)	(0.022)	(0.013)
No incentive, already tenured				
Year 4	0.014	0.019	0.032	-0.003
	(0.012)	(0.013)	(0.021)	(0.012)
Year 5	0.029	0.029	0.044	0.014
	(0.012)	(0.014)	(0.021)	(0.012)
Year 6	0.023	0.025	0.036	0.010
	(0.012)	(0.013)	(0.021)	(0.011)
Year 7	0.033	0.035	0.031	0.035
	(0.013)	(0.015)	(0.023)	(0.013)
Teacher observations	6,998	6,016	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, however, this table reports individual $\hat{\delta}^e$ from the component regressions. P4P = pay for performance (see section 2.2).

Appendix B: Additional details

B.1 Setting and Data

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 $\geq$ 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.

Additional details on pay-for-performance programs—Section 2 "pay for performance" describes the period 2012-2015 drawing 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.

B.2 Identification Strategy

Comparison to two-way FE estimator—The alternative estimator (6), "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 (7) 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

[1] Even if μ_{je} were observed, notice that (6) could itself be carried out using a system of least squares regressions, with appropriately defined sample constraints and weights.

(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 $\{D_{j,e-1}=0, D_{j,e}=0\}$. 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 (7) for each $\hat{\delta}_e$ separately my approach avoids this problem.

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.

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.

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.