CESifo
AREA CONFERENCES 2021

Economics of Education

Munich, 3–4 September 2021

Measuring returns to experience using observations of teaching

Courtney Bell, Jessalynn James, Eric S. Taylor, and James Wyckoff

CESifo

Measuring returns to experience using observations of teaching†

Courtney Bell, University of Wisconsin
Jessalynn James, TNTP
Eric S. Taylor, Harvard University
James Wyckoff, University of Virginia

July 2021

We study the returns to experience in teaching, estimated using evaluation ratings from classroom observations. We describe the assumptions required to interpret changes in observation scores over time as the causal effect of experience on performance. We compare two difference-in-differences style estimation strategies: the within-teacher estimator common in the literature, and an alternative which avoids potential biases in the common approach. Using data from Tennessee and Washington, DC, we show empirical tests relevant to assessing the identifying assumptions and substantive threats—e.g., leniency bias, manipulation, changes in incentives or job assignments—and find our estimates are robust to several threats.
JEL No. I2, J24, M5

† Bell: courtney.bell@wisc.edu. James: jessalynn.james@tntp.org. Taylor (corresponding author): eric_taylor@gse.harvard.edu. Wyckoff: wyckoff@virginia.edu. We first thank the District of Columbia Public Schools, Tennessee Department of Education, and Tennessee Education Research Alliance. Generous financial support was provided by the Spencer Foundation. We also thank Brendan Bartanen, Katherine Castellano, Heather Hill, and Ben Ost for helpful comments on earlier drafts.

Monitoring employee job performance is a fundamental task in personnel management. In particular, understanding how performance improves with experience—the "returns to experience"—is critical to decisions about hiring and turnover, investments in employee training, and others. Consider the choice between retaining a current employee or replacing that employee with a novice new hire; the optimal choice depends not simply on the current performance of the two individuals, but rather on each person's expected future performance over time. However, isolating the causal effects of experience is complicated by imperfect and incomplete performance measures, and selection on performance through hiring and turnover decisions. We examine the case of classroom teachers, and the most common performance measure for public-school teachers: rubric-scored classroom observations.

Our focus is estimating the returns to experience in teaching using data from classroom observations. To be precise, we define "returns to experience" as the causal effect of one additional year of experience on teacher performance. We estimate returns separately for the first year of experience, second year, third year, etc. Our primary objective is evaluating claims about returns to experience for (a) performance of the teaching practice inputs which the observation rubrics are designed to measure. But we also consider inferences about returns to experience on (b) broader output-based measures of teacher performance, like teacher contributions to student outcomes. The extent to which experience affects (a) and (b) differently partly motivates our work, because input-based measures are much more common in schools than output-based measures.

Given the causal inference goal, we make explicit the causal inference features of the problem, including identifying assumptions, threats to those assumptions, and an empirical exploration of those threats. To begin we show returns-to-experience estimates using observation scores and the standard methods for estimating returns to experience in teaching. These estimates are the solid lines in Figure 1 using data from Tennessee and Washington, DC. We then describe how those estimates can be thought of as difference-in-differences style estimates. The identifying assumptions are clarified by combining the diff-in-diff framework with a conceptual framework that relates observation scores to actual teacher performance on different dimensions of performance. Those assumptions require, first, that veteran (comparison) teachers no longer experience returns to an additional year of experience. Second, that the process, explicit or implicit, that maps true performance to scores does not depend on a teacher's years of experience.

We evaluate a number of threats to these identifying assumptions. Most threats are reasons why observation scores might rise (fall) over time even if a teacher's true performance is unchanged. One simple example is when changes are made to the scoring rubric, as happened in Washington, DC Public Schools (DCPS) in 2017. As we discuss in detail, changes to the rubric (or to rater training, or to rater-teacher matching rules) do not necessarily threaten inferences about the causal returns to experience. Veteran teachers—the diff-in-diff comparison group—provide an estimate of the effect of such changes under the first assumption above, and that estimate is a reasonable counterfactual for early-career teachers under the second assumption above. We use similar reasoning, combined with empirical evidence where available, to address other threats: rater leniency bias, raters using information from outside the observation, changes in incentives that distort teacher effort, manipulation behaviors by teachers which raise scores but not performance, the effect of job changes, and others.

We find little evidence that these potential threats compromise a causal interpretation of the typical returns-to-experience estimates, applied to observation scores. Veterans provide a plausible diff-in-diff style counterfactual estimate for several often-stated threats: leniency bias from raters, manipulation by teachers, changes in the evaluation system, changes in teachers' job assignments, and others. Our estimates are robust to changes in the rubric, different rater types, and controlling for student baseline achievement, among other things. Still, there are reasons to remain cautious about a causal interpretation. We find, in one setting, a weakening correlation between teacher observation scores and student test scores as teacher experience grows.

The standard returns-to-experience estimates also carry potential bias from the use of two-way fixed effects methods. We discuss the standard estimates in light of recent insights from de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (in-press), and others. One distinctive feature of the returns-to-experience case is multiple treatments: the first year of experience, second year, third year, etc. Thus, for example, to correctly estimate the first-year effect we must account for any second-year effect (or third-year effect, etc.) occurring in the comparison group. When there are multiple treatments, the two-way fixed effects estimates will be biased if the correlation of treatments is changing over time. We show this potential bias formally by building on the Goodman-Bacon (in-press) framework. In the current case, this bias will occur when the distribution of teacher experience is changing over time.

Empirically, however, these potential two-way fixed effects biases do not substantially alter our conclusions about the returns to experience captured by classroom observations. The dashed line in Figure 1 shows estimates using the alternative diff-in-diff estimator proposed in de Chaisemartin and D'Haultfœuille (2020). As we discuss in the paper, this alternative mechanically avoids the potential bias of the two-way fixed effects method, but at the cost of precision. The differences in in DCPS estimates are largely explained by changes over time in the DCPS distribution of experience, while Tennessee's distribution remains relatively constant.

Understanding observation-based measures is especially salient in the education sector. As differences in teacher effectiveness have become more formally recognized in practice and research, substantial attention has focused on the development, understanding, and application of measures of teacher performance (Goe, Bell, and Little 2008, Kane, Kerr, and Pianta 2014, Rowan and Raudenbush 2016). Despite wide recognition of the importance of effective teaching, there is comparatively little evidence on whether or how teaching improves (Jackson, Rockoff, and Staiger 2014). One exception is that, on average, teaching performance improves over the first few years of a teacher's career. This returns-to-experience finding has been widely replicated, but nearly all existing estimates measure performance with teacher "value added" to student achievement test scores (see, for example, Rockoff 2004, Papay and Kraft 2015). The gains to student achievement shown by returns to early-career experience hint at opportunities for successful teacher training that has long vexed researchers. However, test-score value-added measures are outcomes, and offer little insight on the teaching tasks by which teachers could improve.

Classroom observations offer another measure of teaching performance that may provide insight on the specific skills teachers develop early in their careers. Standardized, rubric-scored classroom observations are now widely used, and most teachers receive at least one observation per year (Cohen and Goldhaber 2016, Steinberg and Kraft 2017). States and school districts can use observations for a variety of purposes, including understanding changes in teaching performance over time. For example, Figure 1 shows the within-teacher average improvement over the first ten years of teaching for several cohorts of Tennessee and DCPS teachers as measured by their performance on teacher observations. Figure 2 shows the same estimates but for value added to student test scores, and Figure 3 for teachers' improvement on student surveys about teachers' practice in DCPS. The pattern of change over time is similar in these three graphs, which raises a

host of questions about whether each reflects true improvement and, if so, the relationship between skill development and teachers' ability to improve student achievement.

This paper makes three contributions to the literature on teacher job performance. First, we report returns-to-experience estimates using classroom observation scores from two evaluation systems—Tennessee and Washington, DC. There are many returns-to-experience estimates in the teacher performance literature, but nearly all existing estimates use student test-score measures of teacher performance (for a review, see Jackson, Rockoff, and Staiger 2014). Our observation score estimates of inputs complement test score estimates of outputs. In that sense, our estimates add to existing, but scarce, efforts to understand the mechanisms behind test-score returns to experience (Kraft and Papay 2014, Ost 2014, Atteberry, Loeb, and Wyckoff 2015). We are aware of only two other papers that use observation scores to measure returns to experience: Kraft, Papay, and Chi (2020) using data from Charlotte, North Carolina, and contemporaneous work by Laski and Papay (2020) also using Tennessee's data.

Second, as described already, we make explicit the causal inference features of our returns-to-experience estimates. Employing observation-based estimates to better understand the potential effect of individual and school factors to teacher development assumes a causal relationship. The identifying assumptions and related threats have not previously been addressed explicitly, at least in the study of observation scores. Moreover, the diff-in-diff features and assumptions we describe have direct parallels from estimates using student test scores. Third, and related, our analysis of threats incorporates concerns about observation scores raised in prior papers, and in many cases provides new empirical evidence. These known concerns include rater leniency bias (Weisberg, Sexton, Mulhern, Keeling et al. 2009; Steinberg and Kraft 2017), the influence of the students in the classroom (Campbell and Ronfeldt 2018), unintended effects of teacher-rater pairings (Chi 2020), among other concerns (Cohen and Goldhaber 2016, Grissom and Bartanen 2019).

In Section 1 we replicate the standard returns to experience estimates in the DCPS and Tennessee settings, including describing the data for the paper. In Section 2 we describe a framework, both conceptual and econometric, to aid in evaluating claims about (inferences from) the standard returns-to-experience estimates. Section 3 discusses threats to a causal interpretation of returns-to-experience estimates, providing empirical tests where we can. Section 4 concludes.

1. Data and setting

The data from DCPS and Tennessee are well-suited to examining the early-career development of teaching skills and are distinctive in several respects. Most importantly, we observe early-career changes in individual educators' teaching practices, using data from rubric-based classroom observation programs that have been used in each setting over many years. In DCPS, the panel begins with the start of its current evaluation system, IMPACT, in 2009-10, and spans through 2018-19. Tennessee's current evaluation system began in 2011-12, and our data run from that start through 2018-19. Both datasets include several measures of teacher performance, including item-level sores for each specific task-based item on the observation rubric, as well composite scores which average across items. Teachers in tested grades and subjects can be linked to their students, and corresponding achievement scores and demographic information. Characteristics of the teachers and their students in out data are summarized in Table 1.

Common features. The DCPS and Tennessee settings share many first-order features. In both locations, all teachers—regardless of experience—are evaluated every year by trained observers. Likewise, for both Tennessee and DCPS, observation scores are a substantial component of a larger set of evaluation measures, including "value added" scores which measure teacher contributions to student achievement.[2] Those larger evaluation systems are used to identify exemplary teachers, those in need of additional support or training, or individuals who will be dismissed. During most of the period we study, teachers in DCPS were observed five times per year. After a change in the rubric in 2017, teachers were observed up to three times per year depending on experience and performance. In Tennessee, the number of evaluations per year varies according to teachers' prior performance (and licensure status), but teachers are typically evaluated multiple times per year. The median novice teacher in Tennessee receives 2.5 formal observations and the median novice teacher in DCPS receives five formal observations.

While the two systems use different rubrics, they both use standards-based observation protocols to assess teaching practice, and both systems assess similar tasks and teaching practices. Tennessee uses the TEAM (Tennessee Educator Acceleration Model) evaluation rubric. Rubric items are divided into three categories of skills: instruction, planning, and environment. Each

[2] In DCPS classroom observations account for 75 percent of overall IMPACT scores for the more than 80 percent of teachers without a value-added score. For teacher with value added as part of their evaluation, observations account for between 30 and 40 percent depending on the year. In Tennessee, classroom observations are 50 and 85 percent of the overall TEAM score for teachers with and without value-added scores, respectively.

category is comprised of multiple items for teaching tasks. Ratings for each item range from 1-5 (5 = significantly above expectations, 1 = significantly below expectations). During most of the period of our analysis, DCPS used an observation rubric called the Teaching and Learning Framework (TLF), a modified version of Danielson's Framework for Teaching (1997). The DCPS rubric has a 1-4 rating scale (4 = highly effective, 1 = ineffective) for items measuring nine teaching tasks.[3] In 2017, DCPS transitioned to the Essential Practices (EP) observation rubric, which covers similar skills to the TLF, but with more concise definitions for each related task and explicit alignment to the Common Core State Standards.

One frequent, but misleading, criticism of such classroom observation systems is that the scores produced have little variation, with most teachers scoring in one or two top categories (Kraft and Gilmour 2017, Weisberg, Sexton, Mulhern, and Keeling 2009). This lack of variation arises in part because final scores are rounded off to integer values. In this paper we use observation scores that average across many item scores (several items and several observations of a given item), and those scores vary meaningfully, with a relatively Gaussian density (as shown in Appendix Figure A1).

Differences between the two evaluation systems. While both evaluation systems share many features, there are a number of useful differences. First, while both places use trained school-based personnel to conduct evaluations (e.g., principals and assistant principals, or other instructional leaders), through 2016 DCPS additionally employed "master educators"—observers external to the school with subject- and grade-specific expertise. Two of every teacher's annual observations were conducted by a master educator.

Second, the two systems have different incentives and consequences associated with teachers' performance scores. While both DCPS and Tennessee might be considered high-stakes evaluation systems, DCPS's has notably higher stakes. In DCPS, teachers with particularly low performance (those whose composite scores earn them an overall rating of ineffective) and teachers with relatively low performance (those rated lower than effective) who fail to improve are subject to involuntary dismissal—a policy that influences teachers' improvement and their retention decisions, at least at the margins (Dee and Wyckoff 2015, Dee, James, and Wyckoff 2021). There are also stakes in DCPS associated with high performance. Teachers who

[3] The first seven tasks align generally with the domain of instruction, while the final two align with the domains of classroom management and environment.

demonstrate exceptional performance (those rated highly effective) are eligible for substantial bonuses and, if they continue to perform well, large base pay increases. In Tennessee, to earn tenure a teacher must receive a final composite score of "above expectations" or higher (roughly the top two-thirds of teachers) for two consecutive years, after working at least five years total. Tenure can be revoked based on evaluation scores but that is rare: a teacher must score "below expectations" or lower (roughly the bottom 5 percent of teachers) for two consecutive years, and this rule does not apply to teachers who were tenured before 2011-12.

Another difference between the two systems is DCPS's recent addition of another task-based evaluation measure that assesses instructional quality (with data available from 2016-17 through 2018-19)—student surveys of practice. This measure is adapted from the Tripod survey (see Ferguson and Danielson 2015), which ask students' questions about their teachers' practice (e.g., "When explaining new ideas or skills in class, my teacher tells us about common mistakes that students might make"), and is administered to DCPS classrooms in grades 3 and up. The survey assesses teaching across seven categories: Care, Confer, Captivate, Clarify, Consolidate, Challenge, and Classroom management.

Finally, in addition to the specifics of their evaluation systems, DCPS and Tennessee differ from each other in size and many other characteristics. TEAM is used by nearly the entire state of Tennessee, and therefore includes teachers and schools across a range of settings and demographics. Each year the Tennessee data include roughly 84,000 teachers, of whom 5,500 are in their first-year teaching, with 450,000 students at 1,350 schools. DCPS, on the other hand is an urban majority-minority and low-income district, with approximately 3,500 teachers (290 novice) at 125 schools serving 46,000 students each year.

Other data. In addition to classroom observation data for teachers, we have access to other data for teachers and students. For DCPS and Tennessee teachers, we know when they entered teaching, their experience in teaching, and other socio-demographic characteristics. We have information regarding the observation raters and timing of the five observations. In both settings we have the usual information regarding each teacher's students in tested subjects and grades, including eligibility for free or reduced-price lunch, race, ethnicity, and standardized achievement scores, where applicable.

2. Estimates and causal inference

2.1 Standard estimates of returns to experience

Figure 1 depicts the returns to experience in teaching, using the performance measure of classroom observation ratings. The solid-line estimates use an estimation strategy typical of prior papers on the returns to experience in teaching. Though the estimation strategy is common, in nearly all prior papers the performance measure is teachers' contributions to student test scores (or teacher "value added scores").[4] We come to the dashed-line estimates later.

The core of the typical estimation strategy is to focus on variation within individual teachers over time (Rockoff 2004). The regression specification is:

$$\bar{s}_{jt} = h\big(expr_{jt}\big) + \mu_j + \pi_t + \nu_{jt}, \tag{1}$$

where the outcome variable is a measure of teacher performance. In the Figure 1 estimates, $\bar{s}_{jt}$ is the classroom observation score of teacher j in school year t, scaled in teacher standard deviation units (mean 0, s.d. 1 within state-by-year cells).[5] For a given teacher, years of experience, $expr_{jt}$, is colinear with school year, t, unless she takes a leave of absence. Specification 1 includes both teacher fixed effects, μ_j, and school-year fixed effects, π_t, and thus requires some restriction on $h\big(expr_{jt}\big)$ to avoid the age-period-cohort problem. The typical restriction, which we also use in the Figure 1 solid line, is to assume no returns to experience after some number of years, $\bar{e}$. Then $h\big(expr_{jt}\big)$ is a series of indicator variables for years of experience up to $\bar{e}$:

$$h\big(expr_{jt}\big) = \sum_{e=0}^{\bar{e}-1} \delta^e D_{jt}^e$$

where $D_{jt}^e = \mathbf{1}\{expr_{jt} = e\}$. It is common to choose D_{jt}^1, the first year of teaching, as the omitted category, but we omit the "veterans" category, $D_{jt}^{\bar{e}} = \mathbf{1}\{expr_{jt} \geq \bar{e}\}$, and thus the zero line on the

[4] The exceptions we are aware of, as described in the introduction, are Kraft, Papay, and Chi (2020) and Laski and Papay (2020), which also use classroom observation scores but a somewhat different estimation strategy.

[5] We begin with the item-by-observation level data in the original rubric units (integer scores 1-4 in DCPS and 1-5 in Tennessee); these are the data as recorded by the observers. All of the following steps are carried out separately for DCPS and Tennessee data: (i) We standardize the item-by-observation level scores by year so that each item is mean 0, standard deviation 1. (ii) For each teacher j and item, we average item-by-observation level scores to create an item average score for year t. We then re-standardize the item-average scores. (iii) For each teacher j, we average her annual item-average scores to create the overall annual average score. Finally, we again standardize the overall average scores by year. Thus $\bar{s}_{jt}$ is mean 0 and standard deviation 1 each year, across all teachers in the state of Tennessee (District of Columbia Public Schools), regardless of experience.

Figure 1 y-axis is average veteran performance.[6] The δ^e estimates are plotted in Figure 1's solid line. The vertical lines mark cluster-corrected 95 percent confidence intervals, with teacher clusters.

Our main outcome variable $\bar{s}_{jt}$ is teacher j's "classroom observation score" in year t. More precisely, $\bar{s}_{jt}$ is the average of several task-specific scores, $\bar{s}_{jt} = \frac{1}{K}\sum_{k=1}^{K} s_{jt}^{k}$. The Tennessee rubric includes $K = 19$ items and DCPS $K = 9$. Our focus on the average observation score is motivated by an empirical constraint: While the tasks being scored are distinct—for example "teacher content knowledge" and "managing student behavior"—in practice the scores across tasks are highly correlated. In our Tennessee data, the mean correlation between items is 0.53 with a standard deviation of 0.05; in a factor analysis the first factor explains 95 percent of the variation in item scores. This correlation of items is common in classroom observation rubric scores (e.g., Kane et al. 2011).

In Figure 2 we plot analogous estimates where the performance measure is teachers' contributions to student test scores (also known as teacher value added scores). To obtain these estimates we fit a specification analogous to specification 1:

$$A_{ijst} = h(expr_{jt}) + \mu_j + \pi_t + b(A_{is(t-1)}) + u_{ijst}, \quad (2)$$

where A_{ijst} is the end of year t test score for student i in subject s taught by teacher j. Test scores are in student standard deviation units (mean 0, s.d. 1 within state-by-year-by-subject-by-grade cells). The $h(expr_{jt})$ function is the same as in specification 1, and μ_j and π_t are again teacher and year fixed effects. We continue to estimate standard errors using a cluster (teacher) correction. The function $b(A_{is(t-1)})$ is a flexible function of student i's prior year test score in subject s. Our analysis sample includes grades 4-8 in math and English language arts.[7]

[6] Alternatives to the version of $h(expr_{jt})$ in 2 include: (i) Specifying h as cubic, or other higher-order polynomial, in $expr_{jt}$, though often still with $expr_{jt}$ top-coded at some point (e.g., Rockoff 2004). (ii) Dividing $expr_{jt}$ into bins, e.g., 1–2, 3–4, 5–9, 10–14, 15–24, and 25+ where $\bar{e}$ would have otherwise been = 10 (e.g., Harris and Sass 2011). (iii) Using the non-standard age-experience progressions, e.g., leaves of absence, to estimate specification 1 without restrictions on h (e.g., Wiswall 2013).

[7] Years 2015-16 and 2016-17 are excluded for Tennessee because students were not tested in 2015-16. In Tennessee if the student had two or more teachers in a given subject and year, we include one observation per teacher and weight each observation by the proportion of responsibility allocated by the state to the teacher. Three quarters of students had one teacher in a given subject. If the student's prior year test score is missing, we replace it with zero and include an indicator for missing in the function b.

In Figure 3, we plot estimates where the performance measure is based on student survey responses in DCPS. The Figure 3 estimates use the same teacher-by-year specification 1 as for Figure 1, except for the change in outcome variable. The outcome is teacher j's average survey score in year t; the average across seven survey items known as the "7Cs." This score is scaled in teacher standard deviation units.[8]

2.2 Causal inferences

Labeling estimates like those in Figure 1 as "returns to experience" is a causal claim. The "returns" are the effect of the treatment "experience" on some outcome often left implicit. Threats to that claim depend on the outcome. For example, contrast claims about observation scores *per se* and broader claims about teacher job performance. Causal claims may also be threatened by the statistical properties of the estimators or by substantive institutional details. This section describes some conceptual and econometric structure to aid in evaluating claims about returns to experience estimates.

2.2.1 Difference-in-differences framework

Returns to experience estimates, like those in Figure 1, can be thought of as difference-in-differences estimates. Moreover, the standard estimation strategy, described in Section 2.1, is a two-way fixed effects diff-in-diff estimator. The diff-in-diff framework provides a well-known structure for evaluating causal inference claims.

To see the diff-in-diff features, start by focusing on a simple case: an estimate of (i) the effect of the first year of teaching experience on observation scores, using (ii) data from just two years, $t = 2012$ and $t + 1 = 2013$. This simple case is a classic two-group, two-period ("2x2") diff-in-diff setup. The classic 2x2 graph for this case is shown in Figure 4 by the pair of blue lines at the far left, using data from Tennessee. The solid blue line's two end points (filled circles) plot the mean observation scores for the "treated group" of teachers who had zero experience in 2012 and one year of experience in 2013. The dashed blue line plots a "comparison group" of veteran teachers with 10 or more years of experience by 2012. Figure 4 is measured in the original 5-point Tennessee rubric scale, without any standardizing or regression adjustments. The diff-in-diff

[8] We do not have access to student-level survey responses. We begin with teacher-by-survey-item data on each of seven items, "7Cs," measured by the Tripod survey. (i) Within year, we first standardize the teacher-by-item scores by year so that each category is mean 0 and standard deviation 1 for a given year. (ii) For each teacher j, we then average her annual item-level scores to create the overall annual average score. Finally, we again standardize the overall average scores by year.

estimate is $0.12 = (3.72 - 3.47) - (3.97 - 3.84)$ rubric scale points, or about one-third of a teacher standard deviation. Here the dashed line provides the counterfactual. Assume that there are no returns to additional experience after 10 years. Then the slope of the dashed line reflects trends in scores unrelated to true teacher performance or experience, and thus the dashed slope is a plausible counterfactual for the novice teacher scores if there were no returns to the first year of experience.[9] The rest of Figure 4 shows several 2x2 diff-in-diff plots, one for each cohort of new first-year teachers. And we could apply this exercise to examine the effect of the second year of experience, the third year, and so on.

Our narration of Figure 4 matches the intuition often given for the standard returns to experience estimation strategy, though here we have been more explicit about the diff-in-diff features. What often remains unexplained is how to aggregate the several 2x2 cases into an overall estimate, and in that aggregation the intuition can break down in ways we describe next.

2.2.2 Two-way fixed effects and potential bias

The standard estimation strategy is a two-way fixed effects diff-in-diff estimator, but with multiple treatments. Recall specification 1:

$$\bar{s}_{jt} = h\big(expr_{jt}\big) + \mu_j + \pi_t + \nu_{jt}, \tag{1}$$

and the similar specification 2. The more familiar example of a two-way FE diff-in-diff has a single binary treatment variable in the place of $h\big(expr_{jt}\big)$. Thus, a specification like:

$$\bar{s}_{jt} = \delta D_{jt} + \mu_j + \pi_t + \nu_{jt}.$$

Several recent papers describe the properties of the two-way FE estimator; our analysis of the returns-to-experience case here draws primarily on de Chaisemartin and D'Haultfœuille (2020) and Goodman-Bacon (in-press).[10] A key insight is that the two-way FE estimate, $\hat{\delta}_{fe}$, is a *weighted* average of all the possible classic two-group, two-period ("2x2") diff-in-diff estimates, $\hat{\delta}_{2x2\,(rr')}$, nested in the data. The subscript rr' indexes 2x2 cases with treated group r and comparison group r'.[11] The $\hat{\delta}_{2x2(rr')}$ are analogues to the several 2x2 cases in Figure 4.

[9] Using terms we will introduce in Section 2.2.4, changes in the dashed line represent changes in the function g, if there are no returns beyond year 10.

[10] Other papers in this fast-growing literature include Borusyak and Jaravel (2017), Sun and Abraham (2020), Athey and Imbens (2018), Callaway and Sant'Anna (2020), and Imai and Kim (2020).

[11] Goodman-Bacon (in-press), among others, formalizes this insight in the case of a single binary treatment. The same insight largely holds when there are multiple treatments, as in the current case of $h\big(expr_{jt}\big)$. We discuss the "largely" qualification below, but the insight holds sufficiently enough that the same potential bias is relevant. Each of the 2x2

The two-way FE estimator comes with two key sources of potential bias, and a third when there are multiple treatments. The first potential bias occurs when groups who receive the same treatment but at different times have different treatment effects—heterogeneity across groups. If treatment effects are homogeneous across the 2x2 cases, rr', then the two-way FE weights are irrelevant to the averaging and the weights maximize precision.[12] If treatment effects are heterogeneous, then the weights do not return the true average effect.

Still, even with heterogeneity across groups, the weights still return a sensible estimate in the returns-to-experience case. Because of the precision maximizing objective, the two-way FE weights are increasing in the sample size and treatment variance of a group rr'. Thus, first, the standard estimates will give more (less) weight to larger (smaller) cohorts. Second, the standard estimates will give more (less) weight to more recent (older) cohorts r. This occurs because, for a given treatment D^e, all cohorts r have the same number of pre-treatment years and the number of post-treatment years is increasing with cohort age; this creates greater treatment variance for more recent cohorts.

The second potential bias occurs when the effect of a treatment grows (shrinks) over time—heterogeneity over time within groups. The problem occurs when (i) a comparison group r' received the treatment at some earlier time, and (ii) the treatment effect is growing or shrinking with time. The change in outcomes for r' caused by (ii) is nevertheless counted as counterfactual change and subtracted off the observed change for r to calculate $\hat{\delta}_{2x2(rr')}$. Feature (i) is certainly a feature of the returns-to-experience case: comparison groups are (largely) composed of more-experienced teachers. Identifying Assumption 2, discussed in more detail below, is partly motivated by this second potential bias.

The third potential bias is specific to settings with multiple treatments, and it occurs when the distribution of treatments changes across groups—heterogeneity of treatment probabilities. To

---

estimates takes the classic form $\hat{\delta}_{2x2(rr')} = \left(\bar{y}_r^{post} - \bar{y}_r^{pre}\right) - \left(\bar{y}_{r'}^{post} - \bar{y}_{r'}^{pre}\right)$. However, because group r' might itself have a change in treatment status during run of the available data, the "pre" and "post" periods for rr' need to be defined such that group r's treatment status changes but r''s does not (see Goodman-Bacon in-press for more details).

[12] In this paper's setting, homogeneity across groups is not implausible. Indeed, the example data in Figure 4 show relative homogeneity across cohorts r. The nature of the "treatment" is likely more stable over time, though heterogeneity could arise from changes in starting skill level of novice cohorts. Absent some new intervention targeted specifically to early-career teachers, the returns to experience likely reflect something about skill development which is more fundamental and not changing discontinuously over time.

simplify the exposition, consider the case with just two treatments. In the notation of our setting, the two-way FE specification would be:

$$\bar{s}_{jt} = \delta^1 D^1_{jt} + \delta^2 D^2_{jt} + \mu_j + \pi_t + \nu_{jt}. \quad (3)$$

With multiple treatments the two-way FE estimate, $\hat{\delta}^1_{fe}$, is no longer the same weighted average of all the possible classic 2x2 estimates, $\hat{\delta}^1_{2x2(rr')}$, nested in the data. The difference comes from how the multiple treatment case deals with the co-occurrence (or covariance) of treatments. We can write $\hat{\delta}^1_{2x2(rr')}$ as a function of three key inputs:

$$\hat{\delta}^1_{2x2(rr')} = a\left(\hat{\delta}^{\sim 1}_{2x2(rr')}, \hat{\delta}^{\sim 2}_{2x2(rr')}, \hat{\rho}^1_{2x2(rr')}\right).$$

The first term $\hat{\delta}^{\sim 1}_{2x2(rr')}$ is the 2x2 estimate of treatment 1's effect on $\bar{s}$ *without controlling* for treatment 2, and vice versa for $\hat{\delta}^{2\sim}_{2x2(rr')}$. The third term $\hat{\rho}^1_{2x2(rr')}$ captures the covariance of treatments 1 and 2. Specifically, $\hat{\rho}^1_{2x2(rr')}$ is the estimated probability $\hat{E}\left[D^1 | D^2 = 1, \mu_j, \pi_t\right]$ for group rr'.[13] However, the two-way FE estimator is a precision-weighted average of:

$$\check{\delta}^1_{2x2(rr')} = \check{a}\left(\hat{\delta}^{\sim 1}_{2x2(rr')}, \hat{\delta}^{\sim 2}_{2x2(rr')}, \hat{\rho}^1_{fe}\right),$$

which uses $\hat{\rho}^1_{fe}$ instead of $\hat{\rho}^1_{2x2(rr')}$. The $\hat{\rho}^1_{fe}$ term is itself a two-way FE estimate of ρ^1 from the specification: $D^1_{jt} = \rho^1 D^2_{jt} + \mu_j + \pi_t + \eta_{jt}$. Thus, the third potential bias occurs when there is heterogeneity in $\hat{\rho}^1_{2x2(rr')}$. If there is homogeneity then $\hat{\rho}^1_{2x2(rr')} = \hat{\rho}^1_{fe}$ and $\check{\delta}^1_{2x2(rr')} = \hat{\delta}^1_{2x2(rr')}$, even if there is still heterogeneity in the treatment effects, $\hat{\delta}^1_{2x2(rr')}$. Moreover, heterogeneity in $\hat{\rho}^1_{2x2(rr')}$ may be more problematic than heterogeneity in treatment effects; both $\check{\delta}^1_{2x2(rr')}$ and $\check{\delta}^2_{2x2(rr')}$ are separately "identified" and thus contribute to $\hat{\delta}^1_{fe}$ and $\hat{\delta}^2_{fe}$ even if the two treatments are collinear in group rr'.[14]

For returns-to-experience estimates, the distribution of treatments is the distribution of teacher experience. Thus, the third potential bias is caused by changes over time in the proportion

---

[13] In other words, $\hat{\rho}^1_{2x2(rr')}$ is a diff-in-diff estimate of the "effect" of treatment 2 on the dependent variable D^1.

[14] Our analysis here draws on Goodman-Bacon (in-press); section IV.B discusses "controls" and additional treatments are a specific case of additional controls. The key term in Section IV.B is the term in brackets in equation 26 labeled $\hat{\beta}^d_{b,kl}$. Using our notation (including rr' in place of kl) and a little algebra the $\hat{\beta}^d_{b,kl}$ term can be written:

$$\check{\delta}^1_{2x2(rr')} = \left[\frac{var_{rr'}\left(D^1_{jt}\right) * \tilde{\delta}^{\sim 1}_{2x2(rr')} - var_{rr'}\left(D^2_{jt}\right) * \tilde{\delta}^{\sim 2}_{2x2(rr')} * \hat{\rho}^1_{fe}}{var_{rr'}\left(D^1_{jt} - D^2_{jt} * \hat{\rho}^1_{fe}\right)} \mid \mu_j, \pi_t\right] = \left[\frac{cov_{rr'}\left(\bar{s}_{jt},\ D^1_{jt} - D^2_{jt} * \hat{\rho}^1_{fe}\right)}{var_{rr'}\left(D^1_{jt} - D^2_{jt} * \hat{\rho}^1_{fe}\right)} \mid \mu_j, \pi_t\right],$$

where we are assuming no "within-timing-group" variation.

of teachers in their first, second, third, …, $\bar{e}$th year of teaching. For example, because the state (district) is hiring more or fewer novices from year to year.

2.2.3 Alternative diff-in-diff strategy

The dashed line in Figure 1 uses an alternative identification strategy, designed to avoid the potential bias of the two-way FE approach. Here we are using the DID_M estimator proposed by de Chaisemartin and D'Haultfœuille (2020). The alternative is estimated with:

$$\hat{\delta}^{e}_{didm} = \frac{1}{N^e}\sum_t N^{et}\,\hat{\delta}^{et}_{didm}, \text{ where}$$

$$\hat{\delta}^{et}_{didm} = \left[\frac{1}{N^{et}} \sum_{\substack{j:expr_{j,t}=e,\\ expr_{j,t-1}=e-1}} \left(\bar{s}_{j,t} - \bar{s}_{j,t-1}\right)\right] - \left[\frac{1}{M^{et}} \sum_{j:expr_{j,t-1}\geq \bar{e}} \left(\bar{s}_{j,t} - \bar{s}_{j,t-1}\right)\right] \tag{4}$$

Each $\hat{\delta}^{et}_{didm}$ is a diff-in-diff estimate for treatment e in year t. The first expression in square brackets is the mean first difference in $\bar{s}_{jt}$ for the sample of "treated" teachers, those for whom $expr_{j,t} = e$ and $expr_{j,t-1} = e-1$. In the second pair of square brackets is the mean first difference for the "comparison" veteran teachers, those for whom $expr_{j,t-1} \geq \bar{e}$. The number of treated teachers is N^{et} and comparison teachers is M^{et}.[15]

This alternative strategy avoids the potential biases of the two-way FE strategy. First, the alternative strategy avoids the changing distribution of treatments problem. The comparison group only includes veteran teachers whose treatment status does not change from $(t-1)$ to t by construction.[16] Further treatment status can only change once for the treated group—i.e., from $D^e_{j,t-1} = 0$ to $D^e_{j,t} = 1$—because e cannot change faster than t and the alternative estimator only uses two years $(t-1)$ and t. In other words, the alternative estimator matches the intuitive example of Figure 4. Moreover, second, these same features limit the potential bias from heterogeneity of effects over time. Finally, the alternative strategy weights simply by the sample size of treated teachers, N^{et}, to average across cohorts. Recall the standard two-way FE weights

[15] In practice we estimate each $\hat{\delta}^{et}_{didm}$ with a weighted least squares regression, stacking the several et cases into a system of equations with one equation for each $\hat{\delta}^{et}_{didm}$. The stacked regressions approach allows us to estimate standard errors which are cluster (teacher) corrected across $\hat{\delta}^{et}_{didm}$. We undo the implicit precision-maximizing weights before applying the N^{et} weights as shown in equation 4.

[16] Implicit in "treatment status does not change" is the assumption that there are no returns-to-experience effects beyond $\bar{e}$ years. This assumption is common in the literature on teachers, and we formalize this assumption in Section 2.3.

are also a function of treatment variance, which makes the average estimate biased, or at least difficult to interpret, when effects are heterogenous across cohorts.

The cost of the alternative estimator is a potential loss of precision. However, for our standard errors the first-order consideration is the number of unique teachers (clusters), not the number of teacher-by-year observations. Both the standard and alternative estimators use the same set of unique teachers. Moreover, any change in estimated standard errors is potentially misleading because the precision-maximizing promise of the two-way FE approach requires homogeneity of effects.

Empirically, at least in our setting, conclusions about the returns to experience are not greatly affected by the two-way FE estimator's biases. As Figure 1 shows, the standard and alternative estimates are nearly identical in the Tennessee case. That is not true in the DCPS case where there is a noticeable difference. But two things are important to note. First, that DCPS difference is largely explained by changes in the distribution of teacher experience over time; thus, the difference arises from the two-way FE bias specific to the multiple treatments case. As shown in Appendix Figure A2, the distribution of experience in DCPS has been shifting away from early-career teachers over time, but becoming more stable from 2014-15.[17] If we restrict out analysis to this more-stable more-recent period, the standard and alternative approaches are quite similar, as shown in Appendix Figure A3. Second, the DCPS difference in Figure 1 is more a difference in intercepts, and less a difference in slopes over time. For years 1 through 5-6 the year-to-year changes are nearly identical. That is also the period when returns to experience are the largest. Thus, the experience-distribution bias is not affecting most inferences about change over time.

The two estimation strategies also yield similar estimates when the outcome is teacher contribution to student test scores (or value-added scores), as shown in Figure 2. The same is true when we use the student survey measure of teacher performance, in Figure 3. Smaller sample sizes make these estimates noisier under either strategy, but the patterns across strategies are consistent.

In the remainder of the paper we use the two-way FE estimator. However, the remaining content of the paper applies to both the standard and alternative approaches, including identifying assumptions, threats to those assumptions, and several empirical tests. Moreover, the similarity of

[17] Our data begin in 2009-10 at the start of DCPS's new IMPACT performance evaluation program, which might partly explain the change in experience distribution. However, those years also coincide with the slow labor recovery following the 2008 recession. We do not see the same pattern in Tennessee where the experience distribution has been stable over the years we study.

standard and alternative estimates, after accounting for the DCPS experience distribution changes, suggests other sources of two-way FE potential bias are not first-order.

2.2.4 Inferences from scores to performance

Classroom observation scores, at best, measure only some aspects of a teacher's job. Claims about teacher performance based on observation scores alone require assumptions—in particular, assumptions about the relationships among (i) classroom observation scores, (ii) teachers' actual performance on the job tasks observations claim to measure, and (iii) the student outcomes which are the end goal of schooling. In this subsection we provide a conceptual framework to organize these relationships.

A teacher's primary job responsibility is to improve her students' outcomes. Let μ^y measure a teacher's true contribution to (equivalently, causal effect on) student outcome y. Outcomes like math and reading skills, social and emotional skills, labor market success, citizenship, health, or any other end goal of schooling. Thus, claims about teacher performance are often claims about μ^y. However, in practice, measuring μ^y is difficult, even when y itself is observable. That difficulty is demonstrated by the case of "value-added scores" which estimate a teacher's contribution to student test scores.[18] While we will sometimes use student test scores as y in this paper, this conceptual framework is meant to be general to any y.

If μ^y summarizes a teacher's output, then μ^y is a function of that teacher's performance in many different job tasks:

$$\mu^y = f^y\left(\theta^1, \theta^2, \ldots, \theta^{k'}, \theta^{k'+1}, \ldots, \theta^K\right), \tag{5}$$

where k indexes teacher job tasks. Each task is a unit of work that produces an input to y. Let θ^k measure the teacher's performance in task k. Higher performance in task k—a higher value of θ^k—is synonymous with producing more or higher-quality task k inputs.

Classroom observation rubrics are often loosely described as measuring inputs to μ^y. More sharply, observation rubrics are designed to assess a teacher's actions and decisions in selected job tasks. In other words, observation rubrics are designed to measure θ^k. For example, observers are asked to score the nature and frequency of questions teachers ask students, but observers are not asked to assess whether these questions generated student learning. Observation scores are also

[18] Still, value-added scores are estimates, and a value-added score for one individual teacher is generally quite noisy. For a recent review of the literature on teacher job performance, including measurement, see Jackson, Rockoff, and Staiger (2014).

sometimes described as measures of teachers' skills. We prefer to think of θ^k as performance, which is a function of both skills and effort. Improvement in scores might reflect improvement in skills, or improvement in effort, or both. Rubrics, like the ones we study, typically measure both skills and effort.[19]

Crucially, however, observation scores do not measure all the θ^k which contribute to μ^y. This is the first key feature of the relationship between observation scores and student outcomes. To emphasize this feature, the expression in 5 partitions tasks into two groups: let tasks $\{1,2,\ldots,k'\}$ be tasks the rubric is designed to measure, while tasks $\{k'+1,\ldots,K\}$ are not measured by the observation rubric.

A second key feature is that observation scores are only estimates of θ^k. Let s^k be a rubric score which is designed to measure the latent variable of teacher performance in task k:

$$s_{jt}^k = g^k\left(\theta_{jt}^k\right) + \varepsilon_{jt}^k, \tag{6}$$

where j indexes teachers and t indexes time. The ε^k term is measurement error, with $E\left[\varepsilon_{jt}^k\right] = 0$.

The function g^k represents the evaluation process. The specific form of g^k is unknown, to both the econometrician and the evaluation designer. However, in general terms, g^k is determined first by the explicit features of the evaluation process—e.g., the rubric itself, how evaluators are trained, how evaluators are assigned to teachers, incentives. Such explicit features are (mostly) controllable by those designing and implementing the evaluation. But g^k also includes less-explicit, less-controllable features—e.g., the behaviors teachers or evaluators choose in response to the explicit features. A commonly used alternative to 6 would be $s_{jt}^k = \theta_{jt}^k + \epsilon_{jt}^k$, where all deviations of s from θ are "error" and $E\left[\epsilon_{jt}^k\right] = 0$. We prefer using the function g^k to emphasize the role of the evaluation process itself. Even in large samples, that process will create some difference between $E\left[s_{jt}^k\right]$ and $E\left[\theta_{jt}^k\right]$.

Using data on s^k to make claims about θ^k requires some assumptions about g^k. Our focus in this paper is inferences about how $E\left[\theta_{jt}^k\right]$ changes with experience on the job. Thus our focus is on assumptions about how g^k does or does not change over time or across groups. We describe those identifying assumptions in Section 2.3 below.

[19] While these observation rubrics are designed to measure teacher performance, there is some evidence that empirically they also measure student behavior (Campbell and Ronfeldt 2018). We address this issue in Section 3 as a threat to causal inferences about teacher performance.

However, all claims about θ^k based on s^k require a more fundamental assumption about g^k—an assumption about measurement validity. An ideal condition would be that g^k is such that $corr(s_{jt}^k, \theta_{jt}^k) = 1$, after accounting for measurement error, ε_{jt}^k. While this ideal is unlikely to hold in practice, the correlation may be sufficiently large so that the benefits of using s^k to inform some management decision (or research claim) outweigh the costs of mistakes in inferences about θ^k. Equation 5 suggests one partial test is a predictive validity test: $corr(\bar{s}_{jt}, \mu_{jt}^y)$. Our estimate of that correlation is 0.38 for Tennessee data and 0.30 for DCPS, when μ^y is the teacher's value-added contribution to math and reading test scores.[20] Those estimates of 0.38 and 0.30 are likely to be too small. First, there is the common attenuation because of measurement error. Second, the simple mean $\bar{s}_{jt}$ gives equal weight to each task score s_{jt}^k. But it seems unlikely that the elasticity of μ^y with respect to θ^k is equal for all tasks, k. In other words, if we knew the education production function f^y, we would likely choose un-equal weights. Thus, claims about differences in μ^y based on $\bar{s}$ will likely understate true differences even when $corr(s^k, \theta^k) = 1$.[21]

2.3 Identifying assumptions

Under what identifying assumptions can the estimates in Figure 1 be interpreted as causal returns to experience—specifically the effect of experience on performance, θ, of the tasks which the rubric is designed to measure? The underlying diff-in-diff structure suggests a parallel-trends-style assumption: roughly, that the difference between novice and veteran scores, s, would be constant over time if neither novice nor veteran performance, θ, improved with experience. But we can sharpen the assumptions by using the conceptual framework from subsection 2.2.4.

Interpreting Figure 1 as the causal returns to experience requires two identifying assumptions. *Assumption 1:* The function g does not depend on experience, i.e., $g(\theta, expr) = g(\theta)$. For example, this assumption requires that if an early-career and a veteran teacher have the

[20] Appendix Table A1 reports estimates from regressions where the dependent variable is a student test score, A_{ijst} as in equation 2, and the key predictor variable is the teacher's average observation score, $\bar{s}_{jt}$ as in equation 1. In column 2, the coefficient on $\bar{s}_{jt}$ is 0.081 in Tennessee and 0.098 in DCPS. That coefficient is the predicted increase in teacher value added, μ_{jt}^y, for an increase in teacher observation score, $\bar{s}_{jt}$, of one teacher standard deviation; that coefficient is not the $corr(\bar{s}_{jt}, \mu_{jt}^y)$ because value added is measured in student test score standard deviations. For Tennessee the correlation is $0.38 = 0.08/0.21$, where 0.21 is the estimated standard deviation in value added. For DCPS that denominator is 0.33.

[21] Additionally, the scoring rubrics used in Tennessee and DCPS were both adapted from the well-established *Framework for Teaching* (Danielson 1997), which is based on a carefully-articulated conception of teaching and learning (Dwyer and Villegas 1993), empirical studies (Myford et al. 1994), and a design process that followed established standards (AERA, APA and NCME 2014).

same true task performance, θ, they will have the same observation score, s. *Assumption 2:* True performance is not changing over time, on average, in the comparison group of teachers, i.e., $E\left[\theta_{jt} - \theta_{j(t-1)} | expr_{jt} \geq \bar{e}\right] = 0$.

The value of the comparison group, and thus the second difference, is shown by stating the identifying assumptions that would be required for a first-difference estimate of returns to experience. If we used only early-career teachers' data, we could not separate the returns to experience from changes in g over time, because, as mentioned above, $expr$ and t are colinear within teacher. Estimates based on first differences alone cannot use Assumption 2, and instead require the stronger *Assumption 3:* The function g does not change over time—i.e., $g(\theta, t) = g(\theta)$. In the next section we discuss different substantive threats to these identifying assumptions, but some of the quite-plausible threats are known changes in g over time.

3. Alternative explanations and threats to causal inference

Observation scores may improve (decline) over time for reasons unrelated to a teacher's gains from experience. In this section we describe several alternative explanations for changing scores, and whether an alternative explanation threatens a causal "returns to experience" interpretation of Figure 1. We focus specifically on interpreting changes in observation scores as the causal effect of experience on performance of the tasks which the rubric is designed to measure.

Before taking up specific alternative explanations, we begin with some general evidence relevant to the plausibility of identifying Assumptions 1 and 2. First, Figure 5 reports a partial test of Assumption 1. For this test assume that g does not depend on experience (Assumption 1), and that the f^y function also does not depend on experience. (Recall that f^y relates teacher task performance, θ, to teacher contributions to student outcomes, μ^y.) If both g and f^y are unrelated to experience, then we would expect that the relationship between observation scores, $\bar{s}$, and teacher contributions, μ^y, should be unrelated to experience.

Figure 5 shows, for teachers with e years of experience (x-axis), the predicted increase in test-score value added, μ^y, for a one standard deviation increase in observation score, $\bar{s}$ (y-axis).[22]

[22] To obtain the estimates in Figure 5, we fit a student-test-score regression similar to specification 2, but where $h(expr_{jt})$ is replaced with

$$h(expr_{jt}, \bar{s}_{jt}) = \gamma^{\bar{e}} \bar{s}_{jt} + \sum_{e=0}^{\bar{e}-1} \delta^e D_{jt}^e + \gamma^e (\bar{s}_{jt} \times D_{jt}^e)$$

The solid line uses only within-teacher over-time variation (by including teacher FE), and the dashed line uses both within- and between-teacher variation (omitting teacher FE). To get a sense of the *correlation* between $\bar{s}$ and μ^y, multiply the y-axis by about 5 for Tennessee and 3 for DCPS (see footnote #).

The relationship between $\bar{s}$ and μ^y is (largely) unrelated to experience in Figure 5. With perhaps one exception, there is no clear trend related to experience. And we cannot reject the null that each point estimate is different from the series average, though the DCPS estimates are quite noisy. The exception is the earliest years in Tennessee using only within-teacher variation (solid line series). Those estimates suggest the correlation declines from the first year to the fourth, but then remains stable afterward. Some of the specific threats described below could be a mechanism behind the declining correlation. That decline in correlation could be evidence that g does depend on experience, but it is not necessarily evidence against Assumption 1. Even if the function f^y does not depend on teacher experience, the optimization of f^y to maximize μ^y may depend on experience. For example, perhaps as early-career teachers gain experience they shift more effort to tasks not measured by the observation rubric, i.e., $k = \{k' + 1, \ldots, K\}$, or more subtly shift across tasks in a way not well captured by the simple average $\bar{s}_{jt}$.

Additionally, Figure 5 is only a partial test. We have only one y outcome: teacher contributions to student test scores. Teachers contribute to other important student outcomes, like social and behavioral skills (Jackson 2018), and classroom practices are likely important to those outcomes as well. Related, in DCPS we have student surveys which may capture inputs to test and non-test student outcomes. Appendix Figure A4 repeats the test in Figure 5 with the surveys as outcomes, and we find steady, albeit noisy, correlations between classroom observation scores and the student survey scores.

We can also partially test identifying Assumption 2. That assumption requires that true performance, θ,is not changing over time, in expectation, among the comparison group of teachers, i.e., $E\left[\theta_{jt} - \theta_{j(t-1)} | expr_{jt} \geq \bar{e}\right] = 0$. Our main estimates in Figure 1 set $\bar{e} = 10$ to define the veteran group. If Assumption 2 holds, then our estimates for returns at $e =$ 1-9 should be robust

which includes the interactions $\bar{s}_{jt} \times \mathbf{1}\{expr_{jt} = e\}$ on the right-hand side. Figure 5 plots $(\gamma^e + \gamma^{\bar{e}})$ for each level of experience, e.

to setting $\bar{e}$ above 10. Appendix Figure A5 shows estimates with $\bar{e} = 15$ and $\bar{e} = 20$, and for $e =$ 1-9 the estimates are quite similar to Figure 1. Additionally, while we cannot observe $\Delta\theta = E\left[\theta_{jt} - \theta_{j(t-1)} | expr_{jt} \geq \bar{e}\right]$ directly, we can estimate $\Delta g(\theta) = E\left[g\left(\theta_{jt}\right) - g\left(\theta_{j(t-1)}\right) | expr_{jt} \geq \bar{e}\right]$. Among veteran teachers, the mean first-difference in observation scores is 0.004 standard deviations (st.err. 0.002) in Tennessee and -0.073 standard deviations (st.err. 0.006) in DCPS.[24]. Under what conditions would $\Delta g(\theta) \cong 0$ but $\Delta\theta \neq 0$? Only in the knife-edge case where any change in true performance, θ, is just offset by a change in the g function.*3.1 The evaluation system*

Changes in scores over time may be caused by changes to the evaluation system's tools and procedures. Key features of an evaluation system include the scoring rubric, the training provided to raters, and the rules for assigning teachers to raters.[25] Even if a teacher's performance, θ, remains constant, the score assigned to that performance, s, may go up or down if the system's processes change. In other words, the evaluation system's tools and procedures are key features of the function g. (The incentives or consequences attached to scores are also a key feature of an evaluation system; we return to incentives below.)

The most straightforward example is a change in the scoring rubric. In 2017 DCPS switched from the Teaching and Learning Framework (TLF) rubric to an entirely new Essential Practices (EP) rubric. The new rubric did not measure exactly the same set of tasks, k, as the old rubric. Other rubric changes might be smaller, like word choices, even if the tasks scored remain the same.[26] However, rubric changes, big or small, would not necessarily threaten our identifying

[24] In DCPS, compositional differences in the teaching force over time (Dee and Wyckoff 2015, Dee et al. 2021, James and Wyckoff 2020) could make it appear, with our preferred within-year standardization process, as if experienced teachers were declining over time as the average performance of incoming teachers improves. However, relying on alternative standardization approaches, including standardizing relative to veteran teachers within year and standardizing scores across years, do not change the slopes shown in Figure 1. Differences in point estimates across standardization approaches never exceed 0.037, with an average difference in point estimates across approaches and levels of experiences of 0.005. In rubric units, the average first difference for veteran teachers is also quite small, at -0.014 (st.err. 0.003).

[25] Our language and examples in this discussion mainly imply the evaluation systems designed or used by schools, districts, or states. The features and reasoning also apply to scores collected by researchers or for other purposes.

[26] Our focus here is comparisons over time within a sample of teachers. Rubrics also vary across samples. DCPS and Tennessee use different observation rubrics, and those two rubrics are designed to cover different sets of teaching tasks. DCPS employs different rubrics to assess teachers with different sets of tasks, e.g., general education teachers have a somewhat different rubric than early-childhood or special-education teachers. More generally, school systems and researchers use a wide variety of observation rubrics, which should caution against strong comparisons of observation scores across systems or studies. Early-career experience may improve performance on a task which is measured by one rubric but not another. However, if we take a broader notion of teacher performance, like

assumptions, as long as the changes affect early-career (treatment) and veteran (comparison) teachers equally.

The DCPS changes allow us to compare estimates from different rubrics. In Figure 6 the short dash line shows estimates of returns to experience using only data generated by the TLF, while the long dash line uses only EP. Both dashed lines are limited to scores from school administrators. For both rubrics the average first-year teacher's rating is much lower than the average veteran's rating, but that starting gap is smaller with the EP rubric. In both cases teachers make larger improvements over the first five years compared to the next five, but the improvements are steeper using the TLF rubric. The differences suggest a potential threat to Assumption 1—that g does not depend on experience—at the time of the change in rubrics in DCPS. However, the difference between the dashed (TLF) and long-dashed (EP) estimates could be a compositional change. Starting in 2011, and thus concurrent with our data, DCPS became more selective in both hiring and retention decisions, with strategies based explicitly on performance measure (Dee and Wyckoff 2015, Jacob et al. 2018). There were noticeably fewer early-career teachers by 2017 (Appendix Figure A2). Thus, in Figure 6, the higher scores with the EP rubric may reflect true higher performance because of selection.

Choosing raters is also a key evaluation design decision, which itself may change over time. Figure 6 also compares estimates by rater type for DCPS. The solid red line uses only ratings from the master educator raters, who are external to the school, while the dashed red line uses only ratings from school administrators. Both lines are limited to scores generated by the TLF rubric. The two TLF lines are quite similar, especially over the first five years of a teacher's career. Additionally, in this comparison there is no composition change concern since each teacher was rated by both a school administrator and master educator each school year. Figure 6 does obscure one important difference between master educator scores and school administrator scores. School administrators give higher scores on the 1-4 scale; in other words, the g function does depend on rater type. However, the difference in scores between the rater types is the same for all teachers regardless of experience; thus, the rater type difference in g does not violate Assumption 1.[27]

contributions to student achievement, many rubrics which measure different tasks are similarly good predictors of teacher contributions to test scores (Kane and Staiger 2012).

[27] Figure 6 does not show intercept differences between the three series; each series is estimated separately with the veteran intercept set to zero. Appendix Figure A7 shows the estimates which allow for the intercept comparison.

In general, changes to the evaluation system are changes to the function g. Interpreting Figure 1 as the causal returns to experience does not require that g remain unchanged. The only restriction on g is that g not depend on experience. This applies to obvious changes in g, like the rubric or types of evaluators, and to changes which are more difficult (for the researcher) to observe. One potentially difficult to observe change is to the training of raters. Imagine that system administrators determine, at a given point in time, that raters need to be re-trained on some aspect of scoring. That re-training might be in fact be motivated by administrators' belief that scores, s, are not reflecting performance, θ, as they should. A second example is a change to the rules for assigning teachers to raters. Chi (2020), among others, has documented teacher-rater match effects on scores; when a teacher and rater share a gender or race, the teacher's scores are higher. Imagine the evaluation system administrators decide, at some point, to make gender or race an explicit factor in the rules for making assignments.

3.2 Behavior of the raters

Changes in scores over time may reflect changes in the behavior of the raters. Raters have some discretion within any evaluation system's designed procedures. Observation scores fall somewhere in between the theoretical poles of objective evaluation and subjective evaluation. Raters may also take actions which violate the designed procedures they were trained to follow. The behavior of raters, whether intended or unintended in the system design, is part of the function g.

One behavior that is frequently cited, given rater discretion, is leniency bias—the tendency for raters to give scores which are higher than warranted. The histograms in Appendix Figure A1 would be consistent with systematic leniency bias in observation scores, although such bias is less evident for scores assigned by the master educators in DCPS. The skew in Appendix Figure A1 could also accurately reflect teacher performance using a rubric with ceiling effects. Leniency bias is often cited as a concern in classroom observation scores by both researchers and in public debate (Kraft and Gilmour 2017, Anderson 2013).

However, leniency bias does not necessarily threaten our causal interpretation of Figure 1 as returns to experience. To violate Assumption 1—g does not depend on experience—rater leniency would need to be correlated with teacher experience. For example, imagine that raters are less lenient with a first-year teacher compared to their rating of the same teacher in her second year; then Figure 1 would over-state the returns to the first year of teaching. Such a change in

leniency might be a mechanism behind the early-years decline in Tennessee in Figure 5. However, if it is not correlated with experience, leniency bias will be differenced out in the same way as rubric changes or other evaluation system features.

Another potential mechanism is that raters may use information learned outside a given observation. Consider the case of a teacher scored by her school principal. A few brief classroom observations are a small fraction of the interactions a teacher and principal will have in a school year; the principal likely learns much about the teacher's performance outside of formal observations. Ho and Kane (2013) show evidence that a teacher's own principal scores a video of her classroom differently than a principal from another school in the district scores the same video, perhaps because the teacher's own principal begins the scoring with a prior on the teacher's performance. Additionally, because the rubric covers only some teaching tasks, $k = \{1,2, \dots, k'\}$, a principal may raise (lower) observation scores to reflect the principal's beliefs about the teacher's performance of tasks not covered by the rubric, $k = \{k', \dots, K\}$. A principal using outside information is a potentially rational behavior if the observation scores are used for personnel decisions and the principal cares about μ^y and not $\bar{s}$.

This outside information explanation may threaten Assumption 1—g does not depend on experience—but only if raters have and use different outside information depending on a teacher's years of experience. The number of years a teacher-principal pair has worked together may well be correlated with the teacher's years of experience, but it does not need to be strongly correlated if school principals switch schools frequently. A high correlation would suggest principal raters might have different outside information on early-career and veteran teachers. Empirically the correlation is 0.17 in the DCPS data and 0.15 in the Tennessee data.

We can carry out a test relevant to this outside-information question. Figure 7 shows an event study for a change in school principal. The estimates are from a regression which begins with specification 1, with added controls for year relative to a change in the school's principal. Further the time series is allowed to differ for early-career and veteran teachers. If principals learn about a teacher's performance outside of formal classroom observations, we might expect observation scores to rise or fall. However, scores do not change on average as a principal and teacher work together longer. This pattern holds for both early-career and veteran teachers. In

Tennessee there is some evidence that principals give slightly lower scores in their first year in a new school.[28]

3.3 Incentives and distortion of effort

Changes in scores may reflect changes in the incentives attached to those scores. Still, a change in incentives alone does not threaten inferences about true performance, θ^k, for tasks scored by the rubric. A new or stronger incentive attached to task k's score, s^k, might well induce a teacher to raise her performance of that task, θ^k, through more effort for task k or investing in skills for task k. Indeed, raising θ^k is typically the goal of attaching incentives to s^k. Alternatively, a new or stronger incentive might induce more manipulation behavior of the kind discussed in the next subsection. In short, apart from manipulation, a change in incentives does not threaten casual claims about the effect of experience on performance, θ^k, for tasks scored by the rubric, $k \in \{1,2,\dots,k'\}$.

However, a change in incentives for s^k can threaten inferences about μ^y performance. Recall that μ^y is a teacher's contribution to one or more student outcomes, y, like math achievement, social skills, or earnings as an adult. As discussed earlier, observation rubrics comprise only some of the teaching tasks which are inputs to μ^y. Thus, the act of evaluation creates incentives to give more effort or attention to those scored tasks, $\{1,2,\dots,k'\}$ in f^y, and less effort to other un-scored tasks, $\{k'+1,\dots,K\}$ in f^y. Those incentives might be formally linked to scores, like monetary bonuses or the threat of dismissal, or more general career concerns. This asymmetry between scored tasks and un-unscored tasks is a case of the well-known multi-task distortion problem (Holmstrom and Milgrom 1991).

Using evaluation and incentives to shift teacher effort away from some tasks and toward other tasks is not necessarily distortion. There is (quasi-)experimental evidence that rubric-based classroom observations can improve teachers' contributions to student test scores, even when teachers are not evaluated based on those test scores (Taylor and Tyler 2012, Briole and Maurin 2020, Burgess, Rawal, and Taylor in-press). In DCPS specifically, teacher performance improves

[28] On additional note on rater behavior. As described in Section 1.2, the item level observation scores for specific tasks s^k are strongly correlated, in these data and most teacher observation data. This fact is sometimes interpreted as evidence that raters do not actually differentiate between tasks, k, but instead score teachers on some single general dimension of teaching performance. This seems unlikely given that the item level correlations are not = 1. A more plausible explanation is that the rubrics define tasks where true performance is in fact strongly correlated. Whatever the explanation, this issue is not central to our analysis in this paper which focuses on the average score. This issue does limit our ability to make conclusions about how experience may affect tasks differentially.

more when the teacher spends more of the year anticipating an unannounced rater visit (Phipps 2018, Phipps and Wiseman 2021).

While incentives do not necessarily threaten our causal interpretation of Figure 1, changes in incentives may be a mechanism behind the improvements seen in Figure 1. The simplest example is tenure rules. In Tennessee, teachers can earn tenure after five years, but tenure requires sufficiently high observation scores in years four and five.[29] Thus, teachers have somewhat more incentive to focus effort on the rubric-measured tasks in years four to five compared to one to three, which might contribute to the pattern in Figure 1. Still, it seems unlikely a teacher concerned about tenure would wait until year four to pay attention to the rubric, and the slope from years three to four in Figure 1 is not obviously a departure from the trend suggested by the other year-to-year slopes.

Unlike Tennessee, for most of the period of this analysis, the evaluation incentives in DCPS were not explicitly a function of years of experience, but could be correlated with experience. DCPS teachers are dismissed if rated "Minimally Effective" (the second-lowest rating) in two consecutive years or if they fail to exceed a "Developing" rating (the third-lowest rating) within three consecutive years. Before fall 2012, teachers could receive permanent salary increases after two consecutive years of being rated "Highly Effective" (the top rating). Figure 8 shows the proportion of teachers in each rating category by years of experience, suggesting the incentives are not strongly correlated with experience.[30]

3.4 Manipulation of scores

Observation scores may reflect changes in teachers' actions unrelated to their job performance. Teachers, like professionals in any other occupation, may adopt behaviors or actions which do raise their scores, s, but do not raise their job performance, θ. In the literature on job performance evaluation these actions are known as manipulation.[31] This manipulation of observation scores might occur, for example, because classroom observations are infrequent and brief; thus, a teacher could prepare a special lesson or even rehearse the lesson with his students

[29] More precisely, tenure requires being rated "4. Effective" or "5. Highly Effective" on the 1-5 integer scale. While only one input to that rating, classroom observation scores get a weight of 50-85 percent for the teachers.

[30] Also studying DCPS, Adnot (2016) reports evidence that teachers facing the two-consecutive-years-minimally-effective dismissal threat shift effort across tasks within the rubric toward tasks which are more likely raise their scores. This is a sort of distortion within measured tasks, but suggests that teachers are aware of this margin.

[31] Empirical examples of manipulation by teachers include cheating on student tests (Jacob and Levitt 2003) and intentionally excluding low-scoring students from high-stakes tests (Jacob 2005, Cullen and Reback 2006, Figlio 2006, Figlio and Getzler 2006).

in advance of the rater's visit. By contrast, if the evaluation process or incentives prompted a teacher to improve her lessons on all (many of) the days the rater would not be present, that would be an improvement in performance and not manipulation.

Manipulation plausibly threatens our casual returns-to-experience interpretation of Figure 1. In our framework, teacher manipulation results from the evaluation system's procedures and incentives, and is part of the function g. A teacher's awareness of how to manipulate likely grows as he gains experience with the evaluation system—e.g., extraordinary preparation in response to an announced visit or the use of a "lesson in a box" in response to an unannounced visit. That suggests a plausible correlation between manipulation and general teaching experience, which threatens Assumption 1 that g is invariant to experience. However, that correlation might be weakened if more-experienced teachers share their manipulation strategies with newly-hired teachers.[32] If the manipulation component of observation scores is unrelated to general experience, then manipulation will be differenced out in Figure 1.

The decline in correlation over years 1-4 in Tennessee in Figure 5 may be explained by increasing manipulation over the first few years of a teacher's career. However, we cannot rule out other mechanisms, such as, for example, raters becoming more lenient as a teacher moves from first to second to third year. And there are other limitations to the test in Figure 5, as discussed above. On the other hand, while underpowered, the evidence in DCPS (panel B) does not indicate a decline in the relationship between classroom observation scores and student achievement over experience. In addition, the relatively stable correlation between classroom observation scores and student survey scores across levels of teaching experience in DCPS (Appendix Figure A4) provide evidence against the presence of manipulation, unless teachers were similarly able to manipulate scores on both measures across levels of experience.

Dee and Wyckoff (2015) examine whether DCPS school administers manipulate observation scores, s, in the face of increased incentives. Consider the teachers who received their first Minimally Effective rating in 2010-11, and thus were under a significant threat of dismissal during 2011-12. Observation scores did improve in 2011-12 for these teachers, on average.[33]

[32] Strong or widespread manipulation would threaten the basic informativeness assumption about g.

[33] These improvements in 2011-12 were likely the result of being rated Minimally Effective (ME) in 2010-11. The improvements for ME teachers were larger than improvements for teachers rated Effective in 2010-11. The category Effective was the next-highest category in those years. Dee and Wyckoff (2015) provide regression-discontinuity estimates using the cutoff between Minimally Effective and Effective.

However, master educators also scored these teachers as having improved, and the increase in observation scores was similar across both types of raters. Additionally, these teachers under dismissal threat also improved on their test-score value added, μ^y. Taken together, these results suggest that the dismissal threat did not improve observation scores through manipulation alone.

3.5 Changes in job assignments

Changes in a teacher's scores may reflect changes in her job assignment. A teacher's observation score, s, might decline (improve) after a job change for either of two reasons: First, the teacher's actual performance, θ, could decline (improve) because of the job change. Using student-test-score measures of μ^y, Ost (2014) provides evidence that teaching skills and experience are not fully transferable across grade levels. Switching from 3rd to 5th grade, for example, likely requires some adjusting of questioning techniques, or shifting effort to new lesson plans at the expense of in-class performance.

Let a and a' be two different job assignments; θ_{jta} is the actual performance of teacher j in school year t and job assignment a (omitting the k superscript to simplify). We can write:

$$E\left[\theta_{jt} - \theta_{j(t-1)}\right] = \underbrace{E\left[\theta_{jta} - \theta_{j(t-1)a}\right]}_{\Delta^t} + p \times \underbrace{E\left[\theta_{j(t-1)a} - \theta_{j(t-1)a'}\right]}_{\Delta^a}, \quad (6)$$

where p is the probability of switching from job a' to a.

The intuitive notion of "returns to experience" implies that the job is constant and experience increases, which matches Δ^t in expression 6. If identifying Assumption 2 holds—no returns for veterans—then Figure 1 reports estimates of $(\Delta^t + p\Delta^a)$. Assuming further that job changes reduce performance, $\Delta^a < 0$, then Figure 1 underestimates the intuitive Δ^t, a negative bias. Alternatively, some researchers or policymakers may be interested $(\Delta^t + p\Delta^a)$, which we could describe as the "returns to experience including job changes typical of early-career teachers."

Job changes do threaten identifying Assumption 2, which requires that $E\left[\theta_{jt} - \theta_{j(t-1)} | expr \geq \bar{e}\right] = 0$ in our comparison group of veteran teachers. A veteran's performance might change because of a job change, $\Delta^a \neq 0$, even if her performance would not have otherwise changed, $\Delta^t = 0$. If job changes do reduce comparison teacher performance, $\Delta^a < 0$, then the estimates in Figure 1 overstate the intuitive Δ^t for novices. This bias is positive, and the bias

described in the prior paragraph is negative, but the two would only cancel each other out under the assumption that p and Δ^a do not depend on experience.[34]

The second reason scores might change is that the function g might differ across jobs. For example, typically the same rubric is used for all teachers, leaving any adaptation to grade-level or subject circumstances up to the rater or training process. More subtly, the function g might depend on the students in the classroom (Campbell and Ronfeldt 2018). Students are themselves an important feature of a teacher's job assignment, and a feature which can change even if grade level or subject do not. The threat to identification parallels other features of g discussed above. As long as job-specific differences in g are unrelated to experience, this second reason is not a serious threat to identification. A job-specific difference might be, for example, if raters are more lenient with novices after a job change than they are with veterans.

In Figure 9 we test the robustness of Figure 1 to changes in the students a teacher is assigned, even if she remains in the same subject and grade level. Using data from Tennessee and DCPS, we plot returns-to-experience estimates with and without controls for student baseline test scores.[35] Accounting for changes in students assigned does not affect our estimates. The similarity of all the estimates in Figures 1 and 9 is partly because they all use only within-teacher variation. The function g might well depend on the students in the classroom (Campbell and Ronfeldt 2018), but most of the variation in students assigned is between teachers or schools, not within teachers over time.

3.6 Performance improvements among veteran teachers

The actual performance of veteran (comparison-group) teachers may change over time—violating Assumption 2—even if there are no returns to experience for veterans. For example, veterans may increase their effort in response to incentives. How would interpretation change if Assumption 2 was violated in this way, but Assumption 1 held? If the veteran gains were only among veterans, then the estimates in Figure 1 would likely understate the true returns to

[34] This assumption is sufficient, but not strictly necessary. We only require that the product $p\Delta^a$ not depend on experience, which should be a weaker assumption.

[35] These estimates come from a student-level regression. The specification is identical to (2) except that the dependent variable is $\bar{s}_{jt}$ instead of A_{ijst}. The dashed-line series in Figure 9 omits the $b(A_{i(t-1)})$ controls. The solid-line series includes $b(A_{i(t-1)})$.

experience for early-career teachers. The veterans' improvements would be subtracted off any improvements for early-career teachers.[36]

3.7 Turnover

One final consideration in interpreting Figure 1 is turnover or attrition from our estimation sample. The estimates in Figure 1 use only within-teacher variation in observation scores. This feature addresses a first-order potential bias: average observation scores might rise with experience, even if each individual teacher's scores remain constant, if lower-scoring teachers are more likely to leave teaching (or at least leave the district or state).

Still, even with teacher fixed effects, Figure 1 is still partly determined by turnover. In Figure 1 the slope between year one and year two is an average of $N_{1,2}$ different individual teacher slopes, where $N_{1,2}$ is the sample of individuals who are observed in year one and year two (and perhaps future years). Similarly, the slope between year four and year five uses only the $N_{4,5}$ sample. However, these are not the same samples: $N_{4,5} \neq N_{1,2}$. First, for any given cohort of novice hires, attrition from the profession over time will make $N_{4,5} \subset N_{1,2}$. Second, experienced teachers who transfer into the system from elsewhere may contribute to $N_{4,5}$ even if they do not contribute to $N_{1,2}$. The slope from year one to year two in Figure 1 might be different if we could estimate it with the $N_{4,5}$ sample.

Empirically, however, our Figure 1 estimates are not strongly influenced by this second-order composition concern. Figure 10 shows our returns-to-experience estimates using subsamples defined by when the teacher leaves teaching in Tennessee or DCPS. Much of the difference is attributable to differences in intercepts. The changes from year one to two, two to three, etc. are quite similar across samples. The exception is that scores decline in a teacher's final year before leaving teaching in the state or district.

4. Conclusion

We conclude that the typical estimates of returns to experience, applied to observation scores, can reasonably be interpreted as the causal effect of additional experience on teachers' job

[36] This subtraction might be desirable in specific cases. Imagine, for example, that veterans improved because of some new training, and that training was given to all teachers, early-career and veteran. If, roughly, the effect of the training was similar for all teachers, then the subtraction makes the Figure 1 estimates returns to experience controlling for any general training effects.

performance—specifically, performance of the input tasks covered by the rubric. The typical estimates are effectively difference-in-differences estimates, where veteran teachers are the comparison group. Veterans provide a plausible counterfactual estimate for several often-stated threats, including for example, leniency bias from raters, manipulation by teachers, changes in the evaluation system, and changes in teachers' job assignments. Our estimates are robust to changes in the rubric, different rater types, and controlling for student baseline achievement, among other things. Still, there are reasons to remain cautious about a causal interpretation. We find, in one setting, a weakening correlation between teacher observation scores and student test scores as teacher experience grows. That weakening is consistent with some threats to the identifying assumptions, but it would also be consistent with changes in optimal teaching strategies as experience increases.

Finally, our primary focus is on inferences about returns to experience as measured by classroom observation used in two settings. Our analyses should be interpreted carefully. We focus on the performance of the input tasks covered by the rubric. Stronger assumptions are required for inferences about teacher performance measured by contributions to student outcomes. Taking differences in scores over time addresses several concerns which are left unaddressed in results based on score levels at a single point in time.

References

Adnot, Melinda. 2016. "Teacher Evaluation, Instructional Practice, and Student Achievement: Evidence from the District of Columbia Public Schools and the Measures of Effective Teaching Project." PhD diss., University of Virginia, Charlottesville.

American Educational Research Association [AERA], the American Psychological Association [APA], and the National Council on Measurement in Education [NCME]. 2014. *The Standards for Educational and Psychological Testing.* Washington, DC: AERA.

Anderson, Jenny. 2013. "Curious Grade for Teachers: Nearly All Pass." *New York Times,* March 30.

Athey, Susan, and Imbens, Guido W. 2018. "Design-Based Analysis in Difference-in-Differences Settings with Staggered Adoption." NBER Working Paper 24963.

Atteberry, Allison, Susanna Loeb, and James Wyckoff. 2015. "Do First Impressions Matter? Predicting Early Career Teacher Effectiveness." *AERA Open* 1(4):1–23.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting Event Study Designs." Working paper.

Briole, Simon, and Éric Maurin. 2020. "There's Always Room for Improvement: The Persistent Benefits of Repeated Teacher Evaluations." Paris School of Economics working paper.

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. In press. "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools." *Journal of Labor Economics.*

Callaway, Brantly, and Pedro H. C. Sant'Anna. 2020. "Difference-in-Differences with Multiple Time Periods*." Journal of Econometrics.* Forthcoming.

Campbell, Shanyce L., and Matthew Ronfeldt. 2018. "Observational Evaluation of Teachers: Measuring More Than We Bargained For?" *American Educational Research Journal* 55(6): 1233–67.

Chi, Olivia L. 2020. "A Classroom Observer Like Me: The Effect of Demographic Congruence between Teachers and Raters on Observation Scores." Working paper.

Cohen, Julie, and Dan Goldhaber. 2016. "Building a More Complete Understanding of Teacher Evaluation Using Classroom Observations." *Educational Researcher* 45(6): 378–87.

Cullen, Julie B., and Randall Reback. 2006. "Tinkering Toward Accolades: School Gaming Under a Performance Accountability System." In *Improving School Accountability: Check-Ups or Choice,* Advances in Applied Microeconomics 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 1–34. Bingley, UK: Emerald Group Publishing Limited.

Danielson, Charlotte. 1997. "Enhancing professional practice: A framework for teaching." Alexandria, VA: Association for Supervision and Curriculum Development.

de Chaisemartin, Clément, and Xavier d'Haultfoeuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review* 110(9): 2964-96.

Dee, Thomas S., Jessalynn James, and Jim Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from DCPS." *Education Finance and Policy*. Forthcoming.

Dee, Thomas S., and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from Impact." *Journal of Policy Analysis and Management* 34(2): 267–97.

Dwyer, Carol A., and Ana M. Villegas. 1993. "Guiding Conceptions and Assessment Principles for the Praxis Series: Professional Assessments for Beginning Teachers." Educational Testing Service Research Report RR-93-17. doi.org/10.1002/j.2333-8504.1993.tb01528.x

Ferguson, Ronald F., and Charlotte Danielson. 2015. "How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching." In *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project*, ed. Thomas J. Kane, Kerri A. Kerr, and Robert C. Pianta, 98-143. San Francisco, CA: Jossey-Bass.

Figlio, David N. 2006. "Testing, Crime and Punishment." *Journal of Public Economics* 90(4-5): 837–851.

Figlio, David N., and Lawrence Getzler. 2006. "Accountability, Ability, and Disability: Gaming the System?" In *Improving School Accountability: Check-Ups or Choice,* Advances in Applied Microeconomics 14, edited by Timothy J. Gronberg and Dennis W. Jansen, 35–49. Bingley, UK: Emerald Group Publishing Limited.

Goe, Laura, Courtney Bell, and Olivia Little. 2008. *Approaches to Evaluating Teacher Effectiveness: A Research Synthesis*. National Comprehensive Center for Teacher Quality.

Goodman-Bacon, Andrew. 2020. "Difference-in-Differences with Variation in Treatment Timing." Working paper.

Grissom, Jason A., and Brendan Bartanen. 2019. "Strategic Retention: Principal Effectiveness and Teacher Turnover in Multiple-Measure Teacher Evaluation Systems." *American Educational Research Journal* 56(2): 514–55.

Harris, Douglas N., and Tim R. Sass. 2011. "Teacher Training, Teacher Quality and Student Achievement." Journal of Public Economics 95(7-8): 798-812.

Ho, Andrew D., and Thomas J. Kane. 2013. "The Reliability of Classroom Observations by School Personnel." Research paper. Measures of Effective Teaching project," Bill & Melinda Gates Foundation.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, & Organization* 7(1): 24-52.

Imai, Kosuke, and In Song Kim. 2018. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data*." Political Analysis* 1-11.

Jackson, C. Kirabo, Jonah E. Rockoff, and Douglas O. Staiger. 2014. "Teacher Effects and Teacher-Related Policies." *Annual Review of Economics* 6(1): 801–25.

Jacob, Brian A. 2005. "Accountability, Incentives, and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(5–6): 761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 166(1): 81–97.

Jacob, Brian A., Jonah E. Rockoff, Eric S. Taylor, Benjamin Lindy, and Rachel Rosen. 2018. "Teacher applicant hiring and teacher performance: Evidence from DC Public Schools." *Journal of Public Economics* 166:81-97.

Kane, Thomas J., Kerri Kerr, and Robert Pianta. 2014. *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project.* San Francisco, CA: Jossey-Bass.

Kane, Thomas J., and Douglas O. Staiger. 2012. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Measures of Effective Teaching project, Bill & Melinda Gates Foundation research paper.

Kraft, Matthew A., and Allison F. Gilmour. 2017. "Revisiting The Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness." *Educational Researcher* 46(5): 234–49.

Kraft, Matthew A., and John P. Papay. 2014. "Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience." *Educational Evaluation and Policy Analysis* 36(4):476–500.

Kraft, Matthew A., John P. Papay, and Olivia L. Chi. 2020. "Teacher Skill Development: Evidence from Performance Ratings by Principals." *Journal of Policy Analysis and Management* 39(2): 315–47.

Laski, Mary, and John Papay. 2020. "Understanding the Dynamics of Teacher Productivity Development: Evidence on Teacher Improvement in Tennessee." Presentation, 2020 Association for Public Policy Analysis and Management Fall Research Conference.

Myford, Carol, Ana Maria Villegas, Anne Reynolds, Roberta Camp, Charlotte Danielson, Jacqueline Jones, Joan Knapp, Penny Lehman, Ellen Mandinach, Lori Morris, Alice Sims-Gunzenhauser, and Barbara Sjostrom. 1994. "Formative Studies of Praxis III: Classroom Performance Assessments an Overview" Research Report RR-94-20. https://doi.org/10.1002/j.2333-8504.1994.tb01593.x

Ost, Ben. 2014. "How Do Teachers Improve? The Relative Importance of Specific and General Human Capital." *American Economic Journal: Applied Economics* 6(2): 127–51.

Papay, John P., and Matthew A. Kraft. 2015. "Productivity Returns to Experience in the Teacher Labor Market." *Journal of Public Economics* 130(1): 105-119.

Phipps, Aaron R. 2018. "Incentive Contracts in Complex Environments: Theory and Evidence on Effective Teacher Performance Incentives." PhD diss., University of Virginia, Charlottesville.

Phipps, Aaron R., and Emily A. Wiseman. 2021. "Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation." *Education Finance and Policy*. Forthcoming.

Rockoff, Jonah H. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94(2): 247-252.

Rowan, Brian, and Stephen W. Raudenbush. 2016. "Teacher Evaluation in American Schools" In *Handbook of Research on Teaching,* 5, edited by Drew H. Gitomer and Courtney A. Bell, 1159-1216. Washington, DC: American Education Research Association.

Steinberg, Matthew P., and Matthew A. Kraft. 2017. "The Sensitivity of Teacher Peformance Ratings to the Design of Teacher Evaluation Systems." *Educational Researcher* 46(7): 378–96.

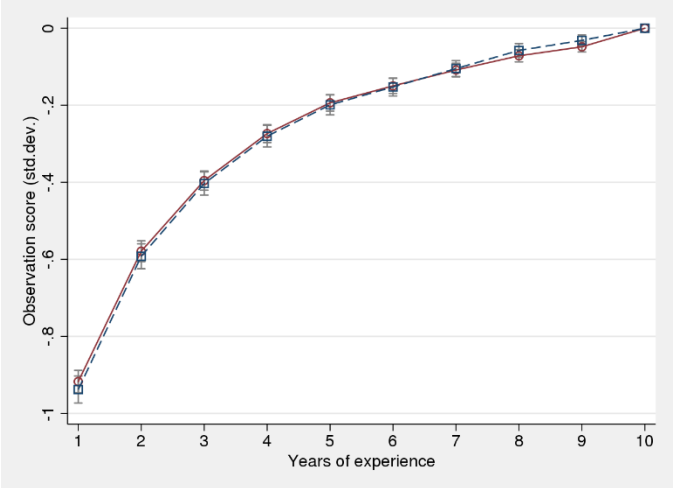
Sun, Liyang, and Abraham, Sarah. 2020. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics*. Forthcoming.

Taylor, Eric, and John Tyler. 2012. "The Effect of Evaluation on Teacher Performance." *The American Economic Review* 102(7): 3628–51.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, David Keeling, Joan Schunck, Ann Palcisco, and Kelli Morgan. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness*. New Teacher Project.

Wiswall, Matthew. 2013. "The Dynamics of Teacher Quality." *Journal of Public Economics* 100: 61-78.

Panel A. Tennessee



Panel B. DCPS

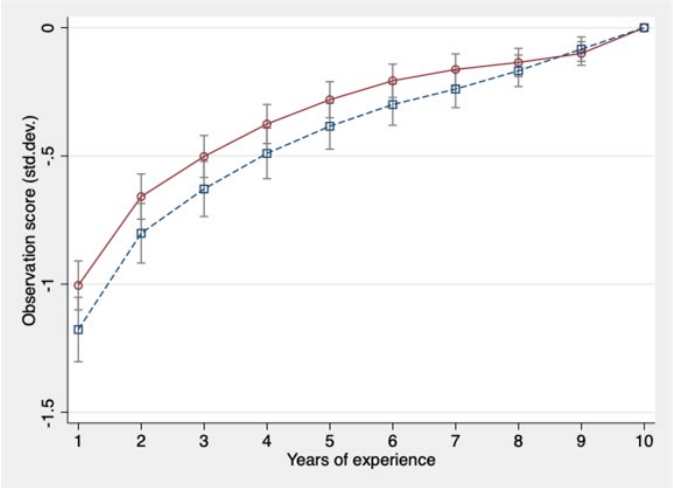
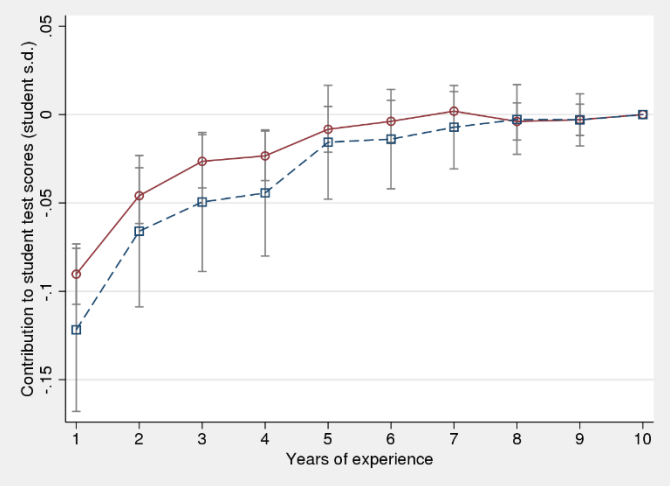


Figure 1—Returns to experience measured in classroom observation scores

Note: The solid line reports estimates using the two-way fixed effects approach described in Section 1.2. The dashed line reports estimates using the alternative diff-in-diff strategy described in Section 2.3. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). In both cases the outcome variable is teacher j's classroom observation score, $\bar{s}_{jt}$, which is an average of several item-level scores recorded during a given school year t. Observation scores are standardized (mean 0, st.dev. 1) by school year using the distribution of all teachers in the jurisdiction, Tennessee or DCPS respectively. For the solid line estimates we fit a single two-way fixed effects regression, with teacher j and school year t fixed effects. The specification includes indicators for years of experience 1 through 9 individually, with ≥ 10 years the omitted category, but no other controls. The plotted points are the coefficients on the experience indicators. The dashed line estimates are the difference between two means: (a) The average first-difference, $(\bar{s}_{jt} - \bar{s}_{j,t-1})$, among "treated" teachers—those with e years of experience (x-axis) in school year t, and $e-1$ years in school year $t-1$. (b) The average first-difference, $(\bar{s}_{jt} - \bar{s}_{j,t-1})$, among "comparison" teachers—those with ≥ 10 years of experience in both year t and $t-1$. The (a) minus (b) second-difference is calculated separately for each unique combination of e and t in the data. Then the plotted points are the weighted average across t for a given e, where the weights are the number of "treated" teachers. The sample size for the solid line in Tennessee is 375,072 teacher-by-year observations for 81,847 unique teachers; and similarly 349,920 and 66,156 for dashed line Tennessee, 33,484 and 7,268 for solid line DCPS, and 33,040 and 7,201 for dashed line DCPS.

Panel A. Tennessee



Panel B. DCPS

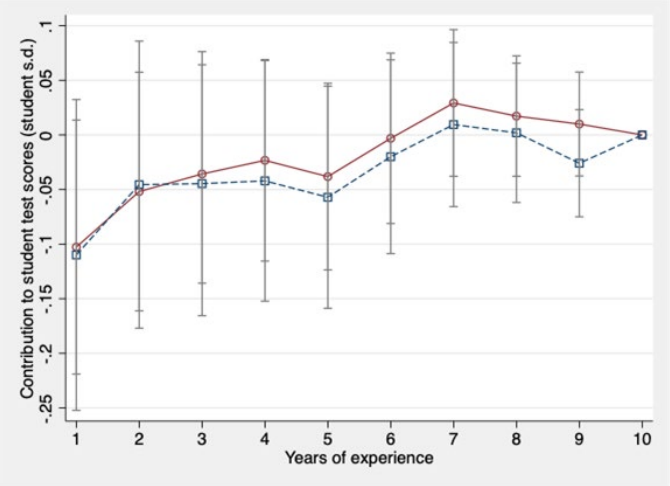


Figure 2—Returns to experience for contributions to student achievement

Note: The solid line reports estimates using the two-way fixed effects approach described in Section 1.2. The dashed line reports estimates using the alternative diff-in-diff strategy described in Section 2.3. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). In both cases the outcome variable is student i's test score, A_{ijst}, in subject s and school year t. Test scores are standardized (mean 0, s.d. 1) within each grade-by-subject-by-year cell using the distribution for all students in the jurisdiction, Tennessee or DCPS respectively. For the solid line estimates we fit a single two-way fixed effects regression, with teacher j and school year t fixed effects. The specification includes indicators for years of experience 1 through 9 individually, with ≥ 10 years the omitted category. Additional controls are a quadratic in prior-year test score, where the parameters are allowed to differ across grade-by-subject-by-year cells, $b\left(A_{is(t-1)}\right)$. The plotted points are the coefficients on the experience indicators. For the dashed line estimates, we begin by estimating teacher contributions to student test scores, $\hat{\mu}_{jt}$. We fit a regression of student scores A_{ijst} on the same prior score controls, $b\left(A_{is(t-1)}\right)$, and teacher fixed effects; and then obtain the residuals $A_{ijst} - \hat{b}\left(A_{is(t-1)}\right)$. Our estimate $\hat{\mu}_{jt}$ is the average residual for teacher j in year t. The dashed line estimates are the difference between two means: (a) The average first-difference, $\left(\hat{\mu}_{jt} - \hat{\mu}_{j,t-1}\right)$, among "treated" teachers—those with e years of experience (x-axis) in school year t, and $e-1$ years in school year $t-1$. (b) The average first-difference, $\left(\hat{\mu}_{jt} - \hat{\mu}_{j,t-1}\right)$, among "comparison" teachers—those with ≥ 10 years of experience in both year t and

$t-1$. The (a) minus (b) second-difference is calculated separately for each unique combination of e and t in the data. Then the plotted points are the weighted average across t for a given e, where the weights are the number of "treated" teachers. The sample size for the solid line in Tennessee is 4,222,939 student-by-subject-by-year observations and 92,403 teacher-by-year observations for 34,395 unique teachers; and similarly 71,474 and 20,954 for dashed line Tennessee, 247,005, 5,413 and 2,268 for solid line DCPS, and 4,249 and 1,280 for dashed line DCPS.

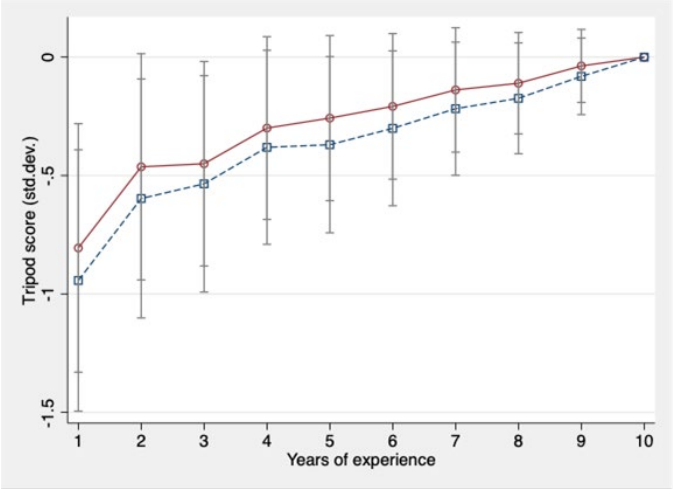


Figure 3—Returns to experience measured in scores from student surveys (DCPS)

Note: The solid line reports estimates using the two-way fixed effects approach described in Section 1.2. The dashed line reports estimates using the alternative diff-in-diff strategy described in Section 2.3. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The details of estimation are identical to Figure 1 except that the outcome variable in Figure 3 is based on student survey responses to the Tripod survey. The dependent variable is the teacher j's Student Surveys of Practice (SSoP) score for school year t. SSoP scores are standardized (mean 0, s.d. 1) by school year using the distribution for all teachers in DCPS. The survey was administered to all DCPS students in grade 3 and above from 2016-17 to 2018-19. The sample size for the solid line is 4,406 teacher-by-year observations for 1,687 unique teachers, and similarly 4,312 and 1,640 for the dashed line.

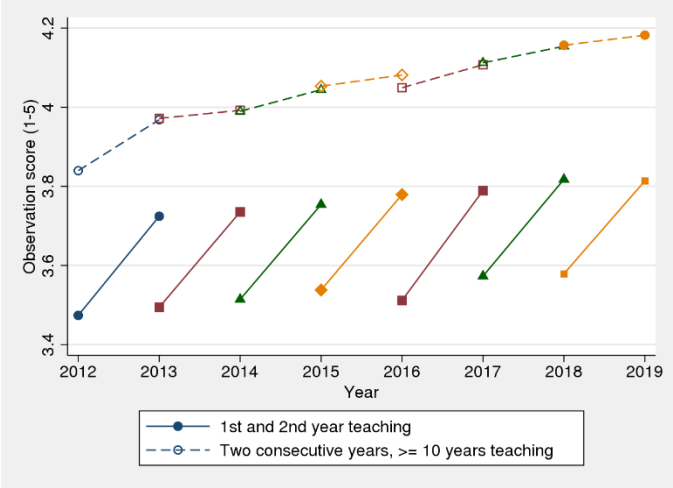
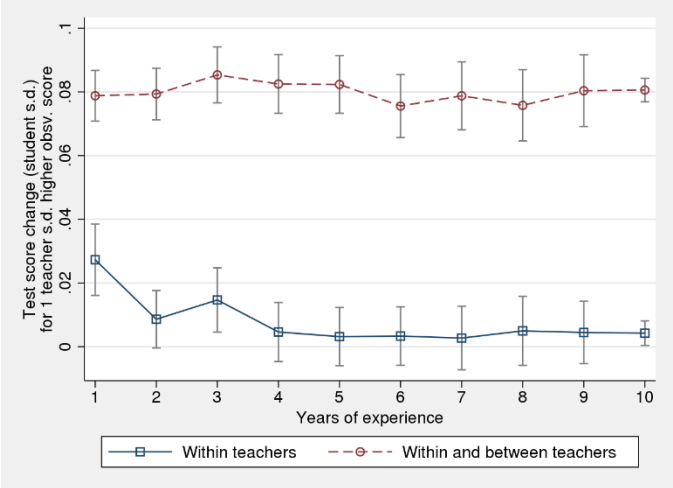


Figure 4—Illustrating 2x2 difference-in-differences components (Tennessee)

Note: Each plotted point is the average observation score for a group of teachers, measured in original rubric units without any adjustments. Consider the solid blue line at the far left. The filled circle markers are teachers whose first year teaching was 2012, and the two marker points are mean observation scores for those teachers in their first year, 2012, and second year, 2013. The dashed blue line at far left, with unfilled circle markers, are teachers with 10 or more years of experience as of 2012. Each pair of lines, matched by color and marker shape, replicates this for cohorts hired in 2013-2018.

Panel A. Tennessee



Panel B. DCPS

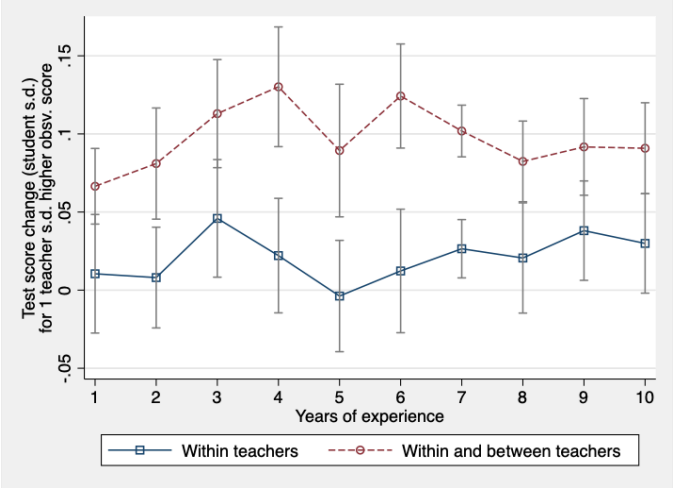


Figure 5—Predicting student test scores with teacher observation scores by years of teacher experience

Note: The solid and dashed lines each report estimates from a separate linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). In both cases the outcome variable is student i's test score, A_{ijst}, in subject s (maths or English language arts pooled) and school year t. Test scores are standardized (mean 0, s.d. 1) within each grade-by-subject-by-year cell using the distribution for all students in the jurisdiction, Tennessee or DCPS respectively. In both cases the specification includes (a) indicators for years of experience 1 through 9 individually, with ≥ 10 years the omitted category; (b) classroom observation score, $\bar{s}_{jt}$; and (c) the interactions of (a) and (b). Each plotted point is sum of the coefficient on the (a)*(b) interaction for e years of experience (x-axis) plus the main-effect coefficient on (b). Additional controls are a quadratic in prior-year test score, where the parameters are allowed to differ across grade-by-subject-by-year cells, $b(A_{is(t-1)})$. The solid line specification includes year and teacher fixed effects. The dashed line includes only year fixed effects, omitting the teacher fixed effects. The sample size the same for the two lines; in Tennessee 4,222,939 student-by-subject-by-year observations and 92,403 teacher-by-year observations for 34,395 unique teachers, and similarly in DCPS 252,400, 5,429, and 2,274.

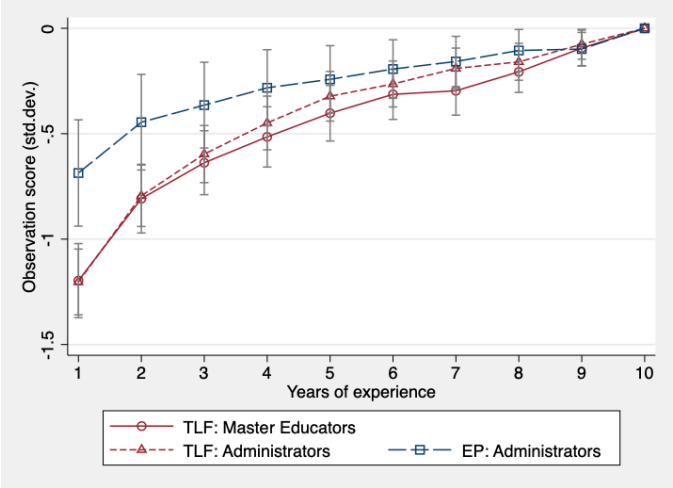
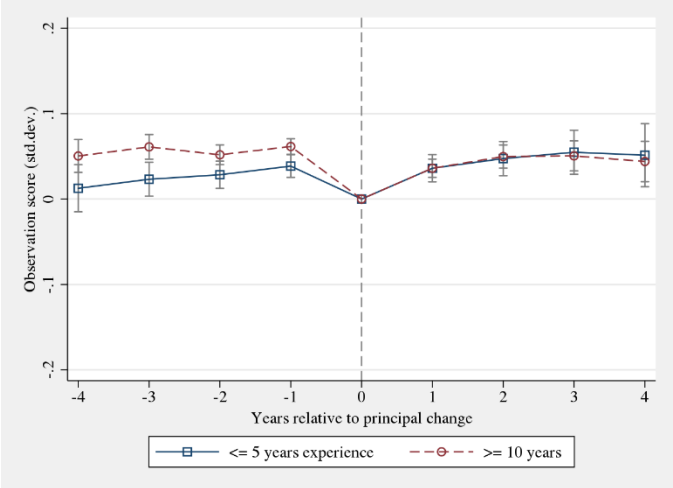


Figure 6—Estimates using different rubrics and rater types (DCPS)

Note: Each of the three lines reports estimates from a separate linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The details of estimation are identical to the solid line in Figure 1 with the following exceptions. First, the estimation sample is limited by the type of rater: external "Master Educators" for the solid line, and school administrators for the dashed and long dashed lines. Second, the estimation sample is limited by the rubric used: TLF from 2010-2016 and EP from 2017-2019. The sample size for the solid line is 18,715 teacher-by-year observations for 5,118 unique teachers; and similarly 21,080 and 5,380 for dashed line, and 10,190 and 3,726 for the long dash line.

Panel A. Tennessee



Panel B. DCPS

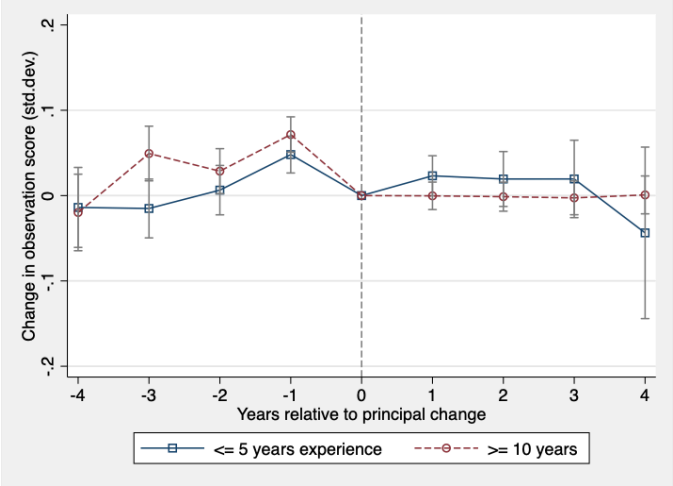


Figure 7—Event study of a change in school principal

Note: All estimates are from a single linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The dependent variable is teacher j's classroom observation score, $\bar{s}_{jt}$, which is an average of several item-level scores recorded during a given school year t. Observation scores are standardized (mean 0, st.dev. 1) by school year using the distribution of all teachers in the jurisdiction, Tennessee or DCPS respectively. The specification includes (a) indicators for year relative to a change in school principal; (b) an indicator $= 1$ if teacher j has ≤ 5 years of experience, and $= 0$ if teacher j has ≥ 10 years; and the interaction of (a) and (b). The new principal's first year, x-axis $= 0$, is omitted for both groups defined by (b). The specification also includes indicators for years of experience, with ≥ 10 years omitted, plus teacher and year fixed effects. If a teacher experiences two (or more) principal changes, we stack the data to include each teacher-by-event-study case in the data. DCPS observation scores in Panel B represent administrator-assigned scores only, but can include multiple administrators (i.e., principals and assistant principals) within a given teacher-year. The sample size for the solid line in Tennessee is 72,850 teacher-by-year observations for 29,193 unique teachers; and similarly 136,443 and 32,244 for dashed line Tennessee, 6,927 and 2,511 for solid line DCPS, and 9,597 and 2,406 for dashed line DCPS.

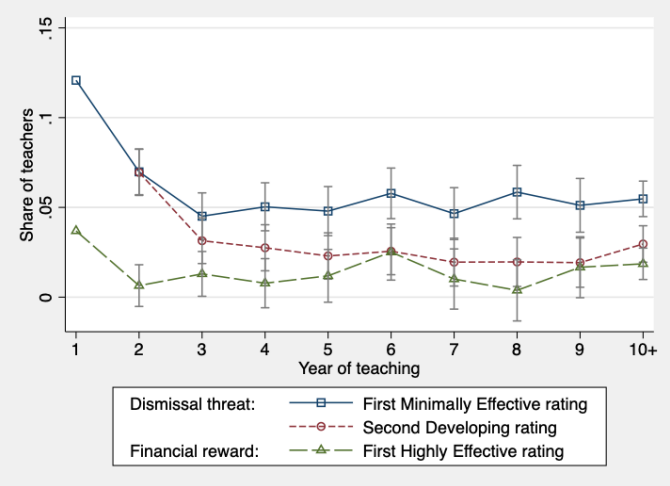
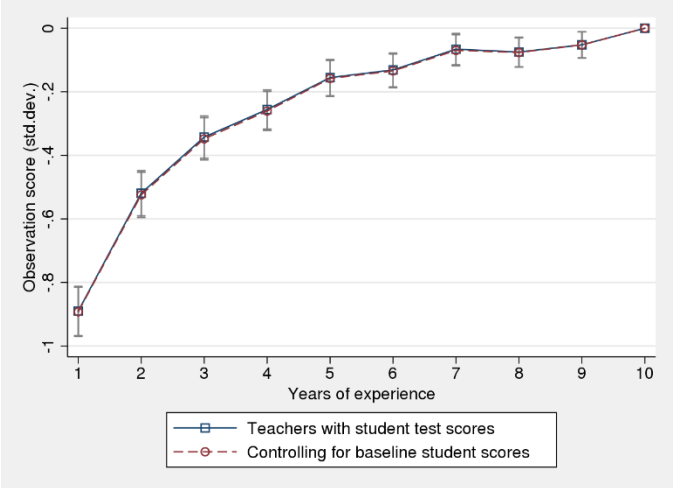


Figure 8—Incidence of consequential performance ratings (DCPS)

Note: Each plotted series reports the percentage of teachers scoring at the relevant consequential rating level. In DCPS, teachers who receive their first Minimally Effective rating must improve the following year or risk dismissal. Beginning in 2012-13, teachers who have earned a second consecutive Developing rating are likewise subject to dismissal if they fail to improve. Through spring 2012, Highly Effective teachers were conversely eligible for large financial rewards. The share of teachers facing each performance incentive are estimated only within the respective years in which the incentive was in place. The sample for the solid line includes 35,672 teachers-by-year and 9,455 unique teachers; and similarly for the dashed line 22,344 and 6,936, and for the long dashed line 10,004 and 4,755.
.

Panel A. Tennessee



Panel B. DCPS

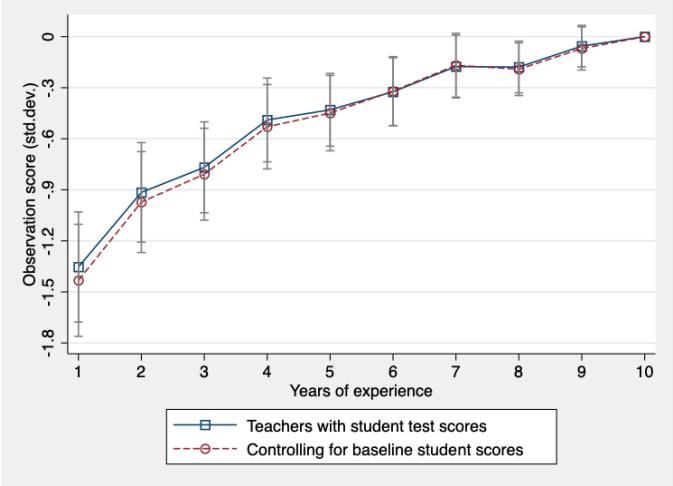
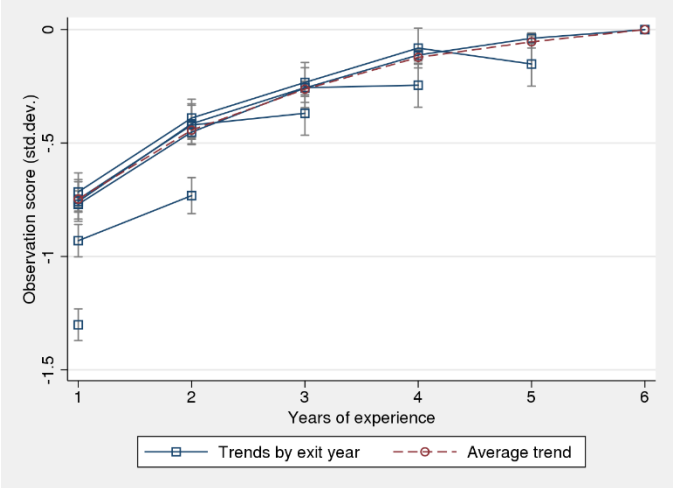


Figure 9—Estimates controlling for student baseline test scores

Note: The solid and dashed lines each report estimates from a separate linear regression, but with the same estimation sample. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). For the dashed line "controlling for baseline student scores" estimates, the specification is identical to the two-way fixed effects regression in Figure 2 but the dependent variable is different. The dependent variable is the teacher observation score (the dependent variable in Figure 1) for the student i's teacher j in subject s and year t. Just as in Figure 2, the observations are student-by-subject-by-year; and the controls are teacher experience indicators, student prior test scores $b(A_{is\,(t-1)})$, and teacher and year fixed effects. For the solid line "teachers with student test scores" estimates, all details are identical to the dashed line except that the solid line omits the prior test score controls $b(A_{is(t-1)})$. The sample size the same for the two lines; in Tennessee 3,076,946 student-by-subject-by-year observations and 65,750 teacher-by-year observations for 25,017 unique teachers, and similarly in DCPS 250,377, 5,369 and 2,258.

Panel A. Tennessee



Panel B. DCPS

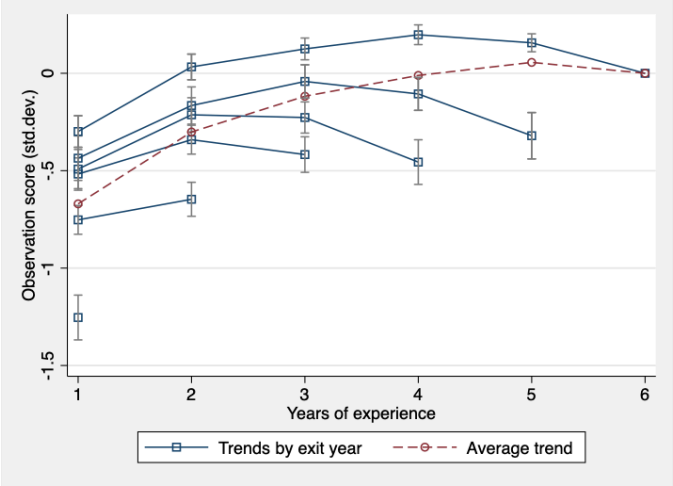


Figure 10—Estimates by year of exit

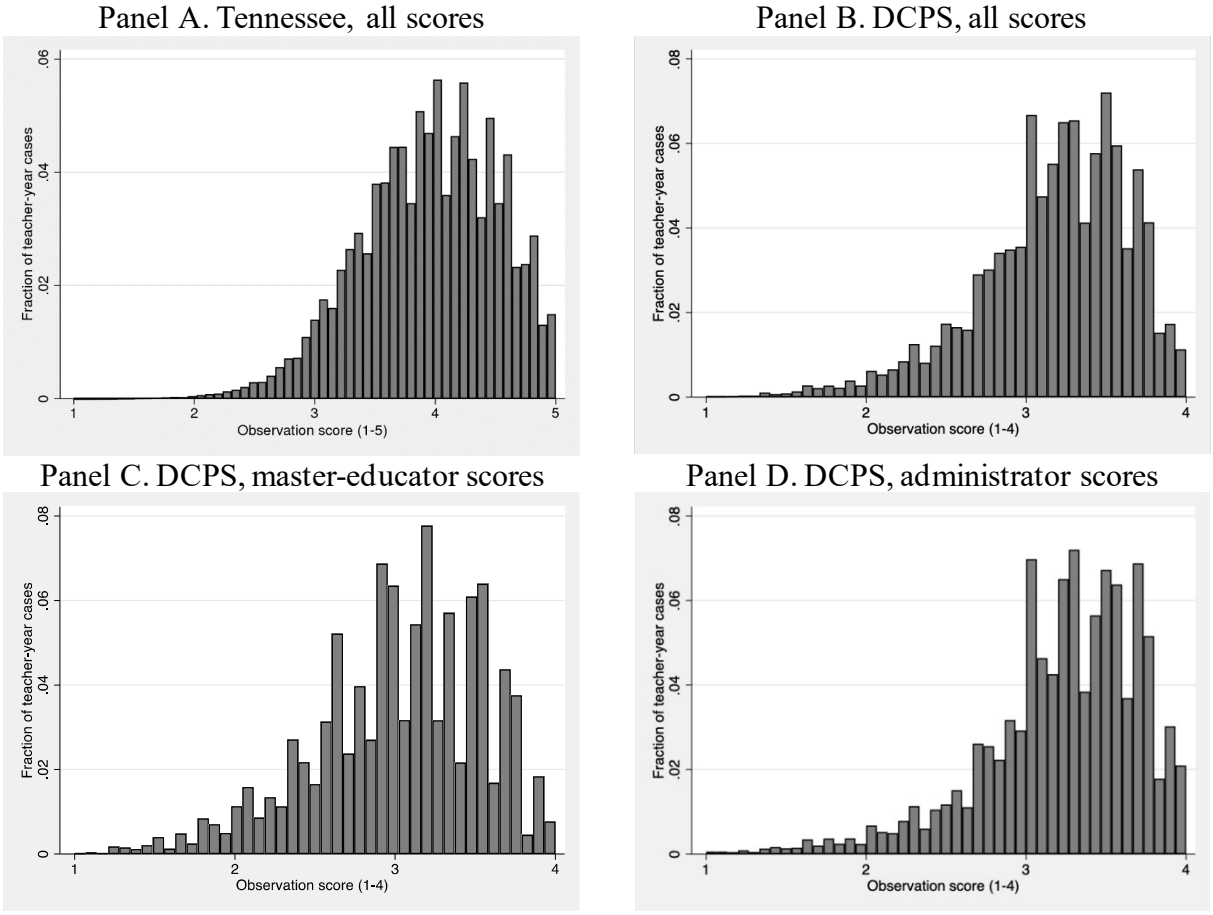
Note: Each panel reports estimates from two separate linear regressions, but with the same estimation sample. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The first regression is shown in the circles and dashed lines. For this "average trend" the details of estimation are identical to Figure 1, with one exception. The top code indicator is for experience ≥ 6 years (instead of ≥ 10 years). The second regression is shown in the five series that use squares and solid lines. For these "trends by exit year" the details of estimation are identical to the "average trend" regression, with the following exceptions: We divide teachers into five sub-samples defined by when they exited the state (district). We then interact the experience indicators with indicators for exit year. The sample size the same for the two series; in Tennessee 27,853 teacher-by-year observations for 6,613 unique teachers, and similarly in DCPS 31,785 and 8,931.

Table 1—Characteristics of the two samples

	Tennessee	DCPS
	(1)	(2)
(A) Students		
At or above proficiency on NAEP		
Math, grade 4	0.39	0.31
Math, grade 8	0.30	0.18
Reading, grade 4	0.34	0.27
Reading, grade 8	0.32	0.20
Race/ethnicity		
Black	0.22	0.64
Hispanic	0.09	0.18
White	0.64	0.13
Other or multiple race or ethnicity	0.05	0.04
Urbanicity		
City	0.34	1.00
Suburb	0.25	0.00
Town	0.14	0.00
Rural	0.27	0.00
Share of school-age population in poverty	0.22	0.28
English language learner	0.04	0.10
Special Education	0.13	0.17
(B) Teachers		
Observation score (original units)	3.94	3.17
	(0.57)	(0.47)
Observation score, administrators	3.94	3.22
	(0.57)	(0.49)
Observation score, master educators		3.02
		(0.53)
In student test score sample	0.23	0.15
Female	0.79	0.74
Race/ethnicity		
Black	0.06	0.51
Hispanic	0.00	0.05
White	0.86	0.32
Other or multiple race or ethnicity	0.08	0.04
Graduate degree	0.55	0.69
Years of experience		
Mean	11.83	10.86
Standard deviation	(9.61)	(8.25)
Categorical		
1st year teaching	0.06	0.07
2nd	0.06	0.07
3rd	0.06	0.07
4th	0.05	0.06
5th	0.05	0.06
6th	0.05	0.05
7th	0.04	0.05
8th	0.04	0.04
9th	0.04	0.04
10th or more	0.55	0.48

Note: Panel A: National Assessment of Educational Progress (NAEP) scores are the simple mean of NAEP tests which occurred during the years in our analysis sample. Descriptive statistics for students are form the from National Center for Education Statistics' Common Core of Data. The exception is the "in poverty" statistic which comes from US Census Bureau Small Area Income and Poverty Estimates. Panel B: Authors calculations using administrative data.

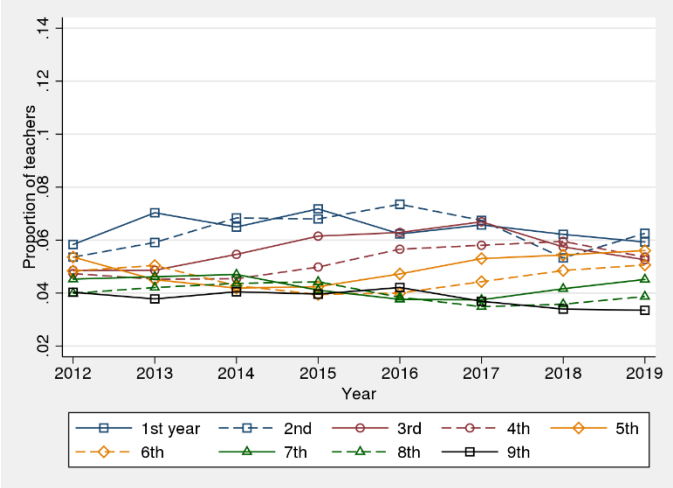
Appendix A. Additional figures and tables



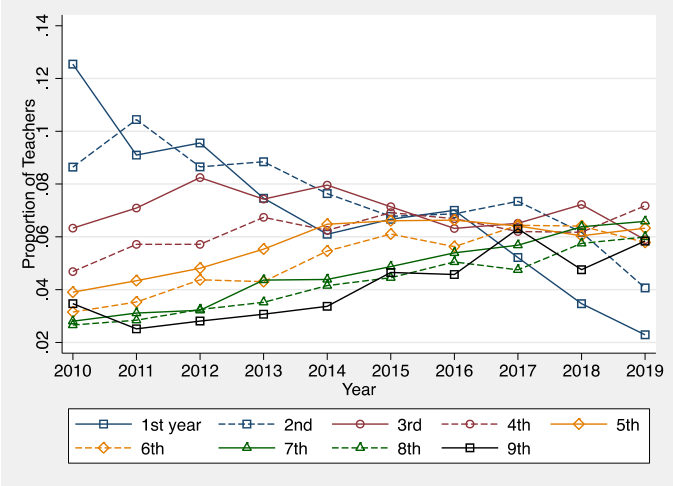
Appendix Figure A1—Distribution of observation scores

Note: Histograms of teacher-by-year observations. The x-axis is a teacher's annual observation score, which is an average of scores for different items or tasks, in the original rubric-scale units. Data are from the Tennessee TEAM rubric 2011-12 through 2018-19, and DCPS TLF rubric 2009-10 through 2015-16. The sample size for Tennessee in Panel A is 375,072 teacher-by-year observations; and similarly for DCPS 35,672 in Panel B, 34,898 in Panel C, and 21,086 in Panel D.

Panel A. Tennessee



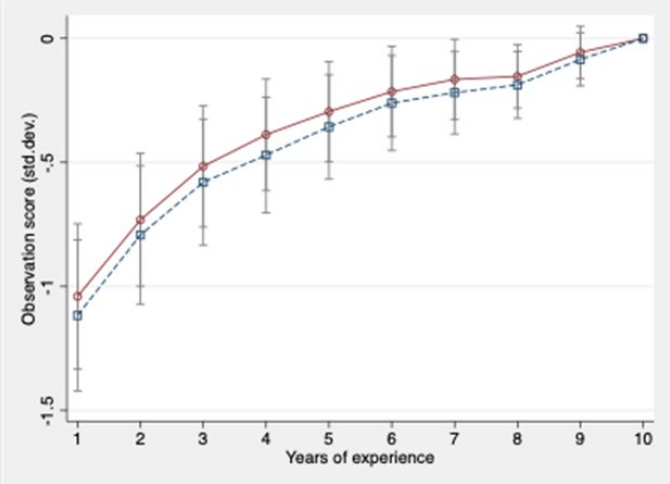
Panel B. DCPS



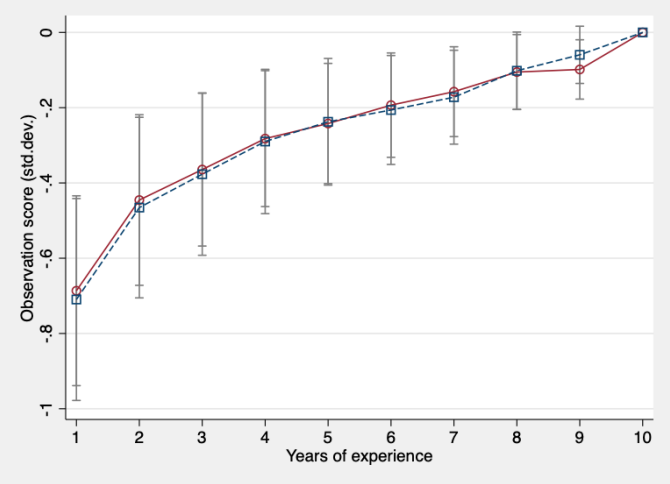
Appendix Figure A2—Distribution of teacher experience over time

Note: Each line measures the proportion of teachers (y-axis) in a given school year (x-axis) who are in their *e*th year of teaching. The estimation sample is the same as Figure 1. The estimation sample for Tennessee includes 375,072 teacher-by-year observations for 81,847 unique teachers, and similarly for DCPS 35,672 and 9,455.

Panel A. School years 2013-14 to 2015-16, TLF rubric

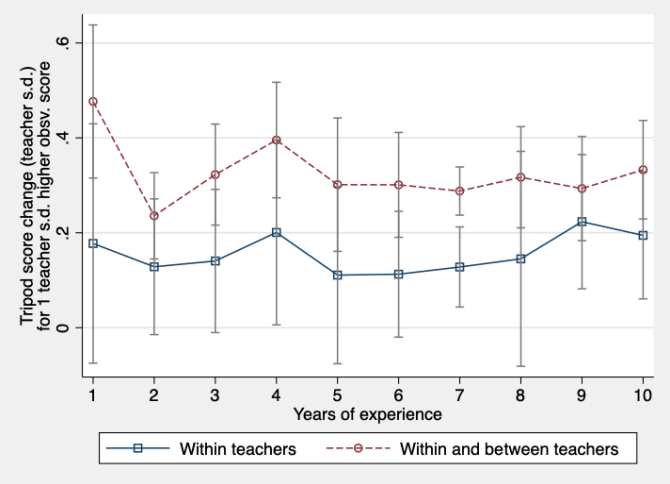


Panel B. School years 2016-17 to 2018-19, EP rubric



Appendix Figure A3—Estimates when the distribution of experience is relatively stable (DCPS)

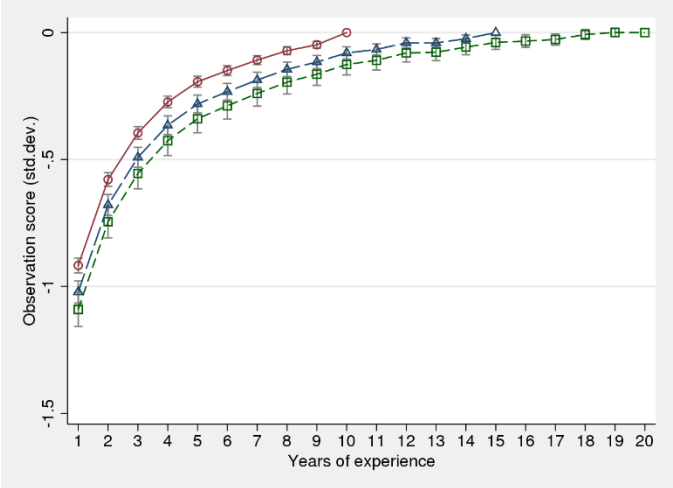
Note: The solid line reports estimates using the two-way fixed effects approach described in Section 1.2. The dashed line reports estimates using the alternative diff-in-diff strategy described in Section 2.3. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The details of estimation are identical to Figure 1 except that the estimation samples here are each a subset of Figure 1's estimation sample. Panel A uses only data from 2013-14 to 2015-16, and panel B only 2016-17 to 2018-19. Starting in 2016-17 DCPS switched from the TLF rubric to the new EP rubric. The sample size for the solid line in panel A is 24,125 teacher-by-year observations for 7,726 unique teachers; and similarly 21,558 and 5,452 for dashed line in panel A, 11,547 and 5,083 for solid line in panel B, and 10,116 and 3,689 for dashed line panel B.



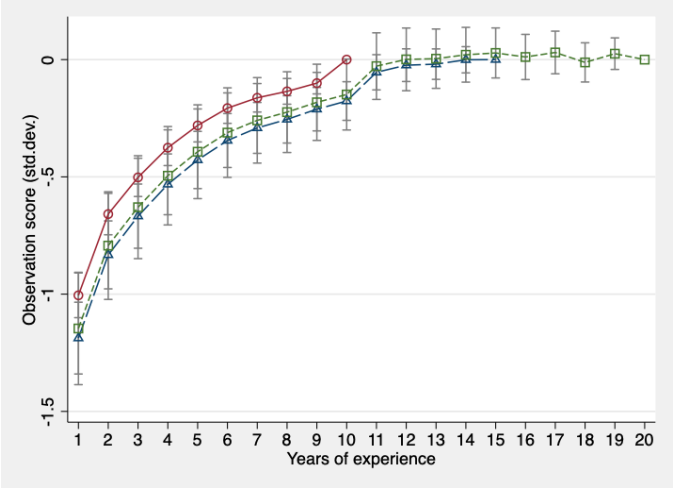
Appendix Figure A4—Predicting student survey scores with teacher observation scores by years of teacher experience (DCPS)

Note: The solid and dashed lines each report estimates from a separate linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). In both cases the outcome variable is teacher j's Student Surveys of Practice (SSoP) score for school year t. SSoP scores are standardized (mean 0, s.d. 1) by school year using the distribution for all teachers in DCPS. In both cases the specification includes (a) indicators for years of experience 1 through 9 individually, with ≥ 10 years the omitted category; (b) classroom observation score, $\bar{s}_{jt}$; and (c) the interactions of (a) and (b). Each plotted point is sum of the coefficient on the (a)*(b) interaction for e years of experience (x-axis) plus the main-effect coefficient on (b). The solid line specification includes year and teacher fixed effects. The dashed line includes only year fixed effects, omitting the teacher fixed effects. The sample size for both lines is 5,362 teacher-by-year observations for 2,643 unique teachers.

Panel A. Tennessee



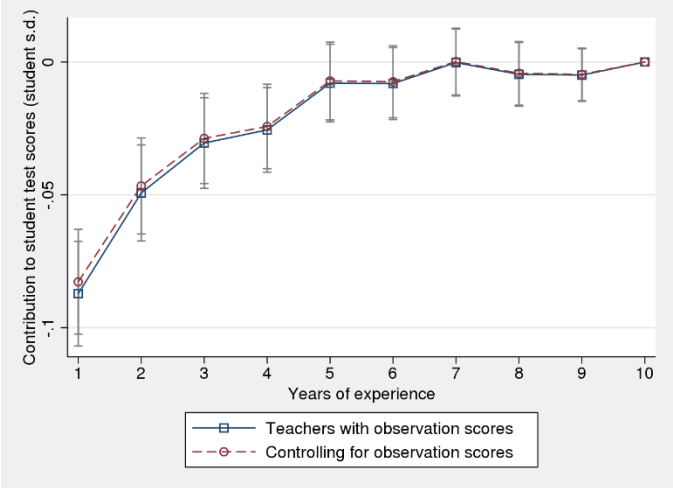
Panel B. DCPS



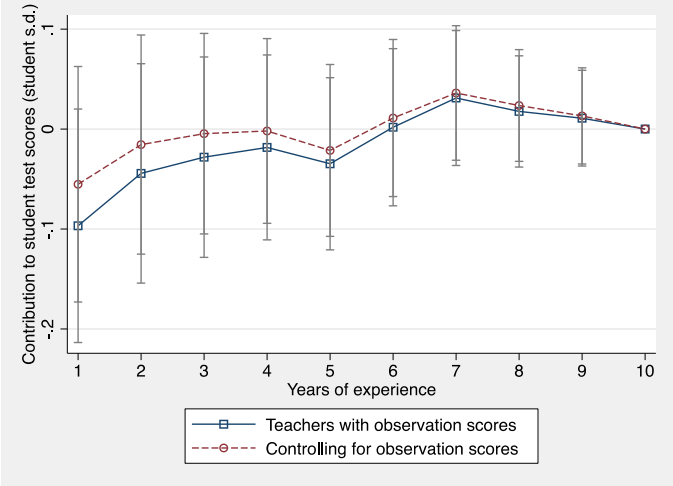
Appendix Figure A5—Estimates by definition of comparison group

Note: Each of the three lines reports estimates from a separate linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The solid line is identical to the solid line in Figure 1. For the two dashed lines, the details of estimation are identical to the solid with one exception. For the solid line, the comparison group is teachers with ≥ 10 years of experience, $\bar{e} = 10$. The two dashed lines show $\bar{e} = 15$ and $\bar{e} = 20$ respectively. The sample size the same for all three lines; in Tennessee 375,072 teacher-by-year observations for 81,847 unique teachers, and similarly in DCPS 33,484 and 7,267.

Panel A. Tennessee

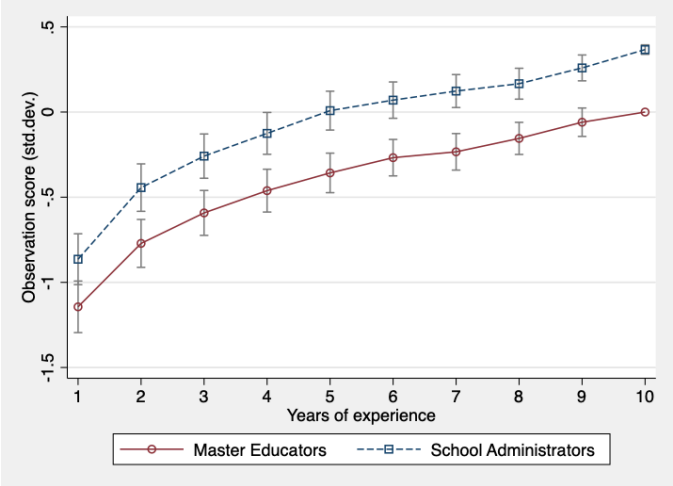


Panel B. DCPS



Appendix Figure A6—Returns to experience for contributions to student achievement controlling for classroom observation score

Note: The solid and dashed lines each report estimates from a separate linear regression, but with the same estimation sample. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). For the solid line "teachers with observation scores" estimates, the details of estimation identical to solid line in Figure 2 (the two-way fixed effects estimates) except that the estimation sample here is a subset of the Figure 2 sample. For the solid line here, the estimation sample is limited to teacher-by-year cases where we have an observation score. For the dashed line "controlling for observation scores" estimates, we add a control for observation score. The sample size the same for the two lines; in Tennessee 3,076,946 student-by-subject-by-year observations and 65,750 teacher-by-year observations for 25,017 unique teachers, and similarly in DCPS 244,696, 5,350 and 2,249.



Appendix Figure A7—Estimates by rater type (DCPS)

Note: Estimates are from a single linear regression. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). The details of estimation are identical to the solid line in Figure 1 with the following exceptions. First, the estimation sample is limited to the TLF years in DCPS from 2010-2016. Second, the experience indicators are interacted with an indicator for rater type: master educator or administrator. The omitted category is master educator and ≥ 10 years of experience. The sample size for the solid line is 18,715 teacher-by-year observations for 5,118 unique teachers; and similarly 21,080 and 5,380 for dashed line.

Appendix Table A1—Predicting student test scores with teacher observation scores

	(1)	(2)	(3)	(4)
(A) Tennessee				
Observation score (st.dev.)	0.166	0.081	0.009	0.005
	(0.003)	(0.001)	(0.002)	(0.002)
(B) DCPS				
Observation score (st.dev.)	0.196	0.098	0.029	0.025
	(0.012)	(0.006)	(0.007)	(0.008)
Student prior test score controls		√	√	√
Teacher experience controls				√
Teacher fixed effects			√	√

Note: Each column within panels reports results of a separate least-squares regression. Standard errors in parentheses are corrected for clustering (teacher). The dependent variable is student i's test score, A_{ijst}, in subject s (maths or English language arts pooled) and school year t. Test scores are standardized (mean 0, s.d. 1) within each grade-by-subject-by-year cell using the distribution for all students in the jurisdiction, Tennessee or DCPS respectively. The key independent variable is teacher j's classroom observation score, $\bar{s}_{jt}$, which is an average of several item-level scores recorded during a given school year t. Observation scores are standardized (mean 0, st.dev. 1) by school year using the distribution of all teachers in the jurisdiction, Tennessee or DCPS respectively. The "student prior test score controls" are a quadratic in prior-year test score, where the parameters are allowed to differ across grade-by-subject-by-year cells, $b\left(A_{is(t-1)}\right)$. The "teacher experience controls" are a set of indicators for years of experience 1 through 9 individually, with ≥ 10 years the omitted category. The sample size the same across columns; in Tennessee 4,222,939 student-by-subject-by-year observations and 92,403 teacher-by-year observations for 34,395 unique teachers, and similarly in DCPS 252,400, 5,429, and 2,274.