Measuring returns to experience using supervisor ratings of observed performance: The case of classroom teachers†

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

We study the returns to experience in teaching, estimated using supervisor ratings from classroom observations. We describe the assumptions required to interpret changes in observation ratings over time as the causal effect of experience on performance. We compare two difference-in-differences strategies: the two-way fixed effects estimator common in the literature, and an alternative which avoids potential bias arising from effect heterogeneity. 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. The authors have no relevant conflict of interest.

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.

Supervisor ratings of observed performance—a ubiquitous job performance measure—present a particular challenge when measuring returns to experience. For example, the relative subjectivity of supervisor ratings creates scope for leniency bias (Prendergast 1999), and supervisors' leniency bias may itself depend on the employee's years of experience. Subjective supervisor ratings are quite common in public-sector jobs where organizational objectives are diffuse and often difficult to quantify. We examine the case of classroom teachers, and the most common performance measure for public-school teachers: ratings by the school principal based on classroom observations.

Teacher personnel management is central to education policy. Teacher salaries dwarf other public-school expenses, consuming three out of every five dollars, and teachers' contributions to student academic and social development dwarf other contributions by schools.

Understanding the causal effects of experience on teaching—as measured by observation ratings—can improve teacher policy in two ways. First, as an input to developing new policy options. Over the last two decades, scholars have produced an expansive literature on how to measure teacher performance at a given point in time. But there remains comparatively little evidence on how teaching skills improve. Consequently, policymakers have struggled to develop policies or practices which consistently yield improvements in teaching effectiveness.[1] Second, as an input to implementing existing policies. Consider, for example, school systems where tenure and salary decisions are now linked to observation ratings. The optimal rules for granting tenure or setting salaries depend critically on expected performance over time, not just at a single point in time.

Our empirical focus in this paper is estimating the returns to experience in teaching using classroom observation ratings. We define "returns to experience" as the causal effect of one additional year of teaching experience on teacher performance, estimating returns separately for the first year of experience, second year, third year, etc. We define experience broadly to include whatever professional experiences occur over the course of a teacher's first year (or second 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 teachers' value-added contributions to student achievement scores. 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.

We use a difference-in-differences framework to make explicit the causal inference features of the returns-to-experience estimates. Our preferred estimates come from applying a diff-in-diff strategy proposed by de Chaisemartin and D'Haultfœuille (2020, 2022a). Briefly, the first difference is the observed change in a teacher's observation rating, $(\bar{s}_{jt} - \bar{s}_{j,t-1})$, from year $(t-1)$ to t when her experience changes from $(e-1)$ to e. The second difference is between early-

[1] For reviews of the related research literature see Goe, Bell, and Little (2008), Kane, Kerr, and Pianta (2014), Jackson, Rockoff, and Staiger (2014), Garrett, Citkowicz, and Williams (2019), James and Wyckoff (2020), and Taylor (2023).

career (treated) teachers and veteran (comparison) teachers. These estimates are the solid lines in Figure 1 using data from Tennessee and the Washington, DC Public Schools (DCPS). The identifying assumptions require: First, that, on average, 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 ratings does not depend on a teacher's years of experience.

We find that teacher job performance—as measured by classroom observation ratings—improves by roughly one standard deviation over the first ten years of a teaching career, as shown in Figure 1. One-third of the growth comes in just the first year. The improvements measured by observation ratings are somewhat larger in magnitude compared to improvements measured by teachers' contributions to student achievement scores. In our sample, value-added improves by 0.10 student standard deviations over the first ten years, as shown in Figure 2. One teacher standard deviation in value-added is between 0.10-0.20 student standard deviations. Still, this pair of estimates of the returns to experience—one standard deviation in ratings and 0.10 in value-added—is consistent with prior cross-sectional estimates of the ratings-to-value-added relationship (Kane et al. 2011, 2013, Araujo et al. 2016, Burgess et al. 2023).

The paper goes on to evaluate several threats to the two identifying assumptions. Most threats are reasons why observation ratings might rise (or 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 DCPS in 2017. As we discuss in detail, changes to the rubric (or to rater training, or to rater-to-teacher matching rules) do not necessarily threaten causal inferences about returns to experience estimates. 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 our estimates.

Our preferred estimation method is new to the literature on teacher returns to experience. Thus, we compare our estimates to estimates which use the conventional strategy. That strategy is also a difference-in-differences strategy using a two-way fixed effects estimator, and both strategies require the same two main identifying assumptions. However, the conventional two-way FE strategy requires additional assumptions.

This is the first paper, to our knowledge, that studies the causal returns to experience reflected in supervisor ratings of observed performance. That contribution to the literature comes from combining explicit causal inference reasoning with panel data on performance ratings. Many prior papers have contributed causal estimates of the returns to experience using other measures of performance, for example, wages (e.g., Angrist 1990, Altonji and Williams 1992, Grogger 2009) or outputs like teacher value-added to student achievement scores (e.g., Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Ost 2014).

Our focus is teachers and there is already a large literature on the returns to experience in teaching (see Taylor 2023 for a review). Still, we contribute to that literature in three ways. First, our paper provides a thorough discussion of causal inference considerations—identification strategies, assumptions, and threats—and a new estimation strategy. Claims of having estimated "returns to experience" are causal claims, but the contemporary tools of causal inference are mostly implicit (or entirely absent) in prior papers. We make explicit the difference-in-differences nature of returns to experience estimates, which clarifies identifying assumptions and threats. We further highlight how the typical estimator is a two-way FE estimator, which will be biased if the returns to experience change over time or the distribution of experience changes over time. Finally, we demonstrate a new estimation strategy for the returns to experience in teaching. That new strategy incorporates recent developments in difference-in-differences methods (for reviews see de Chaisemartin and D'Haultfœuille 2022b, Roth et al. 2022), and avoids the potential bias of the common two-way-FE strategy.

Second, we contribute estimates of the returns to experience using a novel teacher performance measure: classroom observation ratings. Existing estimates of the returns to experience in teaching nearly all use value-added measures of performance. Our estimates of how observation ratings (inputs) change with experience complement estimates of how value-added scores (outputs) change, in part by contributing to efforts to understand the mechanisms behind teachers' improvements in value added (Kraft and Papay 2014, Ost 2014, Atteberry, Loeb, and Wyckoff 2015). Two other papers also estimate returns to experience using observation ratings: Kraft, Papay, and Chi (2020) and work concurrent to ours by Laski and Papay (2020). Those two papers focus on understanding how the returns to experience vary across teachers, schools, etc. By contrast, our paper focuses on whether (or under what circumstances) the estimates should be interpreted as the causal effects of experience.

Third, we explain and examine additional identifying assumptions and threats specific to the observation ratings case. These additional considerations are not relevant to the test-score value-added case. Our examination includes several novel empirical tests relevant to identification threats; tests which can be repeated in other settings.[2]

A final, broader contribution of our paper is to current policy debates involving teacher classroom observations ratings. First, most of our empirical

[2] Some of the estimate in Kraft, Papay, and Chi (2020) and Laski and Papay (2020) are also relevant to identification threats, though neither paper discusses causal inference explicitly.

results are framed by causal inference threats. However, those same results are also relevant to broader concerns about measurement in classroom observations, and in some cases our tests provide new evidence. For example, rater leniency bias (Kraft and Gilmour 2017, Steinberg and Kraft 2017), the influence of the students in the classroom (Campbell and Ronfeldt 2018), unintended effects of teacher-rater pairings (Chi 2021), among other concerns (Cohen and Goldhaber 2016, Grissom and Bartanen 2019). Second, effective policy decisions depend on establishing a causal relationship between a (proposed) policy and a valued outcome. This paper provides an important foundation for policymakers and policy researchers who employ observed performance ratings to better understand teacher development policy.

1. Data and setting

Both DCPS and Tennessee maintain panel data on teachers, including ratings from classroom observations over several years. In DCPS, the panel begins with the start of its current evaluation system, IMPACT, in 2009-10, and we use data through 2018-19. Tennessee's current evaluation system began in 2011-12, and our data run from that start date through 2018-19. In both cases the data include item-level ratings for several specific teaching tasks evaluated in a given observation visit. Teachers in tested grades and subjects can be linked to their students and achievement scores. Characteristics of the teachers and their students in our data are summarized in Table 1.

1.1 Features common to both settings

The DCPS and Tennessee settings share many features. In both locations, all teachers, regardless of experience level, are evaluated every school year by trained observers. The resulting observation ratings are a highly-weighted component, among a larger set of evaluation measures including value-added scores which measure teacher contributions to student achievement.[3] The 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 observation rubrics, both rubrics assess similar tasks and teaching practices, and both rubrics have roots in Danielson's Framework for Teaching (1997). Tennessee uses the TEAM (Tennessee Educator Acceleration Model) evaluation rubric.[4] The TEAM rubric's 19 items are divided into three categories of skills: instruction, planning, and environment. Each 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). The TLF rubric has a 1-4 rating scale (4 = highly effective, 1 = ineffective) for items measuring nine teaching tasks.[5] In 2017, DCPS transitioned to the Essential Practices (EP) observation rubric, which covers similar skills to the TLF, but with more concise

[3] 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.

[4] Not all Tennessee districts use the TEAM rubric, but our analysis in this paper uses only data from the TEAM rubric.

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

definitions for each related task and explicit alignment to the Common Core State Standards.

One frequent, but potentially 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 et al. 2009). This criticism, and most policy discussions, focus on the final end-of-year "summative" scores which end up in a teacher's personnel file. These final scores lack variation in part because final scores are rounded off to integer values. In this paper we use observation scores that average across many item ratings (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).

1.2 Differences between the two settings

While both evaluation systems share many features, there are a number of useful differences. First, both places use trained school administrators as raters (e.g., principals and assistant principals, or other instructional leaders). However, until a change in 2017, in DCPS teachers were also observed and rated by "master educators"—specialized observers external to the school with subject- and grade-specific expertise. Two of each 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 low performance (a final annual score below effective) are subject to involuntary dismissal. Prior work documents that these incentives influences teachers' behavior at work and their decision about remaining at DCPS (Dee and Wyckoff 2015, Dee, James, and Wyckoff 2021). There are also rewards in DCPS for high performance. Teachers who demonstrate exceptional performance (a final annual score of 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.

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.

1.3 Additional data

In addition to classroom observation ratings, 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 demographic characteristics. We have information regarding the observation raters and timing of the observation visits. In both settings we have the usual information regarding each teacher’s students, for tested subjects and grades, including eligibility for free or reduced-price lunch, race and ethnicity, and standardized achievement scores.

Additionally, DCPS began using student surveys in 2016-17 as teacher performance measures. This measure is adapted from the Tripod survey (Ferguson and Danielson 2015), which ask students’ questions about their teachers’ practice.

An example question: "When explaining new ideas or skills in class, my teacher tells us about common mistakes that students might make."

2. Returns to experience estimates

2.1 Estimation methods

We estimate the "returns to experience"—the improvement in performance caused by additional experience—using a difference-in-differences strategy. Our measure of performance, $\bar{s}_{jt}$, is the classroom observation score for teacher j in school year t. At the start of year t, teacher j has $expr_{jt}$ years of prior teaching experience. Using these inputs, we apply the diff-in-diff estimator proposed by de Chaisemartin and D'Haultfœuille (2020, 2022a). Later in Section 2.4 we compare this strategy to the more-common estimation approach in the literature.

Let δ_e be the improvement in performance caused by gaining the eth year of teaching experience. Our estimate of δ_e is:

$$\hat{\delta}_e = \frac{\sum_t N_{et}\,\hat{\delta}_{et}}{\sum_t N_{et}}$$

$$\hat{\delta}_{et} = \left[\frac{1}{N_{et}} \sum_{\substack{j:expr_{j,t}=e,\\ expr_{j,t-1}=e-1}} \left(\bar{s}_{j,t} - \bar{s}_{j,t-1}\right)\right] - \left[\frac{1}{M_{et}} \sum_{j:expr_{j,t-1}\geq \bar{e}} \left(\bar{s}_{j,t} - \bar{s}_{j,t-1}\right)\right] \qquad (1)$$

where δ_{et} is simply the δ_e effect for a specific school year t. The number of treated teachers is N_{et} and comparison teachers is M_{et}. Because a teacher, j, may contribute to several $\hat{\delta}_e$, our standard error estimates correct for clustering at the teacher level.[6]

[6] In practice, to facilitate standard error estimation, we estimate the many individual $\hat{\delta}_{et}$ terms simultaneously in one system of regressions, stacking together one regression for each $\hat{\delta}_{et}$. The stack includes one regression (synonymously, one layer) for each unique combination of e and t, where $e \in \{1,2,\ldots,\bar{e}\}$ and t is a school year in our data. Each regression in the stack has the same

The treatment is gaining the eth year of experience. Thus, teacher j is in the treated sample if she had e years of experience at the start of school year t, but only $(e-1)$ years of experience at the start of school year $(t-1)$. Inside the brackets on the left is the average first-difference in observation score, $\bar{s}_{jt}$, for the sample of treated teachers. That first-difference is the observed change in a teacher j's score between year $(t-1)$ to t when her experience changes from $(e-1)$ to e. Still, a given teacher's scores may change over time for reasons unrelated to her own experience, which motivates the second difference between treated and comparison teachers.

Our comparison sample is veteran teachers—teachers who have at least $\bar{e}$ years of experience. One identifying assumption, which we formalize below, is that past $\bar{e}$ years of teaching experience, there are no longer any returns to experience for the average teacher. Inside the brackets on the right is the average first-difference for the veteran comparison teachers. Any observed change from $(t-1)$ to t among veterans is, by assumption, unrelated to experience and differenced out. Our main estimates set $\bar{e}=9$, but our estimates are robust to higher values of $\bar{e}$ as we show later.

Each $\hat{\delta}_{et}$ estimate uses data from just two school years: one treated year, t, and one pre year, $(t-1)$. A teacher's performance in year t is affected by her eth year of experience. A teacher's performance in year $(t+1)$ is also affected by her eth year of experience, but also affected by her $(e+1)$th year of experience. Thus, we observe the marginal effect of the eth year of experience for one school year, t, after which the eth year is confounded with further experience gains.

simple specification, $\Delta\bar{s}_{jt}=\alpha+\delta T_{jt}+\epsilon_{jt}$, but the estimation sample is limited to only teacher-by-year observations, jt, where either (i) $expr_{j,t}=e$ and $expr_{j,t-1}=e-1$, or (ii) $expr_{j,t-1}\geq\bar{e}$. The indicator variable $D_{jt}=1$ for group (i) and $=0$ for group (ii). When stacked together, the specification becomes: $\Delta\bar{s}_{jt}=\alpha_{et}+\delta_{et}T_{jt}+\epsilon_{jt}$. To obtain $\hat{\delta}_e$ we take the weighted average of $\hat{\delta}_{et}$ terms as shown in Equation 1. This stacked approach allows us to correct for clusters (teachers) across regressions.

The outcome variable, $\bar{s}_{jt}$, is teacher j's classroom observation score for school year t. More precisely, $\bar{s}_{jt}$ is the average of several task-specific scores, $\bar{s}_{jt} = \frac{1}{K}\sum_{k=1}^{K} s_{kjt}$. 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). The $\bar{s}_{jt}$ scores are scaled in teacher standard deviation units, within jurisdiction (Tennessee or DC) by year cells.[7]

2.2 Main results

Teacher performance measured in classroom observations improves with experience. In Figure 1 the solid line plots our returns to experience estimates from the difference-in-differences strategy in Equation 1. Observation scores are scaled in standard deviation units, and, by construction, the zero line on the y-axis is the average score among veteran teachers, $expr_{jt} \geq \bar{e} = 9$. The vertical lines mark cluster-corrected 95 percent confidence intervals.

Just one year of teaching experience improves performance by one-quarter to one-third of a standard deviation. Over the first ten years of a teaching career, performance in observations improves one standard deviation. The patterns in DC

[7] We begin with the item-by-observation-visit data recorded by observers in the original rubric units (integer scores 1-4 in DCPS and 1-5 in Tennessee). Separately for DCPS and Tennessee: (i) We standardize the item-by-visit ratings so that, by school year, each item is mean 0, standard deviation 1. (ii) For each teacher j by item by school year, we calculate the school-year average of the standardized item-by-visit ratings. We then re-standardize the item-average scores. (iii) For each teacher j by school year, we average her item-average scores to create the overall average score, $\bar{s}_{jt}$. Finally, we again standardize the overall average scores by year.

and Tennessee are quite similar. What are the educational or economic consequences of these gains? Improvements in teaching inputs contribute to outputs, including student learning which we can measure with teachers' value added to student test scores. There is evidence that teachers' test-score value-added contributions translate to better longer-run outcomes, including college going and labor market success (Chetty, Friedman, and Rockoff 2014). However, there are currently no estimates linking observation scores to longer-run student outcomes.

The pattern in Figure 1, using classroom observation ratings, is similar to the pattern of returns to experience for teacher value added to student achievement scores. In Figure 2 the solid line plots estimates where the performance measure is a teacher's value-added contribution to student test scores. We first obtain value-added scores, $\hat{\mu}_{jt}$, then apply the estimator in Equation 1 substituting $\hat{\mu}_{jt}$ for $\bar{s}_{jt}$.[8] In Figure 2, the y-axis, $\hat{\mu}_{jt}$, is measured in student standard deviation units, and the sample is limited to teachers of grades 4-8 in math and English language arts. The pattern for Tennessee in Figure 2 matches estimates from several other places (see Taylor 2023 for a review). The DC estimates are much nosier but consistent with the typical pattern. In both cases teacher value-added improves by about 0.10 student test-score standard deviations over the first ten years of teaching.[9]

The improvements measured by observation ratings are somewhat larger in magnitude compared to improvements measured by value added. One teacher standard deviation in value-added is between 0.10-0.20 student test-score standard deviations. There are many potential explanations for the difference, given the variety of teacher skills and tasks not captured by observation scores or by test-score value-added. Still, there are several existing cross-sectional estimates of the relationship between observation ratings and value-added (Kane et al. 2011, 2013,

[8] Appendix B provides details of our value-added estimation methods.

[9] Additionally, Appendix Figure A2 reports results using student survey measures of teacher performance, and again the pattern of returns to experience is quite similar.

Araujo et al. 2016, Burgess et al. 2023). In those estimates a one standard deviation increase in observation ratings predicts a 0.05-0.11 increase in value-added. Our pair of estimates of the returns to experience, Figures 1 and 2, are quite consistent with those prior cross-sectional estimates.

2.3 Causal inference

The difference-in-differences setup provides a familiar framework for evaluating causal claims about the estimates in Figure 1. Stated in general terms, the identifying assumption in this case is: Any change over time we observe in veteran (comparison) teachers' scores is the same change we would see in early-career (treated) teachers' scores if there were no returns to experience. We can clarify the identifying assumption further with the help of a simple conceptual framework.

2.3.1 Observation scores and true performance

A teacher's job involves many tasks—learning content, planning lessons, asking questions in class, responding to misbehavior, grading, communicating with parents, any many more. Each of those tasks produces some input to the production of student achievement or other goals of schooling. Let θ_k measure true performance of task k. Higher performance is synonymous with producing more or higher-quality task k inputs.

Classroom observation rubrics are designed to measure task performance, θ_k, at least for some subset of a teacher's tasks. Rubrics are not designed to measure outcomes like student achievement. 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 sometimes described as measures of a teacher's skills. But an observation score is a function of both skills and effort, thus we prefer describing those scores as measures of performance.

Still, classroom observations are an imperfect way to measure performance. An observation score, s_k, is inevitably some combination of true performance, θ_k, and other factors unrelated to performance, ν_k. For exposition we assume:

$$s_k = g(\theta_k, \nu_k) = \theta_k + \nu_k \tag{2}$$

Those other factors, ν_k, include much more than just classical measurement error. Even as the number of observations grows, features of the evaluation process will create some difference between $E[s_k]$ and $E[\theta_k]$. First, ν_k includes explicit features of the evaluation process, for example, the rubric itself, how evaluators are trained, how evaluators are assigned to teachers, incentives attached to scores. Such explicit features are (mostly) controllable by those designing and implementing the evaluation. But ν_k also includes less-explicit less-controllable features, for example, the behaviors teachers or evaluators choose in response to the explicit features.

2.3.2 Identifying assumptions

Interpreting Figure 1 as the returns to experience—the causal effect of teaching experience on true task performance—requires two identifying assumptions. *Assumption 1:* Factors which contribute to observation scores but are unrelated to performance, ν_k in Equation 2, do not depend on teaching experience. Specifically, $E\left[\nu_{kjt} | k, t, expr_{jt}\right] = E\left[\nu_{kjt} | k, t\right]$. This assumption requires that if an early-career and a veteran teacher both have the same true task performance, θ_k, they will have the same observation score, s_k. *Assumption 2:* True performance is not changing over time, on average, in the comparison group of teachers. Specifically, $E\left[\theta_{kjt} - \theta_{kj(t-1)} | expr_{jt} \geq \bar{e}\right] = 0$.

The importance of a comparison group is shown by stating the assumption that would replace Assumption 2 in the absence of a comparison group. *Assumption 3:* The ν_k factors do not change over time. Specifically, $E\left[\nu_{kjt} | k, t, expr_{jt}\right] = E\left[\nu_{kjt} | k\right]$. If we used only early-career teachers' data, we could not separate the returns to experience from changes in ν_k over time, because $expr_{jt}$ and t are colinear within teacher. In Section 3 we discuss several different substantive threats to these identifying assumptions, but some of the quite-plausible threats are known changes in ν_k over time.

These are the assumptions required for claims about performance of the teaching tasks which classroom observations are designed to measure. We might also be interested in claims about other aspects of teacher performance, like teachers' value-added to student achievement scores. Imagine a production process for student achievement; some of the inputs will be the teaching tasks described in an observation rubric. However, to make any inference from observation scores to value-added would require a much better understanding of that production process than currently exists.[10] Later we provide some new empirical evidence relevant to that broader inference.

2.4 Alternative estimation methods

Our estimation methods, described in Section 2.1, are new to the literature on returns to experience in teaching. Here we compare our estimation strategy to the conventional estimation strategy—the strategy which, to date, has been most common in that literature (see Taylor 2023 for a review).[11] The conventional strategy is also a difference-in-differences strategy, but using a two-way fixed effects estimator, though it is not often described in those terms. Both strategies

[10] The literature does include many estimates of the correlation between observation scores and teacher value-added, which is typically much less than 0.50. In our data that correlation is 0.38 for Tennessee and 0.30 for DCPS. Appendix Table A1 reports on these estimates in detail. However, 0.38 and 0.30 are likely to underestimate the true correlation. First, there is the common attenuation because of measurement error. Second, the simple mean $\bar{s}_{jt}$ gives equal weight to each task k, but it seems unlikely the elasticity of value-added, μ, with respect to θ_k is equal for all k. If we knew the production function for student achievement, we would likely choose un-equal weights.

[11] Though the conventional strategy is common, in nearly all prior papers the performance measure is teachers' value-added contributions to student test scores.

require the same core set of identifying assumptions, but the conventional strategy requires additional assumptions about effect heterogeneity.

In the conventional approach, estimates of the returns to experience come from a least-squares regression. The basic specification is:

$$\bar{s}_{jt} = h(expr_{jt}) + \lambda_j + \pi_t + \varepsilon_{jt} \quad (3)$$

where the outcome is a measure of teacher performance, $\bar{s}_{jt}$ in our case.

Selective attrition is a fundamental threat to any returns-to-experience estimate. Attrition from teaching is likely negatively correlated with performance. In response to that threat nearly all estimation strategies focus on variation within individual teachers over time. Our main strategy uses only within-teacher variation by first differences. The conventional approach uses teacher fixed effects (e.g., Rockoff 2004).

However, for a given teacher, years of experience, $expr_{jt}$, is colinear with school year, t, unless she takes a leave of absence. Specification 3 includes both teacher fixed effects, λ_j, and school year fixed effects, π_t, and thus requires some restriction on $h(expr_{jt})$ to avoid the age-period-cohort problem. The typical restriction is to assume no returns to experience after some number of years, $\bar{e}$. Then $h(expr_{jt})$ is a series of indicator variables for years of experience up to $\bar{e}$:

$$h(expr_{jt}) = \sum_{e=0}^{\bar{e}-1} \beta_e \times \mathbf{1}\{expr_{jt} = e\} \quad (4)$$

$$\text{and } \delta_e = \beta_e - \beta_{e-1}.$$

The omitted category is veterans, $\mathbf{1}\{expr_{jt} \geq \bar{e}\}$.[12] This restriction maps to identifying Assumption 2, as stated earlier. That required assumption is well known

[12] It is more common in the literature to make the first year of teaching, $expr_{jt} = 0$, the omitted category in $h(expr_{jt})$. We prefer to omit veterans in part to make the comparison with our main estimates easier. Nevertheless, the choice of omitted category does not change the estimates of interest obtained from fitting the two-way FE specification in Equations 3 and 4. If we reproduced

in the literature on returns to experience in teaching; the assumption is sometimes stated explicitly (e.g., Rockoff 2004) and sometimes criticized (e.g., Papay and Kraft 2015).[13]

This conventional estimation strategy uses a two-way fixed effects estimator. Notice in Specification 3 the characteristic group and period fixed effects, λ_j and π_t, and a series of treatment indicators, $h(expr_{jt})$. A recent, growing literature clarifies several properties of two-way FE estimators; in particular, how those estimators can produce biased estimates when treatment effects are heterogeneous (for reviews see de Chaisemartin and D'Haultfœuille 2022b, Roth et al. 2022).

Three types of heterogeneity can create bias in two-way FE estimates. The first two types, and the resulting bias, are now regularly discussed in papers using difference-in-differences methods. The third type is specific to settings with multiple treatments, including the returns to experience estimates which produce $\hat{\delta}_e$ for several e.

The first source of potential bias would arise if the effects of the eth year of experience, δ_e, differ across cohorts of teachers. In other words, variation over time in δ_e, given the link between cohort, time, and experience.[14] Why might δ_e change

---

the dashed line in Figure 1 from scratch, but instead with the omitted category $expr_{jt} = 0$, the only thing that would change is the y-intercept value. All of the slopes between points would remain exactly as they are in Figure 1. .

[13] There are alternative specifications of $h(expr_{jt})$ in the literature: (i) Specifying h as cubic in $expr_{jt}$, or other higher-order polynomial, 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+ (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).

[14] This potential bias arises in part through the implicit weights in the two-way FE estimator (de Chaisemartin and D'Haultfœuille 2022b). When estimating $\hat{\delta}_e$, the weight given to $\hat{\delta}_{et}$ is increasing in N_{et}, M_{et}, and $var(\mathbf{1}\{expr_{jt} = e\})$. By contrast, our preferred estimates only weight by N_{et}. In our current setting, N_{et} and M_{et} will likely covary over time. However, $var(\mathbf{1}\{expr_{jt} = e\})$ will be constant across cohorts given the specification of $h(expr_{jt})$.

over time? First, selection into (or out of) teaching may change over time. For example, schools may get better (or worse) over time at selecting hires on potential job performance (e.g., Jacob et al. 2014), or self-selection by (prospective) teachers may change in response to compensation for potential performance inside or outside schools (e.g., Nagler et al. 2020, Leaver et al. 2021). Second, the nature of treatment itself may change over time. For example, if schools devote more resources to mentoring for early-career teachers (e.g., Rockoff 2008, Kraft et al. 2018).

The second source of potential bias would, theoretically, arise if the effects of the eth year of experience increase (or decrease) over time within a cohort of teachers.[15] However, in practice, this second bias is not a concern in the conventional returns to experience estimates. In Specification 3, with $h(expr_{jt})$ in 4, treated teachers (early-career teachers) are not part of the comparison group until they are beyond $\bar{e}$ (until they become veteran teachers).[16]

The third source of potential bias arises when there are multiple treatments, as detailed in de Chaisemartin and D'Haultfœuille (2022a). In the current setting, first, the estimate for the eth year, $\hat{\delta}_e$, can be biased if the effects in other years, $\delta_{(-e)} \in \{\ldots, \delta_{e-2}, \delta_{e-1}, \delta_{e+1}, \delta_{e+2}, \ldots\}$, differ across cohorts. Second, $\hat{\delta}_e$ can be biased if the distribution of teacher experience is changing over time; in other words, if the probability of the other treatments is changing over time. Even if all

[15] Conceptually, experience gained in the eth year of teaching can affect performance in year $(e + 1)$, $(e + 2)$, $(e + 3)$, and on into the future. Empirically, however, the effects of the eth year on $(e + 2)$ performance cannot be separately identified from the effects of the $(e + 1)$th year on $(e + 2)$ performance.

[16] The specification of $h(expr_{jt})$ is analogous to the common event study specification, which is not subject to this second source of potential bias. In other words, the conventional strategy only uses treated-teacher data from years e and $(e - 1)$ to estimate the effect of the eth year of experience.

the δ_e are not changing over cohorts, this second bias threat remains.[17] The distribution of experience may be fairly constant for an entire state, like Tennessee, but could change over time for a district, like DCPS, with changes in hiring and retention strategies.

Do these potential biases affect our estimates? The dashed line in Figure 1 shows our estimates from the common two-way FE strategy, alongside our preferred strategy. In Tennessee the two lines are nearly identical, suggesting little change from cohort to cohort in the returns to experience. Estimated growth over ten years is just 2 percent of a standard deviation smaller with the two-way FE approach compared to our preferred estimates. By contrast, there is some difference for DCPS, suggesting the two-way FE strategy underestimates the steepness of returns to experience. Over ten years the accumulated difference is about 17 percent of a standard deviation. Estimated growth in any one year, $(e-1)$ to e, differs by 2 percent of a standard deviation on average. Up through the fifth or sixth year, the year-to-year estimated changes are nearly identical.

The differences in DCPS estimates are largely explained by changes in the distribution of teacher experience in DCPS over time. Appendix Figure A3 shows that the distribution of experience shifted away from early-career teachers over time but became more stable from 2014-15 on.[18] If we restrict out analysis to this more-stable more-recent period, the standard and alternative approaches are quite similar,

---

[17] Briefly, first, when estimating δ_{et} for a given e, two-way FE uses observations from both veterans and early-career teachers (as long as $expr_{jt} \neq \{e, e-1\}$) to estimate the year effects, π_t. But the estimation of π_t must account for the fact that there are other treatments occurring—specifically, the early-career teachers are also gaining $\delta_{(-e)t}$. The two-way FE estimator uses the pooled estimate $\hat{\delta}_{(-e)}$ instead of the t-specific estimate $\hat{\delta}_{(-e)t}$. Second, accounting for other treatments also requires knowing the probability of those other treatments in the sample at time t. In the current setting, the probability of other treatments is the proportion of teachers at each $e \in \{0,1,\dots,\bar{e}\}$—the distribution of teacher experience. The two-way FE estimator uses the pooled estimate of the probability instead of the t-specific estimate.

[18] Our DCPS data begin in 2009-10 and thus the early years coincide with the slow labor recovery following the recession. We do not see the same pattern in Tennessee where the experience distribution has been stable over the years we study.

as shown in Appendix Figure A4. The observable changes in the distribution of experience in DCPS may also be correlated with changes in the returns to experience among DCPS teachers.

The two strategies also yield similar estimates when the performance measure is a teacher's value-added contribution to student test scores, as shown in Figure 2. The value-added returns-to-experience estimates are much noisier, given the much smaller samples. For Tennessee, estimated growth over ten years is one-quarter of a standard deviation smaller with the two-way FE approach compared to our preferred estimates. Estimates are more similar for DCPS. However, in both settings, we cannot reject the null hypothesis of no difference between the two strategies.[19]

3. Alternative explanations and threats to causal inference

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

3.1 General evidence

Before taking up specific alternative explanations, we begin with some general evidence relevant to the plausibility of identifying Assumptions 1 and 2. First, consider Assumption 1 which requires: factors which contribute to observation scores but are unrelated to performance, ν_{jt}, do not depend on teacher experience. We cannot test this assumption directly. However, if Assumption 1 is

[19] Appendix B provides details of our value-added estimation methods.

true, we would predict that $cov(\nu_{jt}, \hat{\mu}_{jt})$ should not depend on experience, and thus $cov(\bar{s}_{jt}, \hat{\mu}_{jt})$ should also not depend on experience.[20] Here $\hat{\mu}_{jt}$ is the teacher's value-added contribution to student achievement test scores. We can test this prediction which is relevant to judging Assumption 1.

Why might $cov(\nu_{jt}, \hat{\mu}_{jt})$ depend on experience? Raters may give greater scrutiny to early-career teachers, perhaps with the specific goal of increasing the correlation between observation scores and value added. Alternatively, raters may be more lenient with early-career teachers, reducing $cov(\nu_{jt}, \hat{\mu}_{jt})$. We discuss other possibilities in Section 3.2, 3.3, and beyond. However, Assumption 1 could be violated without affecting $cov(\nu_{jt}, \hat{\mu}_{jt})$. For example, raters might simply add 1 point to all scores for early-career teachers, which, barring any ceiling effects, would not affect the correlation between observation scores and value added conditional on experience.

Figure 3 provides information on the $cov(\bar{s}_{jt}, \hat{\mu}_{jt})$, whether that relationship depends on teacher experience, and thus a test of the prediction outlined above. The x-axis is years of prior experience. The y-axis is the predicted increase in value added, $\hat{\mu}_{jt}$, if a teacher's observation score, $\bar{s}_{jt}$, increases by one standard deviation.[21] As shown in Figure 3, the estimated relationship between

[20] Using Equation 2 we can write:

$$cov(\bar{s}_{jt}, \hat{\mu}_{jt}|expr_{jt}) = cov(\theta_{jt}, \hat{\mu}_{jt}|expr_{jt}) + cov(\nu_{jt}, \hat{\mu}_{jt}|expr_{jt}).$$

Assume quite plausibly that $cov(\theta_{jt}, \hat{\mu}_{jt}|expr_{jt}) = cov(\theta_{jt}, \hat{\mu}_{jt})$. In plain language, this assumption requires that the education production process—which turns teaching input tasks, θ_{jt}, into a teacher's value-added contributions, $\hat{\mu}_{jt}$—does not depend on experience. Under this assumption, the prediction that $cov(\nu_{jt}, \hat{\mu}_{jt}|expr_{jt}) = cov(\nu_{jt}, \hat{\mu}_{jt})$ implies the further prediction that $cov(\bar{s}_{jt}, \hat{\mu}_{jt}|expr_{jt}) = cov(\bar{s}_{jt}, \hat{\mu}_{jt})$.

[21] The estimation details for Figure 3 are summarized in its note and described in Appendix B. The solid line uses only within-teacher over-time variation (by including teacher FE in the estimation), and the dashed line uses both within- and between-teacher variation (by omitting teacher FE). To get a sense of the correlation between observation scores and value added, multiply the y-axis by about five for Tennessee and three for DCPS. Additionally, Appendix Table A1 provides

observation scores and value added is largely unrelated to experience. There is no clear trend over experience, and we cannot reject the null hypothesis that each point estimate is equal to the average of the series it belongs to, though the DCPS estimates are quite noisy. The one exception is the earliest years in Tennessee using only within-teacher variation (solid line series). Those estimates suggest the $cov(\bar{s}_{jt}, \hat{\mu}_{jt})$ may be higher in a teacher's first year of employment. Some of the specific threats described below could be a mechanism behind the first-year correlation.[22] In summary, the lack of a relationship between $cov(\bar{s}_{jt}, \hat{\mu}_{jt})$ and experience in Figure 3 is consistent with Assumption 1, through the prediction outlined above.

We can also partially test identifying Assumption 2. That assumption requires that, on average, true performance, θ_k, is not changing over time among the comparison group of veteran teachers, i.e., $E[\theta_{kjt} - \theta_{kj(t-1)} | expr_{jt} \geq \bar{e}] = 0$. Our main estimates in Figure 1 set $\bar{e} = 9$ to define the veteran group. If Assumption 2 holds, then our estimates for returns at $e =$ 0-8 should be robust to setting $\bar{e}$ above 9.

Our returns to experience estimates are quite robust to changes in $\bar{e}$. The solid line in Figure 4 simply repeats the solid line in Figure 1 for convenient comparison, with $\bar{e} = 9$. The two dashed lines show estimates where $\bar{e} = 14$ and $\bar{e} = 19$. The three lines have different intercepts; the intercept in this case is the average performance among veteran teachers with $expr_{jt} \geq \bar{e}$. Still, the slopes of

complementary evidence. That table reports the average relationship between observation scores and value added, but that average relationship does not change when we control for experience.

[22] An alternative explanation is the following: We assumed that the production process which turns teaching tasks, θ_{jk}, into value added output, $\hat{\mu}_{jt}$, does not depend on experience. Even if the production process is constant, the way in which a teacher chooses to optimize that production process may depend on experience. For example, perhaps as early-career teachers gain experience they shift more effort to tasks which are not measured by the observation rubric, or more subtly shift effort across tasks in a way not well captured by the simple average of ratings, $\bar{s}_{jt}$.

the lines are quite similar across estimates over the range of 0-9 years of prior experience. For example, in Tennessee, the slope between $(e-1)=0$ and $e=1$ is 0.346 standard deviations in the estimates with $\bar{e}=9$ (solid line) and 0.352 when $\bar{e}=14$ (long-dash line), a difference of 0.006. Indeed, that same difference in slope estimates is 0.006-0.007 for all of the pairwise $(e-1)$ to e slopes. The accumulated difference over the first ten years is 0.05. Those slopes are the returns to experience we want to estimate, and those estimated changes are robust to the choice of $\bar{e}$.[23]

Additionally, while we cannot observe $\Delta\theta = E\left[\theta_{kjt} - \theta_{kj(t-1)} | expr_{jt} \geq \bar{e}\right]$ directly, we can observe $\Delta\bar{s} = E\left[\bar{s}_{kjt} - \bar{s}_{kj(t-1)} | expr_{jt} \geq \bar{e}\right]$. 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 $\Delta\bar{s} \cong 0$ but $\Delta\theta \neq 0$? Only in the knife-edge case where any change in true performance, θ_k, is just offset by a change in the ν_k component of scores.

[23] Figure 4 suggests there may be continued returns to experience beyond a teacher's first decade (beyond $e=9$). For example, in Tennessee, the average observation score at 11 years of prior experience is statistically significantly greater than at 9 years but less than at 14 or 19 years. In DCPS the estimates are much nosier. Such continued returns would, strictly speaking, violate Assumption 1. However, first, the continued returns would generate downward bias in our estimates, leading us to understate the returns to experience. Second, the magnitude of bias is empirically small, for example, 0.346 versus 0.352. The bias is small for two reasons: (i) The gains after the first decade are quite small in relative terms. Teacher scores improve more than 1 full standard deviation in the first decade, but only another 10% of a standard deviation the next five years. (ii) Observations with e = 10-14 are a small share of all observations with $e > 9$.

[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).

3.2 The evaluation system

Changes in observation ratings 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_k, remains constant, the rating assigned to that performance, s_k, 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 ν_k in Equation 2 where $s_k = \theta_k + \nu_k$. (The incentives or consequences attached to performance measures are also a key feature of an evaluation system, and we discuss those incentives below.)

The most straightforward example of a change in ν_k 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 EP rubric did not measure exactly the same set of tasks, k, as the old TLF rubric. Changes in other settings might be smaller, like word choices, even if the tasks scored remain the same. Still, large or small rubric changes would not necessarily threaten our identifying assumptions, as long as the rubric changes affect early-career (treatment) and veteran (comparison) teachers equally.

The DCPS changes allow us to compare estimates from different rubrics. In Figure 5 the short dash line shows estimates of returns to experience using only ratings generated by the TLF, while the long dash blue line uses only EP ratings. 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. Using TLF rubric data suggests the average teacher improves by 1.3 standard deviations over the first ten

[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.

years, compared to 1.2 standard deviations using the EP rubric. Though the differences are not statistically significant. The differences suggest a potential threat to Assumption 1—that ν_k 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 selection 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 A3). Thus, in Figure 5, the higher scores with the EP rubric may reflect true higher performance because of selection.

Choosing raters is also a key evaluation design decision, and a decision which itself may change over time. Figure 5 also compares estimates by rater type for DCPS. The solid red line uses only ratings from the master educator raters, who specialize in rating and 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, and there is no composition concern since each teacher was rated by both a school administrator and master educator each year. The slopes of the two TLF lines are quite similar, especially over the first five years of a teacher's career. Using master educator ratings suggests improvement of over 1.6 standard deviations over the ten years, compared to 1.3 using principal ratings. Though again the differences are not statistically significant. Figure 5 does obscure one important difference between master educator scores and school administrator scores: School administrators give higher average scores on the 1-4 scale; in other words, the ν_k component in Equation 2 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 ν_k does not violate Assumption 1.

In general, changes to the evaluation system are changes to the ν_k component in Equation 2. Interpreting Figure 1 as the causal returns to experience does not require that ν_k remain unchanged over time. The only restriction on ν_k is that ν_k not depend on experience. This applies to obvious changes in ν_k, like the rubric or types of evaluators, and to changes which are more difficult (for the researcher) to observe. Thus, while the tests in Figure 5 address some threats, they do not rule out all threats from changes in the evaluation system. One potentially difficult to observe change is changes 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_k, are not reflecting performance, θ_k, 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 observation scores; for example, 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.

3.3 Behavior of the raters

Changes in ratings over time may reflect changes in the behavior of the raters. Raters have some discretion within any evaluation system's designed procedures. Rubric-based classroom observation ratings fall somewhere in between the theoretical poles of truly objective evaluation and purely subjective evaluation. Moreover, 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 ν_k component in Equation 2.

One behavior that is frequently cited, given rater discretion, is leniency bias—the tendency for raters to give scores which are higher than warranted. Histograms of observation ratings (Appendix Figure A1) are consistent with systematic leniency bias in both Tennessee and DCPS, although such bias is less evident for ratings assigned by the master educators in DCPS. The skew in the ratings distribution 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, New York Times 2013), but leniency bias is common in many occupations beyond teaching (Prendergast 1999).

However, leniency bias does not necessarily threaten our interpretation of Figure 1 as the causal returns to experience. To violate Assumption 1—ν_k 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 decline in Figure 3 after the first year, for teachers in Tennessee. 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 an official observation visit. Consider the case of a teacher rated 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 official 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, a principal may raise (or lower) observation scores to reflect the principal's beliefs about the teacher's performance of tasks not covered by the rubric. A principal using outside information is a potentially rational behavior if the observation ratings are used for personnel decisions and the principal cares much less about observation scores than she cares about student outcomes and teacher value-added to those outcomes.

This outside information explanation may threaten Assumption 1—ν_k does not depend on experience—but only if raters both 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 likely will 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 between years-worked-together and experience is 0.17 in the DCPS data and 0.15 in the Tennessee data.

One test relevant to this outside-information question is the event study of ratings in Figure 6. Event time is relative to a change in the school principal, with year zero the new principal's first year, and we allow the time series to differ for early-career and veteran teachers as shown by the two plotted lines. 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 (about 5 percent of a standard deviation lower).

Figure 6 alone cannot exclude the threat. While Figure 6 suggests principals do not use outside information, that interpretation assumes the information principals have is increasing year over year. Perhaps principals get to know teachers very well in just one year, and then information increases very little after the first year. The outside information gained in year one could affect observation scores in future years, but we would not see change over time in Figure 6. We emphasize these points as a reminder that the empirical tests in this section are intended to be informative, about both causal inference considerations and policy debates, but these tests are not dispositive.[26]

3.4 Incentives and distortion of effort

Changes in ratings may reflect changes in the incentives attached to those ratings. Those incentives might be explicitly linked to observation ratings, like monetary bonuses or the threat of dismissal, or less-explicit career concerns incentives. Still, a change in incentives alone does not threaten inferences about true performance, θ_k, for tasks covered by the rubric. A new or stronger incentive attached to task k's score, s_k, can 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. Thus, inferences about true performance, θ_k, of tasks covered by the rubric are not necessarily threatened by a change in incentives attached to ratings, s_k.

However, an increase in effort for tasks covered by the rubric, k, may come at the expense of teacher performance in other tasks not covered, $-k$. This asymmetry between scored tasks and un-unscored tasks suggests scope for the well-known multitask distortion problem (Holmstrom and Milgrom 1991). Given that potential distortion, a change in incentives attached to rubric ratings can threaten inferences about teacher performance beyond the scope of what is covered by the rubric. Recall that the rubric tasks are inputs to the broader education production responsibilities of teachers, including improving student math achievement, social skills, earnings as an adult, etc.

[26] On additional note on rater behavior. As described in Section 2.1, 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 equal to one. 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.

Still, using ratings and incentives to shift teacher effort away from some tasks and toward other tasks will not necessarily lead to 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, Burgess, Rawal, and Taylor 2021, Briole and Maurin in-press). In DCPS specifically, teacher performance improves 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 as the returns to experience, 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 ratings in years four and five.[27] Thus, teachers have somewhat more incentive to focus effort on the rubric-measured tasks in years four and five compared to years one, two, and 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, the evaluation incentives in DCPS were not explicitly a function of years of experience but could have been 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

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

"Highly Effective" (the top rating). Figure 7 shows the proportion of teachers in each rating category by years of experience, suggesting the incentives are not strongly correlated with experience.[28]

3.5 Manipulation of ratings

Observation ratings 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 ratings, s_k, but do not raise their true job performance, θ_k. In the literature on job performance evaluation these actions are known as manipulation.[29] This manipulation of observation ratings 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 in advance of the rater's visit. By contrast, if the evaluation process or incentives prompted a teacher to improve her lessons on all (or 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 ν_k component in Equation 2. A teacher's awareness of how to manipulate likely grows as he gains experience with the evaluation system. That suggests a plausible correlation between potential for manipulation and general teaching experience, which threatens Assumption 1 that ν_k is invariant to experience. However, that correlation might be weakened if more-experienced teachers share their manipulation

[28] 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.

[29] 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).

strategies with newly-hired teachers. 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 after year one in Tennessee in Figure 3 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 the first to the second year. And there are other limitations to the test in Figure 3, as discussed above. On the other hand, while underpowered, the evidence from DCPS in Figure 3 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 ratings and student survey scores across levels of teaching experience in DCPS (Appendix Figure A5) 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_k, 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 ratings did improve in 2011-12 for these teachers, on average. 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. Taken together, these results suggest that the dismissal threat did not improve observation ratings through manipulation alone.

3.6 Changes in job assignments

Changes in a teacher's ratings may reflect changes in her job assignment. A teacher's observation ratings, s_k, might decline (or improve) after a job change for either of two reasons: First, the teacher's actual performance, θ_k, could decline (or improve) because of the job change. Using teacher value-added to student test score, 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_{kjta} is the actual performance of teacher j in task k during school year t with job assignment a. We can write:

$$E\left[\theta_{kjt} - \theta_{kj(t-1)}\right] = \underbrace{E\left[\theta_{kjta} - \theta_{kj(t-1)a}\right]}_{\Delta^t} + p\,\underbrace{E\left[\theta_{kj(t-1)a} - \theta_{kj(t-1)a'}\right]}_{\Delta^a} \quad (5)$$

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 5. If identifying Assumption 2 holds—no returns to additional experience 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. 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 $E\left[\theta_{kjt} - \theta_{kj(t-1)} | expr \geq \bar{e}\right] = 0$ 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 veteran (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.[30]

The second reason scores might change is that the ν_k component in Equation 2 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, ν_k 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 ν_k discussed above. As long as job-specific differences in ν_k 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 8 we test the robustness of Figure 1 to changes in the students a teacher is assigned. Using data from Tennessee and DCPS, we plot returns-to-experience estimates with and without controls for students prior-year test scores.[31] Accounting for changes in students assigned does not affect our estimates. The similarity of all the estimates in Figures 1 and 8 is partly because they all use only within-teacher variation. The ν_k component 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.

[30] 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.

[31] The estimation for Figure 8 is identical to our preferred strategy used in Figure 1 with two exceptions. First, we limit the sample to teacher-by-year, jt, observations where we have prior-year test scores for students assigned to the teacher, grades 4-8 math and language classes. Second, for the dashed line, the outcome variable is the residual from a regression of observation score, $\bar{s}_{jt}$, on the average prior-year test score for students assigned to the teacher.

3.7 Performance improvements among veteran teachers

The true 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 experience for early-career teachers. The veterans' improvements would be subtracted off any improvements for early-career teachers.[32]

3.8 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 ratings might rise with experience, even if each individual teacher's scores remain constant, if lower-rated teachers are more likely to leave teaching (or at least leave the district or state).

Still, even using only within-teacher variation, 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

[32] 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.

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 9 shows our returns-to-experience estimates using subsamples defined by when the teacher leaves teaching in Tennessee or DCPS. The changes from year one to two, two to three, etc. are quite similar across samples. The exception is that the trajectory appears to change in a teacher's final year before leaving teaching in Tennessee or DCPS.

4. Conclusion

The typical estimates of returns to experience, applied to observation ratings, can reasonably be interpreted as the causal effect of additional experience on teachers' job performance—specifically, performance of the input tasks covered by the rubric. The estimates are 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, these tests are not dispositive; 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 in the very first years of teaching. 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.

Our analyses should be interpreted carefully. First, we focus on the performance of the input tasks covered by classroom observation rubrics. Stronger assumptions are required when using observation ratings to make inferences about teacher performance measured by contributions to student outcomes. Second, taking differences in scores over time addresses many concerns. But several of those concerns would remain when making claims based on score levels at a single point in time. Third, the various tests in Section 3 are intended to be informative about potential threats to casual interpretation, but those tests are not dispositive. Partly because some test results have competing interpretations, as with Figure 6, and partly because we have data from specific settings. We hope future work in other settings will repeat these tests and add new ones.

A final note of caution is that our estimates using data from Tennessee and DCPS may differ from estimates in other settings employing teacher observations. Our identification strategy—using differences in scores over time and between early career and veteran teachers—can be applied to other settings. However, the implementation of observations in other settings may open those systems to violations of the identifying assumptions explained and explored here.

Our own prediction is that our results will generalize well to other settings. The Tennessee and DCPS settings are quite similar to other settings on many design features: the detail and content of the observation rubric, the frequency and duration of observation visits, the use of school principals as observers, and others (Steinberg and Donaldson 2016, Kraft and Gilmour 2017). Though other features differ: the formal incentives attached to teachers' observation ratings, DCPS's use of master educators as raters, and others. Comparisons of the actual data produced by different systems—including variation in scores and reliability—are scarce and quite limited.[33] In general, teacher observation ratings have moderate to low reliability (Kane and Staiger 2012, Ho and Kane 2013, Taylor 2023), but there is

[33] The best exception is Kraft and Gilmour (2017), who find that scores from Tennessee may have less variation than other settings. However, the comparison is limited to the percent of teachers who score at or above expectations on the end-of-year "summative" ratings. Another exception is Weisberg et al. (2009) where the data proceed current designs, and indeed helped to spur those designs.

no evidence that Tennessee and DCPS systems produce scores which are more (or less) reliable than other systems.[34] Tennessee and DCPS were early adopters of the redesigned teacher evaluation systems of the last decade, and our estimates may be more relevant to well established systems. The DCPS system provided strong and atypical financial incentives, which may have motivated teachers to improve their skills more quickly.

In general, other settings may have their own new sources of error and bias in observation scores—the ν_k component—when those scores are used to measure teacher performance. But estimates of the returns to experience will be robust to those sources of error and bias under Assumption 1, that ν_k does not vary with teacher experience. One example from our study illustrates this point. It is plausible that the DCPS master educators might produce more reliable scores than school principals, but our estimates of the returns to experience were similar for both types of raters.

Finally, we see two ways for our work to benefit teacher policy. First, a better understanding of how experience causes improvements in teaching provides a necessary foundation for several questions relevant to policy design. For example, does early-career development depend on formal training, either in teacher certification programs or professional development for new teachers? Or does early-career development come though learning by doing? Do teachers improve differentially across the various tasks of teaching, like managing student behavior, planning, or instruction? Is some minimal level of expertise in certain tasks a prerequisite for the development of other tasks? Answering these questions could provide useful insights to school managers and policymakers developing policies

[34] Additionally, the causes of observation scores (un)reliability are not well understood. For example, Ho and Kane (2013) find that a teacher's own principal (direct supervisor) produces a more reliable observation score for her than does a principal from another school. Burgess et al. (2023) find that peer observers who received very little training produce scores as reliable as found in research using highly trained researchers as observers.

and practices intended to help teachers develop new skills. However, without credible evidence that measured performance improvements reflect true performance improvements, the resulting policy insights disappear.

Second, a better understanding of how experience causes improvements in teaching is a critical input to implementing existing policies. For example, many states and districts now use classroom observation ratings to inform tenure decisions. Optimizing performance-based tenure decisions requires understanding how teacher performance improves with experience. That improvement trajectory is relevant to deciding when to make tenure decisions. And that same improvement trajectory is also relevant to understanding the key cost of denying tenure: a dismissed teacher will be replaced by a novice new hire. In the first year, the novice new hire will likely under-preform the dismissed teacher she replaces, but the costs and benefits accrue over an entire career not just one year. Similarly, how teacher performance changes over time is relevant to designing teacher compensation, including bonuses or salary increases linked to observation ratings.

Whether and how teachers learn new skills is central to education policy decisions about selecting and investing in teachers. Our focus in this paper is estimating the returns to experience—improvements in performance caused by teaching experience. But that focus is motivated by the underlying teacher policy decisions and debates. We provide evidence that the gains in classroom observation ratings, for Tennessee and Washington DC teachers, do reflect stronger teaching performance resulting from early career teaching experience.

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.

Altonji, Joseph G., and Nicolas Williams. 1992. "The Effects of Labor Market Experience, Job Seniority, and Job Mobility on Wage Growth." NBER Working Paper 4133.

Angrist, Joshua D. 1990. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." *American Economic Review* 80(3): 313-336.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher quality and learning outcomes in kindergarten." *Quarterly Journal of Economics* 131(3): 1415-1453.

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

Briole, Simon, and Éric Maurin. In-press. "There's Always Room for Improvement: The Persistent Benefits of Repeated Teacher Evaluations." *Journal of Human Resources.*

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. 2021. "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools." *Journal of Labor Economics.*

Burgess, Simon, Shenila Rawal, and Eric S. Taylor. 2023. "Teachers' Use of Class Time and Student Achievement." *Economics of Education Review 94*: 102405.

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.

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

Chi, Olivia L. 2021. "A Classroom Observer Like Me: The Effect of Demographic Congruence between Teachers and Raters on Observation Scores." EdWorkingPapers 21-470.

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.

De Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022a. "Two-way fixed effects and difference-in-differences estimators with several treatments." NBER Working Paper 30564.

De Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022b. "Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey." NBER Working Paper 29734.

Dee, Thomas S., Jessalynn James, and Jim Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from DCPS." *Education Finance and Policy* 16(2): 313-346.

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.

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.

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106-137.

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

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.

Grogger, Jeffrey. 2009. "Welfare reform, returns to experience, and wages: using reservation wages to account for sample selection bias." *Review of Economics and Statistics* 91(3): 490-502.

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.* 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.

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.

James, Jessalynn, and James Wyckoff. 2020. "Teacher Labor Markets: An Overview." In *The Economics of Education (Second Edition)* ed. Steve Bradley and Colin Green. 355-370.

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., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. 2013. *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Bill & Melinda Gates Foundation.

Kane, Thomas J., and Douglas O. Staiger. 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Bill & Melinda Gates Foundation.Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying effective classroom practices using student achievement data." *Journal of Human Resources* 46(3): 587-613.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2018. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." *Review of Educational Research* 88(4): 547-588.

Kraft, 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."

Leaver, Clare, Owen Ozier, Pieter Serneels, and Andrew Zeitlin. 2021. "Recruitment, Effort, and Retention Effects of Performance Contracts for Civil Servants: Experimental Evidence from Rwandan Primary Schools." *American Economic Review* 111(7): 2213-2246.

Nagler, Markus, Marc Piopiunik, and Martin R. West. 2020. "Weak Markets, Strong Teachers: Recession at Career Start and Teacher Effectiveness." *Journal of Labor Economics* 38(2): 453-500.

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

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* 16(2): 283-312.

Prendergast, Canice. 1999. “The Provision of Incentives in Firms.” *Journal of Economic Literature* 37(1), 7-63.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. “Teachers, schools, and academic achievement.” *Econometrica* 73(2): 417-458.

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

Rockoff, Jonah E. 2008. “Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City.” National Bureau of Economic Research Working Paper No. 13868.

Roth, Jonathan, Pedro H. C. Sant'Anna, Alyssa Bilinski, and John Poe. 2022. “What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature.” arXiv:2201.01194

Steinberg, Matthew P., & Morgaen L. Donaldson. 2016. “The New Educational Accountability: Understanding the Landscape of Teacher Evaluation in the Post-NCLB Era.” *Education Finance and Policy 11*(3): 340-359.

Steinberg, 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.

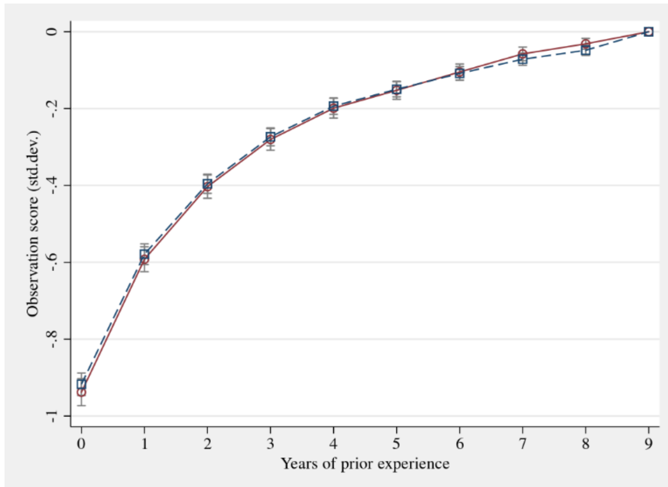
Taylor, Eric. 2023. “Teacher Evaluation and Training.” In *The Handbook of the Economics of Educaiton, Volume 7,* ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann.

Taylor, Eric, and John Tyler. 2012. “The Effect of Evaluation on Teacher Performance.” *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

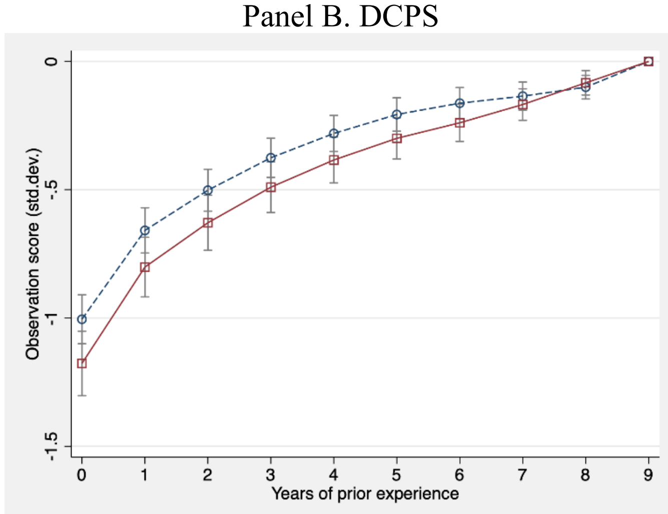
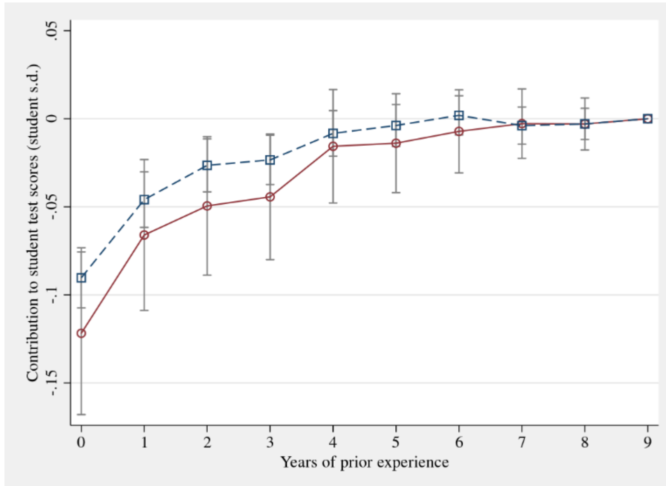


Figure 1—Returns to experience measured in classroom observation ratings

Note: The solid line reports estimates using our preferred diff-in-diff strategy described in Section 2.1. The dashed line reports estimates using the conventional two-way fixed effects approach described in Section 2.4. 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 ratings 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 solid 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 prior 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 ≥ 9 years of prior 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. For the dashed 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 prior experience 0 through 8 individually, with ≥ 9 years the omitted category, but no other controls. The plotted points are the coefficients on the experience indicators. The sample size for the dashed line in Tennessee is 375,072 teacher-by-year observations for 81,847 unique teachers;

and similarly 349,920 and 66,156 for solid line Tennessee, 33,484 and 7,268 for dashed line DCPS, and 33,040 and 7,201 for solid line DCPS.

Panel A. Tennessee



Panel B. DCPS

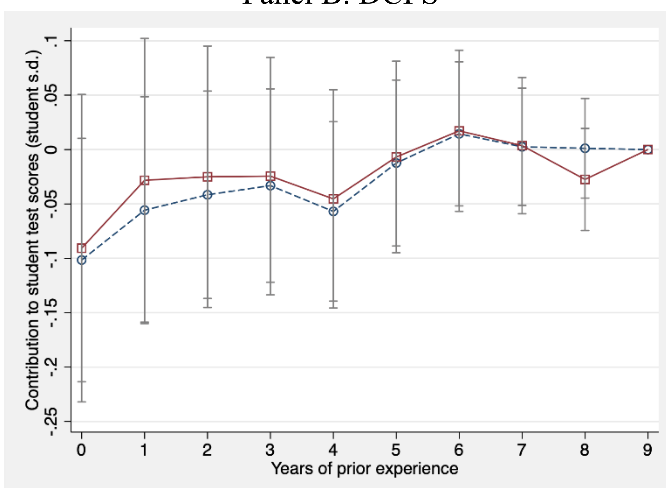
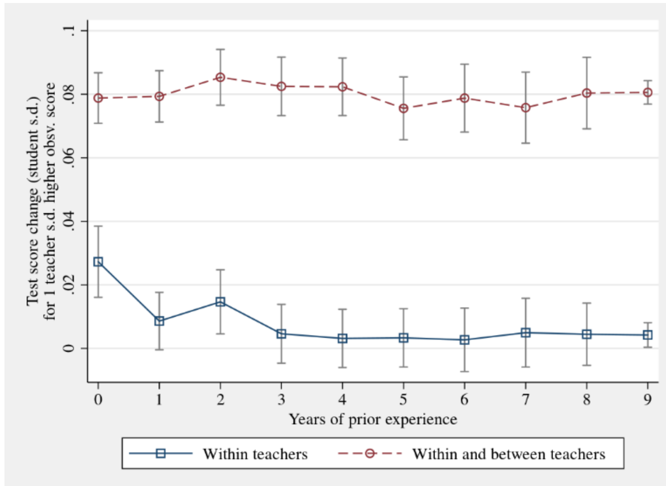


Figure 2—Returns to experience measured in value-added contributions to student achievement

Note: The solid line reports estimates using our preferred diff-in-diff strategy described in Section 2.1. The dashed line reports estimates using the conventional two-way fixed effects approach described in Section 2.4. 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 dashed 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 prior experience 0 through 8 individually, with ≥ 9 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 solid 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 prior 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 ≥ 9 years of prior 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 dashed 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 solid line Tennessee, 247,005, 5,413 and 2,268 for dashed line DCPS, and 4,249 and 1,280 for solid line DCPS.

Panel A. Tennessee



Panel B. DCPS

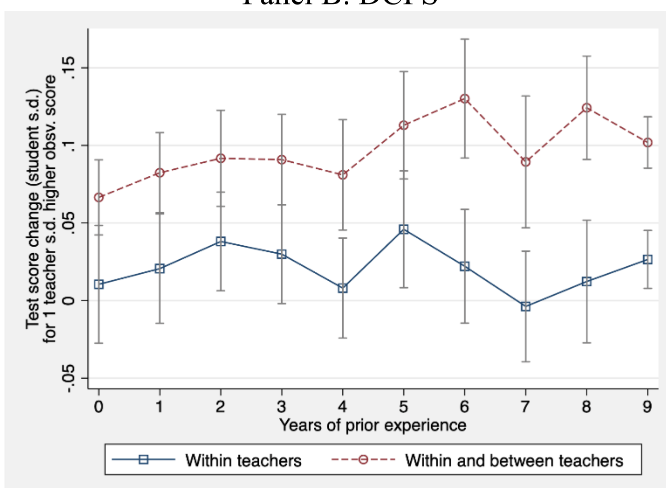
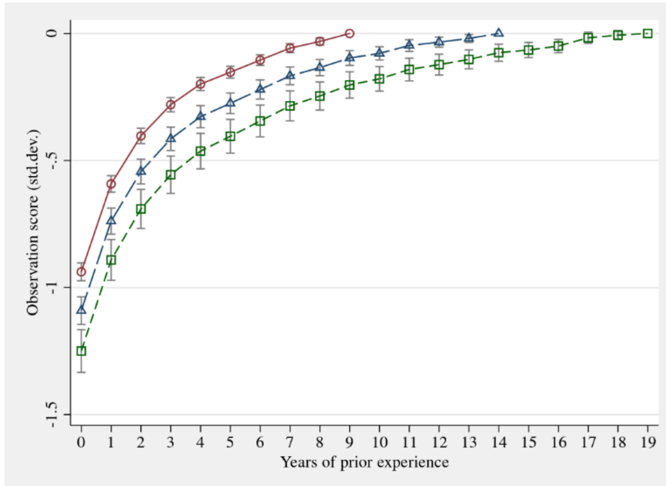


Figure 3—Predicting student test scores with teacher observation scores
by years of teacher experience

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 prior experience 0 through 8 individually, with ≥ 9 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 prior 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.

Panel A. Tennessee



Panel B. DCPS

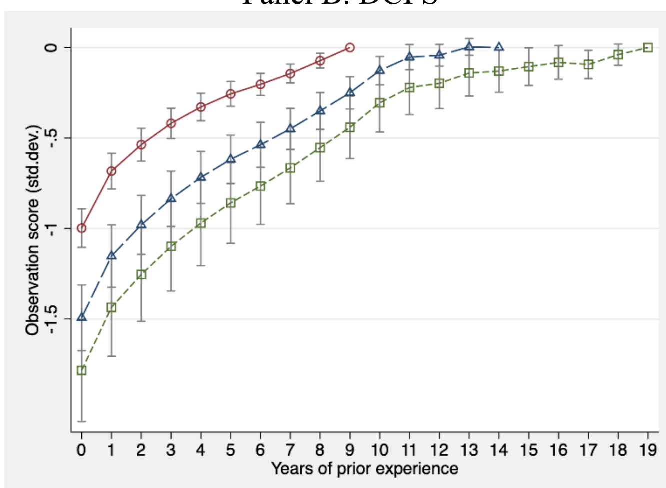


Figure 4—Estimates by definition of comparison group

Note: Each of the three lines reports estimates using our preferred diff-in-diff strategy described in Section 2.1. 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 ≥ 9 years of experience, $\bar{e} = 9$. The two dashed lines show $\bar{e} = 14$ and $\bar{e} = 19$ 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.

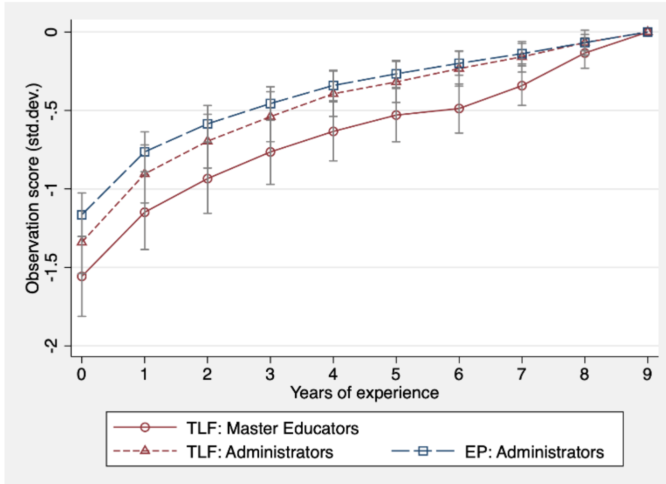
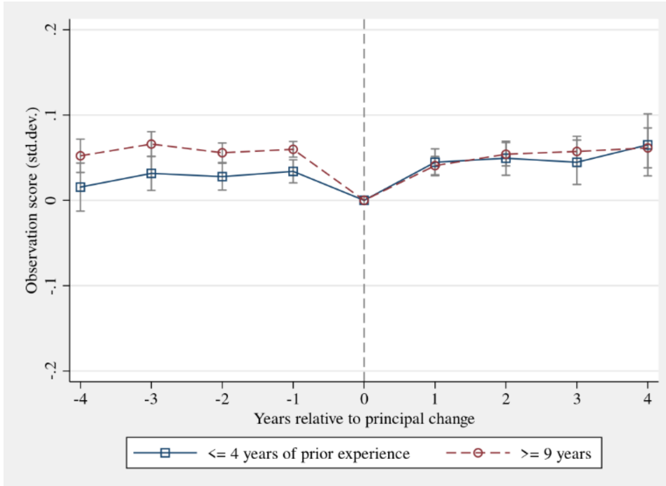


Figure 5—Estimates using different rubrics and rater types (DCPS)

Note: Each of the three lines reports estimates using our preferred diff-in-diff strategy described in Section 2.1. 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

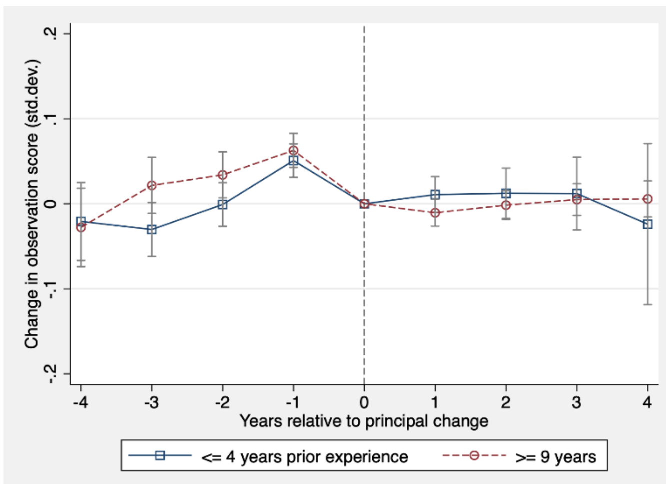


Figure 6—Event study of a change in school principal

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 ≤ 4 years of prior experience, and $= 0$ if teacher j has ≥ 9 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 prior experience, with ≥ 9 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.

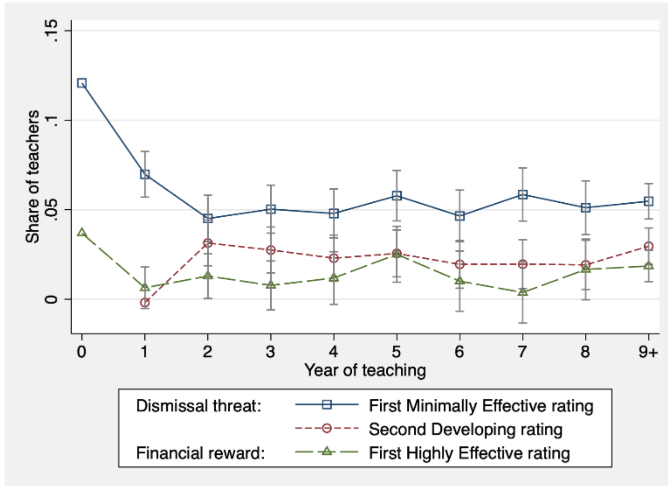
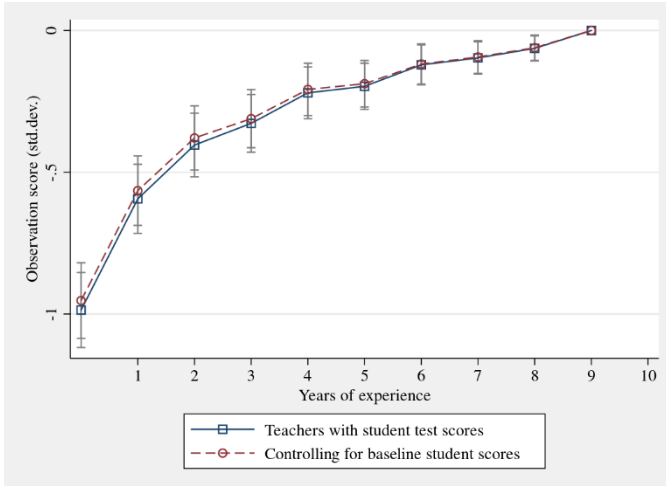


Figure 7—Incidence of consequential performance ratings (DCPS)

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

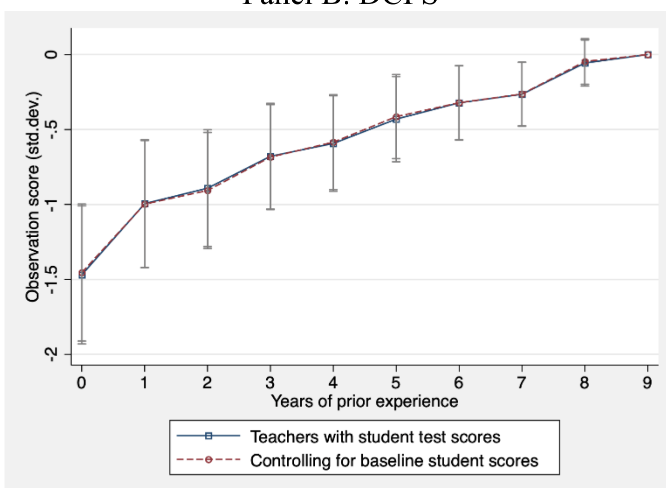
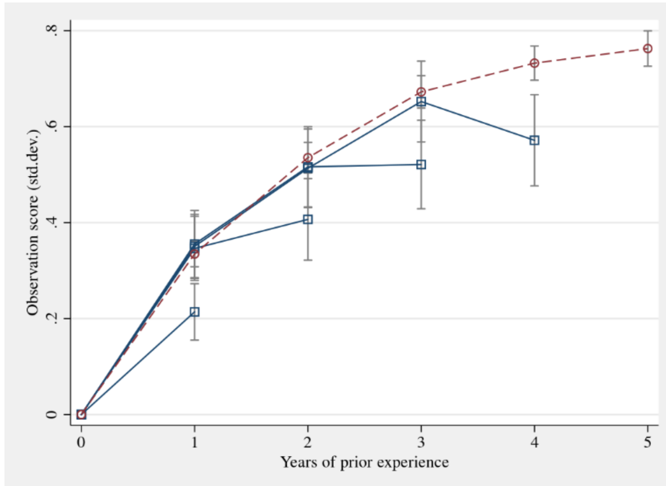


Figure 8—Estimates controlling for student baseline test scores

Note: Both the solid and dashed lines report estimates using our preferred diff-in-diff strategy described in Section 2.1. Both use the same identical sample of teacher-by-year observations. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). For the solid line "teachers with baseline test scores" estimates, the details of estimation are identical to the solid line in Figure 1 except that we restrict the estimation sample. The solid line sample includes only teacher-by-year observations where we have both an average observation rating, $\bar{s}_{jt}$, and baseline test scores, $A_{i(t-1)}$, for the students i assigned to teacher j in year t. For the dashed line "controlling for baseline test scores" estimates, the details of estimation are identical to the solid line except that we first residualize the outcome, $\bar{s}_{jt}$, using the mean baseline test score, $A_{i(t-1)}$, among teacher j's students. 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

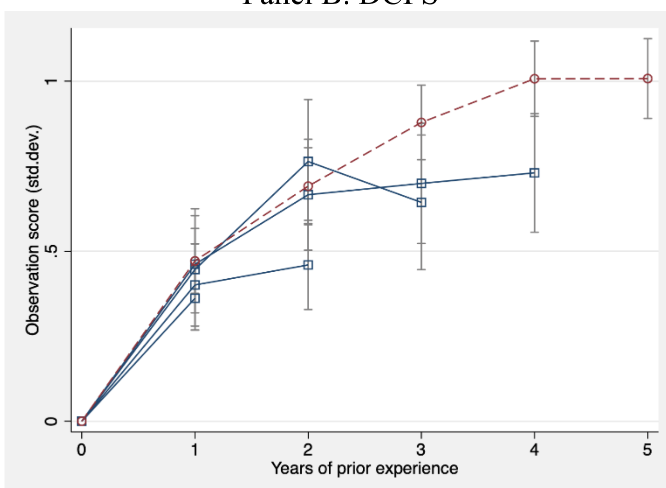


Figure 9—Estimates by year of exit

Note: All lines report estimates using our preferred diff-in-diff strategy described in Section 2.1. The vertical lines mark the 95 percent confidence intervals which are corrected for clustering (teacher). For each line in the figure, the details of estimation are identical to the solid line in Figure 1 except for the estimation sample. The sample for each of the four solid lines is defined by how many years the teacher taught in the jurisdiction (Tennessee or DC). Each teacher is observed for exactly 2, 3, 4, or 5 consecutive years and then not observed in the data subsequently. The dashed line includes teachers observed for 6 or more consecutive years. 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 (1)	DCPS (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.

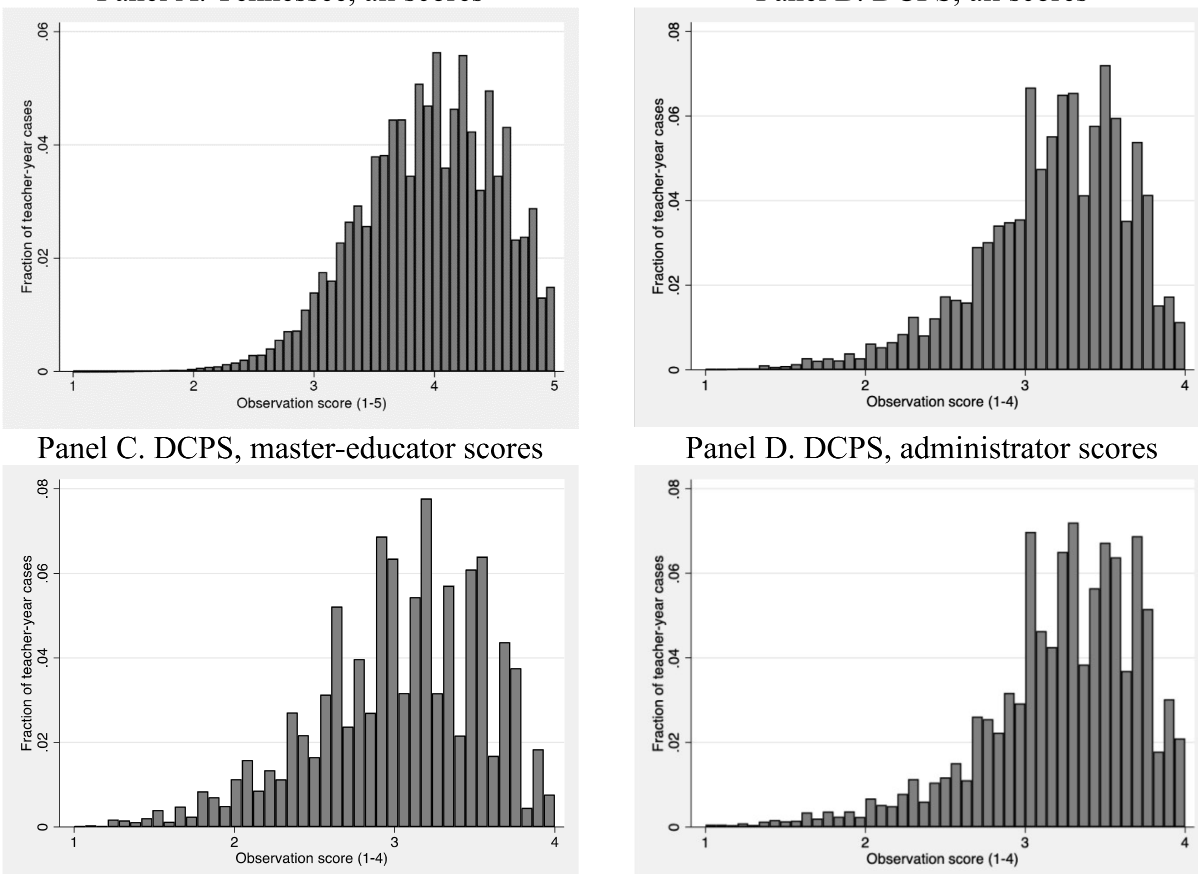
Appendix A. Additional figures and tables

Panel A. Tennessee, all scores

Panel B. DCPS, all scores

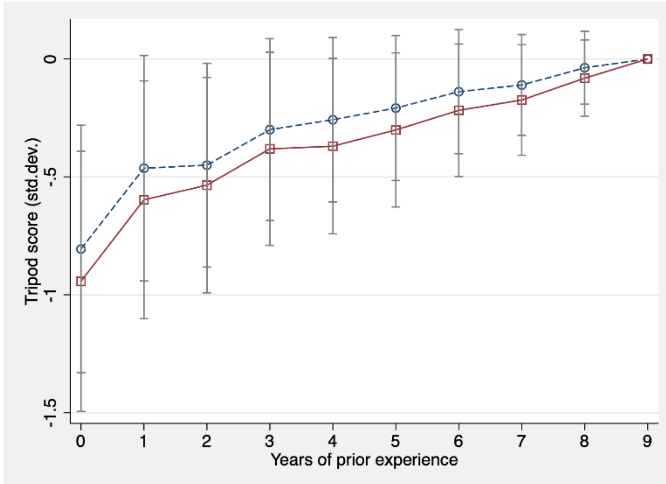
Panel C. DCPS, master-educator scores

Panel D. DCPS, administrator scores



Appendix Figure A1—Distribution of observation scores

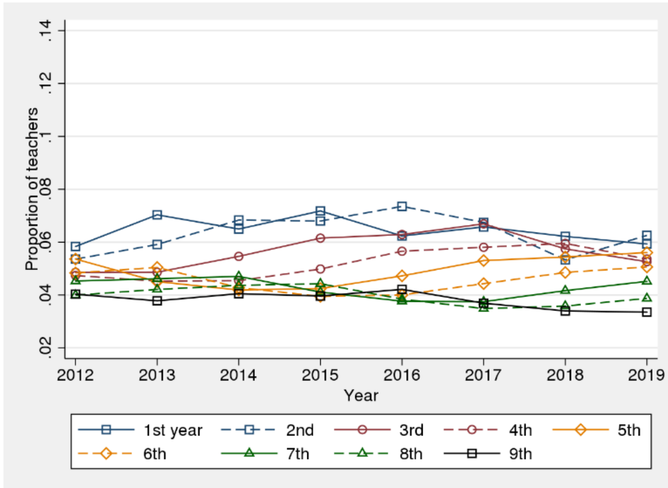
Note: Histograms of teacher-by-year observations. The x-axis is a teacher's annual observation score, which is an average of scores for different items or tasks, in the original rubric-scale units. Data are from the Tennessee TEAM rubric 2011-12 through 2018-19, and DCPS TLF rubric 2009-10 through 2015-16. 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.



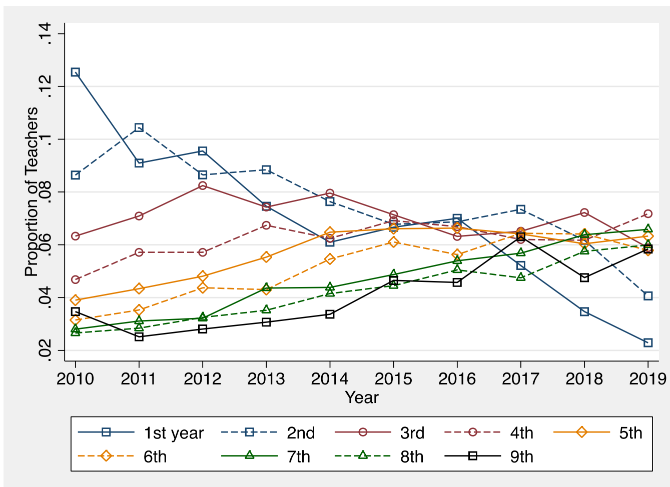
Appendix Figure A2—Returns to experience measured in scores from student surveys (DCPS)

Note: The dashed line reports estimates using our preferred diff-in-diff strategy described in Section 2.1. The dashed line reports estimates using the conventional two-way fixed effects approach described in Section 2.4. 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 this figure 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.

Panel A. Tennessee



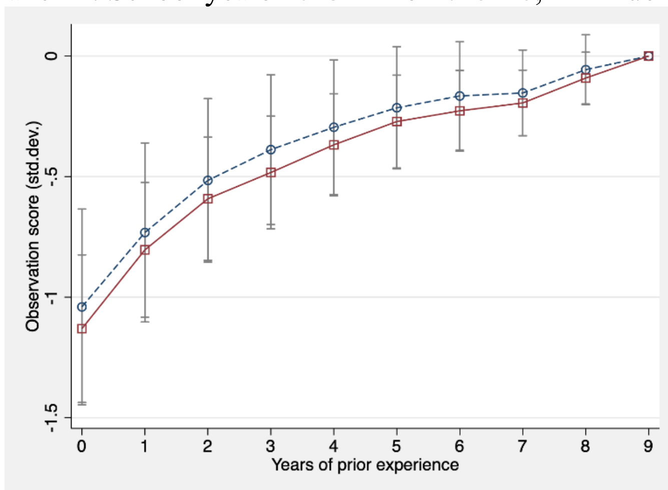
Panel B. DCPS



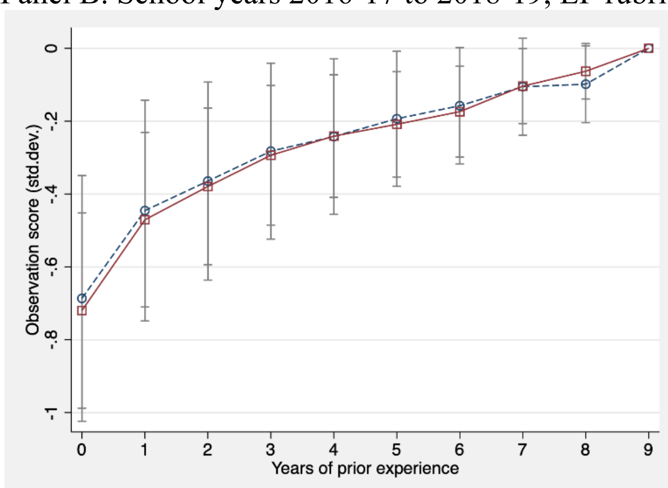
Appendix Figure A3—Distribution of teacher experience over time

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

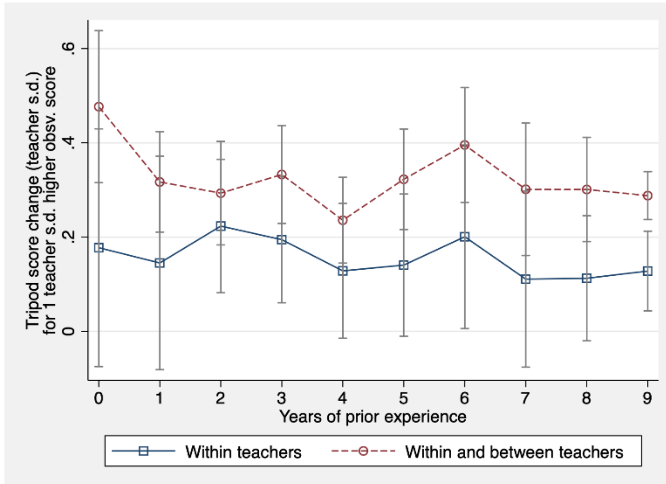


Panel B. School years 2016-17 to 2018-19, EP rubric



Appendix Figure A4—Estimates when the distribution of experience is relatively stable (DCPS)

Note: The solid line reports estimates using our preferred diff-in-diff strategy described in Section 2.1. The dashed line reports estimates using the conventional two-way fixed effects approach described in Section 2.4. 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 A5—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 prior experience 1 through 8 individually, with ≥ 9 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.

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 is 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.

Appendix B: Details of Estimates Involving Teachers' Value-Added Contributions to Student Test scores

B.1 Estimates for Figure 2 Solid Line

The solid line in Figure 2 plots returns-to-experience estimates where the performance measure is a teacher's value-added contributions to student test scores. We first obtain value-added scores, $\hat{\mu}_{jt}$, following the procedure described in the next two paragraphs, then we apply the estimator in Equation 1 substituting $\hat{\mu}_{jt}$ for $\bar{s}_{jt}$. In Figure 2, the y-axis, $\hat{\mu}_{jt}$, is measured in student standard deviation units, and the sample is limited to teachers of grades 4-8 in math and English language arts.

To estimate $\hat{\mu}_{jt}$ we first fit the following regression specification, separately for Tennessee and DCPS data:

$$A_{ijst} = b\big(A_{is(t-1)}\big) + \lambda_j + u_{ijst} \tag{B.1}$$

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 jurisdiction-by-year-by-subject-by-grade cells, where jurisdiction is either the state of Tennessee or the DCPS district). The function $b\big(A_{is(t-1)}\big)$ is a flexible function of student i's prior year test score in subject s, specifically, a quadratic in $A_{is(t-1)}$ where the parameters are free to differ across grade-by-school-year cells. Finally, the λ_j term represents teacher fixed effects.[1]

After fitting Specification B.1, we calculate the modified residuals: $A^*_{ijst} = A_{ijst} - \hat{b}\big(A_{is(t-1)}\big)$ or equivalently $A^*_{ijst} = \hat{\lambda}_j + \hat{u}_{ijst}$. Then our estimate of value added, $\hat{\mu}_{jt}$, is the average residual, A^*_{ijst}, averaging over all students i assigned to

[1] 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.

teacher j in year t, and averaging over subjects s (math and reading) if the teacher taught both. This average residual, $\hat{\mu}_{jt}$, version of a "value added measure" is the same average residual as in step one of the Chetty, Friedman, and Rockoff (2014) or Kane and Staiger (2008) approaches. In the current application we do not "shrink" the estimates because $\hat{\mu}_{jt}$ is the outcome in our analysis.

B.2 Estimates for Figure 2 Dashed Line

The dashed line in Figure 2 plots returns-to-experience estimates where the outcome is also teacher value added, but the estimation methods follow the conventional strategy instead of our preferred strategy. That conventional strategy is described in the next paragraph.

For these estimates we fit a version of the regression specification in Equation 3, but a specification fit with student-level data:

$$A_{ijst} = h(expr_{jt}) + b(A_{is(t-1)}) + \lambda_j + \pi_t + v_{ijst} \tag{B.2}$$

where the function $h(expr_{jt})$ is specified just as it is for the classroom observation outcomes. We repeat Equation 4 here for convenience:

$$h(expr_{jt}) = \sum_{e=0}^{\bar{e}-1} \beta_e \times \mathbf{1}\{expr_{jt} = e\} \tag{4}$$

and $\delta_e = \beta_e - \beta_{e-1}$.

with the omitted category is veterans, $\mathbf{1}\{expr_{jt} \geq \bar{e}\}$. All other details of estimation for B.2 are the same as for fitting B.1. We continue to estimate standard errors using a cluster (teacher) correction.

B.3 Estimates for Figure 3

Figure 3 shows the relationship between observation ratings and test-score value added, and how that relationship changes with teacher experience. The x-axis is years of prior experience. The y-axis is the predicted increase in value added if we increase the teacher's observation score by one standard deviation.

To obtain the estimates in Figure 3 we fit the regression specification in Equation B.2, except that the function $h(expr_{jt})$ is replaced with:

$$h(expr_{jt}, \bar{s}_{jt}) = \alpha_{\bar{e}} \bar{s}_{jt} + \sum_{e=0}^{\bar{e}-1} \beta_e \mathbf{1}\{expr_{jt} = e\} + \alpha_e (\mathbf{1}\{expr_{jt} = e\} \times \bar{s}_{jt}) \tag{B.3}$$

which interacts experience and observation ratings on the right-hand side. Figure 3 plots $(\hat{\alpha}_e + \hat{\alpha}_{\bar{e}})$ for each level of experience, e. The solid line in Figure 3 uses only within-teacher over-time variation, by including teacher fixed effects just as in Specification B.2. The dashed line in Figure 3 uses both within- and between-teacher variation by omitting the teacher fixed effects from the regression specification. As throughout the paper, we estimate standard errors using a cluster (teacher) correction.

References

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates." *American Economic Review* 104(9): 2593-2632.

Kane, Thomas J., & Douglas O. Staiger. (2008). "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.