# ENACTING THE RUBRIC: TEACHER IMPROVEMENTS IN WINDOWS OF HIGH-STAKES OBSERVATION

**Aaron R. Phipps**
(corresponding author)
Department of Social Sciences
U.S. Military Academy at West Point
West Point, NY 10996
aaron.phipps@westpoint.edu

**Emily A. Wiseman**
Curry School of Education
University of Virginia
Charlottesville, VA 22904
ew3kp@virginia.edu

## Abstract

Teacher evaluation systems that use in-class observations, particularly in high-stakes settings, are frequently understood as accountability systems intended as nonintrusive measures of teacher quality. Presumably, the evaluation system motivates teachers to improve their practice—an accountability mechanism—and provides actionable feedback for improvement—an information mechanism. No evidence exists, however, establishing the causal link between an evaluation program and daily teacher practices. Importantly, it is unknown how teachers may modify their practice in the time leading up to an unannounced in-class observation, or how they integrate feedback into their practice post-evaluation, a question that fundamentally changes the design and philosophy of teacher evaluation programs. We disentangle these two effects with a unique empirical strategy that exploits random variation in the timing of in-class observations in the Washington, DC, teacher evaluation program IMPACT. Our key finding is that teachers work to improve during periods in which they are more likely to be observed, and they improve with subsequent evaluations. We interpret this as evidence that both mechanisms are at work, and as a result, policy makers should seriously consider both when designing teacher evaluation systems.

https://doi.org/10.1162/edfp_a_00295

## 1. INTRODUCTION

Improving teacher practice is a policy imperative, as a large body of research shows that high-quality teaching is one of the strongest within-school levers for improving student outcomes (Hanushek 2011). One way to improve teaching practice is through teacher evaluation programs (Taylor and Tyler 2012; Steinberg and Sartain 2015). In particular, researchers argue that including standards-based observation rubrics as an evaluation measure promotes teacher development by providing information to teachers about how they may improve their practice (Papay 2012). There is little empirical evidence, however, addressing the circumstances under which teachers change their practice to align with the information contained in these rubrics: Are improvements made because of feedback received from the evaluation, or does the possibility of evaluation provide motivation and direction to improve practice? This question fundamentally changes the underpinning philosophy and design of teacher evaluation programs. Understanding the factors that lead teachers to improve their practice gives traction to policy makers hoping to replicate the successes of evaluation policy in other contexts. Empirically speaking, a particularly confounding issue is identifying the extent to which in-class evaluations are noninvasive measures of teacher quality as opposed to motivating events that encourage particular practices in the days leading up to an evaluation. Our key result is that teachers work to improve when they are likely to be observed, and they learn from being observed.

Using administrative data from the Washington, DC (hereafter DC) teacher evaluation program called IMPACT, we leverage the exogenous timing of classroom observations to isolate how teachers improve their evaluation score as observation becomes more likely. Importantly, these teaching improvements translate into improved student outcomes, even if the changes to teaching practice in the lead up to an evaluation are not enduring (Phipps 2018). Furthermore, we show how teachers improve their evaluation scores as they gain experience from one evaluation to the next. Taken together, our evidence suggests that teacher evaluation policies should include a classroom observation component to improve student outcomes by encouraging teachers to adopt standards-based practices.

In our period of study, teachers in DC Public Schools (DCPS) experience five classroom observations per year, which take place during district-wide periods of time called windows of observation. Although individual observations are unannounced, the dates at which the windows open and close are publicly available and widely disseminated. Because of the structure of IMPACT, there is random variation in the daily probability of an evaluation within each window, allowing us to causally identify teacher improvements as the result of the increased likelihood of an evaluation as well as the improvements caused by additional experience and feedback.

In making these measured teaching improvements, teachers do not tend to overemphasize one instructional standard to the detriment of others. Effect sizes show that average improvements are brought about by small changes across the nine rubric domains. Additionally, teachers at the high end of the performance distribution appear most responsive to the probability of an evaluation. We hypothesize that high-skill teachers draw on a more robust toolkit of teaching practice to enact rubric standards when an observation is more likely. Finally, we show that probability-based teacher

improvements are consistently positive, independent of the observer's role in the school. However, effect sizes are larger for evaluations conducted by external raters hired by the district, compared with those done by internal raters, like principals.

Our results provide the first quantitative evidence that the classroom observation component of a high-stakes teacher evaluation system encourages instructional best practices. Furthermore, we demonstrate how these improvements come through two key channels: improvements made in anticipation of an evaluation and improvements from having completed an evaluation. In this paper, we synthesize the extant literature about the potential for teacher development within evaluation systems, review the DC context during our period of study, and provide an economic framework for conceptualizing teacher responses to the probability of classroom observation. Then, we discuss our findings and the policy questions that persist.

## 2. BACKGROUND AND CONTEXT

Evaluation systems theoretically may serve two purposes. First, identifying a teacher performance distribution allows for compositional workforce change. High performers may be incentivized with financial rewards and low performers may be sanctioned or dismissed. This accountability mechanism has historically been the focus of the DC teacher evaluation system, and a growing body of research shows that student achievement increases as a result of this accountability mechanism (Dee and Wyckoff 2015; Adnot et al. 2017). However, improving the composition of the teaching workforce through selective dismissal relies on a competitive teacher labor market, and thus the success of this kind of program is variable based on the quality of teachers hired in the place of those dismissed. Researchers found a net negative effect of accountability-oriented, high-stakes evaluation on student achievement in Houston, Texas, and in the Denver, Colorado, program Procomp (Briggs et al. 2014; Cullen, Koedel, and Parsons 2016).

Second, identifying a teacher performance distribution provides information to teachers about their practice. In particular, establishing instructional benchmarks through standards-based observation rubrics and providing feedback relative to those domains may promote self-reflection, collaboration, or other quality-improving behaviors. Taylor and Tyler (2012) theorized this information mechanism as a potential driver of student achievement gains as a result of a low-stakes teacher evaluation system in Cincinnati, Ohio. Similarly, teachers in Chicago, Illinois, improved when they were provided feedback in a low-stakes evaluation system (Steinberg and Sartain 2015). However, this information mechanism relies on input-based measures of practice like classroom observation, which tend to be used more formatively throughout the year. Output-based measures like value added are potentially less useful for informing teachers on how to improve their practice.

Most evaluation systems in the United States use standards-based observation as a primary measure of teacher performance, in part because this component of evaluation enjoys high face validity with teachers (Cohen and Goldhaber 2016; Steinberg and Donaldson 2016). In our own survey of the literature, teacher evaluation systems that use both input- and output-based measures are more successful in effecting student

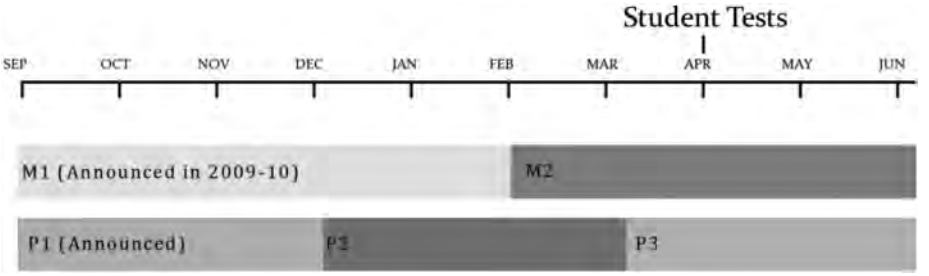
achievement gains than systems using output-based measures alone.[1] This comparison suggests that the information mechanism may be at work in these systems. To that end, recent studies examine the circumstances under which teachers modify their practice in response to standards-based classroom observations. Two studies found that the specificity of the language in observation rubrics may influence teachers to strategically take up practices that are easiest to target. Adnot (2016) found that low-performing teachers facing a dismissal threat in DCPS improved the most on highly specific rubric practices, ranging from a statistically significant effect size of 0.22 to 0.62 standard deviations. Another study found that more explicit, written descriptions of instructional domains correlated with teacher improvement on that rubric component (Kane et al. 2010). Because in-class evaluations may be used as measures of teacher performance as well as formative tools to improve practice, their use in a teacher performance incentive may contribute to teacher improvements both through behavioral changes in preparation for an evaluation (accountability mechanism) as well as through behavioral changes post-evaluation (information mechanism). Our key result is to identify and measure both mechanisms.

Leveraging evaluation systems to improve teaching practice also potentially requires that the information provided to teachers is useful. Whoever conducts the classroom observation, then, may be a crucial lever in communicating this information to teachers. In other contexts, principals still primarily rate teachers as effective despite using more detailed evaluation rubrics with multiple performance categories (Kraft and Gilmour 2016; Grissom and Loeb 2017). Are principals, then, effective as evaluators? Our analysis touches on this issue as well.

### Washington, DC, Context

As part of the growing demand for increased school and teacher accountability, DCPS implemented a high-stakes teacher evaluation program called IMPACT starting in the 2009–10 school year. All teachers in DCPS have large financial incentives that depend on a weighted combination of elements, which mirror many multiple measure teacher evaluation systems of this decade. For most teachers, the largest component of their IMPACT score comes from scored classroom observations based on the district's Teaching and Learning Framework (TLF). The TLF is intended to define criteria that establish effective teaching and addresses various domains, such as maximizing instructional time and checking for student understanding. Where possible, IMPACT scores also include a student test–based value-added score for tested subjects (Math and English Language Arts) in grades 4 through 8, which accounts for only 18 percent of teachers in DCPS. For these teachers, value-added scores make up 50 percent of the final IMPACT score, and in-class evaluations constitute only 35 percent of their final IMPACT score. For the remaining majority of teachers, classroom observation scores make up 75 percent of their final IMPACT score. The other components of IMPACT are small and include an overall school measure of student test score growth, a principal-assessed score of commitment to the school and community, and a teacher's success in reaching instructional

1. For examples of ineffective incentive programs lacking in-class observations, see Springer, Swain, and Rodriguez (2016), Fryer (2013), and Briggs et al. (2014). Examples of effective programs using in-class evaluations include Dee and Keys (2004), Dee and Wyckoff (2015), and Springer et al. (2012).




*Notes:* Each shaded area represents the time-frame in which a teacher must receive an evaluation. Teachers have 5 evaluations that occur in overlapping windows. P1, P2, and P3 are evaluations administered by the principal or assistant principal. M1 and M2 are evaluations administered by a district employee called a master educator.

**Figure 1.** Diagram of Evaluation Windows

goals for grades and subjects ineligible for value-added measures (for more details on the IMPACT program structure, see Dee and Wyckoff 2015).

In DCPS, the overall in-class evaluation score is the average of five in-class evaluations, each of which is weighted equally and conducted in pre-specified time frames, depicted in figure 1. Principals or assistant principals conduct three observations, and external evaluators, typically veteran teachers with demonstrated expertise, conduct the other two.[2] In the first year of the program, the first principal and external evaluations are announced at least a day in advance, though this was changed in subsequent years so that only the first principal evaluation was announced. The remaining four observations (three in the first year) are conducted without notice within a predefined time period or observation window. Each evaluation lasts roughly thirty minutes, and teachers are given a score from 1 to 4 on each of nine components, which are also weighted equally. At the beginning of each year, the rubric guidebook is published publicly for teacher review.

A key element of the IMPACT program is the debriefing conference that evaluators are required to have with each teacher following an in-class evaluation. This meeting usually takes place within one to two weeks of the evaluation, and evaluators lead a conversation about the teacher's scores and comments. Our analysis checks for improvements teachers make as they experience more evaluations, and we hypothesize that the feedback provided during this conference is the main route through which those improvements take place.

The high-stakes nature of this system makes DC a unique context, and previous work seeks to address the effects of these components. Dee and Wyckoff (2015) use the discontinuities in the IMPACT program's reward structure to show that teachers facing dismissal threat significantly improve student achievement gains and in-class evaluation scores. Adnot et al. (2017) find statistically significant student achievement increases as a result of low-performing teacher exits under the IMPACT evaluation policy. In work complementary to our analysis, Phipps (2018) uses a similar identification strategy to disentangle the relative effects of the high-powered incentives and

2. The structure of IMPACT has changed over time, most dramatically in the 2016–17 school year when the standard five classroom observation protocol was decreased to three, and external evaluators were no longer used as observers. The scope of this study includes the 2009–10 school year to the 2011–12 school year.

the feedback provided on observation rubrics from principals and external evaluators. Phipps shows that, by the structure of the IMPACT program, some teachers randomly experience days in which they are guaranteed not to have an evaluation, which he uses to identify the effect of teacher responses to a potential unannounced observation on student test outcomes. He finds that the possibility of an evaluation has substantive effects on student test outcomes in both reading and math. Teachers additionally improve student outcomes after receiving evaluator feedback. Our analysis adds to this body of work by mapping how teachers modify their behavior in response to in-class observations and the TLF rubric.

Underlying our analysis is the assumption that teachers will respond to an evaluation policy efficiently, if at all. There are two reasons that we might question this assumption. First, in-class observations are noisy as indicators of overall teacher practice, which might reduce the attention paid to them. The Gates Foundation's Measures of Effective Teaching study finds that only 35 percent of all observation score variation can be explained by the teacher, suggesting that 65 percent of all variation is a result of factors that produce noise, like occasion and rater variance, as well as error (Kane and Staiger 2012).[3] Second, having multiple independently measured outcomes creates the possibility that teachers will inefficiently shift effort to a subset of outcomes that are easy to improve at the expense of other practices (Holmstrom and Milgrom 1991). We first present a basic mathematical framework of teacher behavior under a high-stakes, multiple domain classroom evaluation. Then, we present the empirical strategy and results.

## 3. MODEL OF TEACHER RESPONSES TO EVALUATIONS

In this study, we examine the extent to which multiple-measure teacher evaluation systems result in teachers shifting their practice, particularly in response to classroom observation. Classroom observations, as implemented in DCPS, are intended to guide and improve teacher practice as well as to provide an objective measure of teacher quality. A critical and unanswered question is the extent to which unannounced evaluations affect teacher behavior in the days leading up to an evaluation. Such behavioral changes can confound attempts to objectively measure teacher quality. On the other hand, if properly directed, teacher responses to high-stakes evaluations may be a valuable tool for improving classroom instruction. Our empirical result relies on the belief that teachers will prepare more as an unannounced classroom evaluation becomes more likely. We therefore present an economic model of behavior to describe how teachers in a high-stakes environment would align their instruction to the observation rubric.

Teachers must choose among a variety of potential teaching styles and lesson structures in order to improve their final rating. The multitasking model originating from Holmstrom and Milgrom (1991) describes the process of allocating time and effort toward a variety of tasks, each of which is rewarded.[4] Let teachers have a set of possible

3. The theoretical result that noisy measures reduce an incentive's effect is expanded in Lazear and Rosen (1981).
4. Holmstrom and Milgrom (1991) include measurement noise on each individual component. Although each TLF component is measured with noise, we ignore measurement noise to facilitate simplicity. The expected qualitative responses are not different.

tasks or practices, called $K$, on which they can focus their time and attention, both in lesson preparation and in the classroom. The work of teaching is incredibly complex, a fundamental point that underscores the importance of considering the ways in which teachers make decisions given numerous tradeoffs and considerations. To simplify, we imagine a teacher who plans her lesson given the needs of her students, the standards on which they are assessed, the curriculum and resources available to her, while taking into account input from colleagues and collaborators, and keeping in mind district and school priorities. She then enacts that lesson using a breadth of instructional knowledge and skills, adjusting her plan based on student mastery of the material. Then let $x_i$ be a teacher's allocation of time toward task $i$, and let $x = [x_1, x_2, \ldots, x_n]^T$ be a vector of all time allocation across tasks, where $n$ is the size of the set $K$. A teacher's allocation of time, $x$, has some utility cost $c(x)$, which has increasing marginal costs for each input. The teacher receives wage $w(x)$ as a result of her score, which is determined solely by $x$ given no measurement noise. Then, with a standard exponential utility function with coefficient of risk aversion $r$, a teacher's utility is

$$U(x) = -\exp\{r(w(x) - c(x))\}.$$

To maximize utility, the first-order conditions require that a teacher chooses $x$ such that the net marginal benefit of each input is zero.

If the bonus system only rewards certain behaviors, say $K' \subset K$, then the marginal benefit of those tasks increases, leading to an increase in how much time and effort a teacher allocates to those tasks. That is, $x_i$ for $i \in K'$ will increase. If the elements in $x$ are cost substitutes, there will also be a decrease in $x_i$ for $i \notin K'$. That is, a teacher will favor elements on the rubric that are easy to adjust over elements that are difficult to adjust or do not earn rewards in the evaluation rubric, a key result of the Holmstrom-Milgrom model. This reflects the limited time available to teachers, in which spending time on preparing one aspect of a lesson or on a specific approach in class naturally requires reducing time spent on another approach.

If evaluations determine a teacher's potential bonus, a teacher will modify her choice of teaching practices, $x$, based on how likely she thinks an evaluation is. When classroom observations have to be conducted once within a prespecified time frame, she can estimate the probability of being evaluated. Intuitively, as the end of the evaluation window approaches, it becomes more likely each day that a teacher who has not already had her observation will be evaluated.

To incorporate the probability of evaluation into the utility function, allow a teacher's utility at the end of the day to be different depending on whether she was evaluated. If she is not evaluated on a given day, her effort exerted on activities that only improved her in-class observation score will not contribute to her utility. In other words, if there is no in-class observation that day, effort toward improving her observation score do not improve her financial reward, $w(x)$. For days in which she is not evaluated, her utility is

$$U(x, m = \text{o}) = -\exp\{-r(-c(x))\},$$

where $m = \text{o}$ indicates she was not evaluated. But if she is evaluated, her utility is as before. To combine these two possible outcomes each day, let $p$ be the probability of an

**Table 1.** School-Level Summary Statistics on Enrollment, Class Size, and School Poverty Status

| | 2009–10 | 2010–11 | 2011–12 |
|---|---|---|---|
| Number of schools | 124 | 121 | 123 |
| Total enrollment | 44,035 | 45,004 | 45,013 |
| Class size | | | |
| Mean | 17.42 | 17.72 | 17.65 |
| Standard deviation | 4.32 | 4.29 | 4.12 |
| Fraction of schools high poverty | 0.771 | 0.750 | 0.772 |

evaluation on the next day. Then in the evening, as a teacher prepares for her next day, her expected utility is

$$EU(x) = pU(x, m = 1) + (1 - p)U(x, m = 0). \quad (1)$$

Equation 1 shows the intuition that as the probability of evaluation increases, there are larger marginal returns to using evaluated practices, $x_i$ for $i \in K'$. As a result, her use of evaluated practices should increase with the probability of an evaluation, leading to an increase in her evaluation score. This result drives our empirical approach: As $p$ increases, teachers will shift their preparation and time toward practices that will improve their evaluation score.

## 4. DATA AND ECONOMETRIC APPROACH

We use administrative data from DCPS from the three school years starting in the fall of 2009 and ending in the spring of 2012. Our data include the date of each of five in-class observations and the subsequent score for each teacher. Using the structure of the IMPACT evaluation program and these data, we identify the potential effects of three variables on a teacher's evaluation score: (1) timing of an evaluation within the school year, (2) the number of prior completed evaluations, and (3) the increased likelihood of an evaluation.

DCPS is a small school district relative to other urban areas, ranking just outside the top 100 school districts nationally by student body. Over the three years studied, there was an average of 45,000 students per year, between 121 and 124 elementary, middle, and high schools, and roughly 3,500 teachers each year. Table 1 summarizes school characteristics over the study period to show that there were no meaningful changes to student or school composition. The average classroom size was constant at about 17.6 students, and the fraction of schools classified as high-poverty ranged between 0.75 and 0.77.

Table 2 provides detailed information on how teacher evaluation scores are distributed over evaluations and years. Because scores are bounded between 1 and 4, we have included percentile measures at the 5th and the 95th percentiles instead of the minimum and maximum. External evaluator scores are lower, on average, than internal evaluator scores but the difference is not statistically significant. The distribution of observation scores is skewed left with the median slightly higher than the mean. Figure 2 depicts the distribution of all evaluation scores across all years with the ratings

**Table 2.** Summary Information on Teacher Observation Scores by Year and Observer

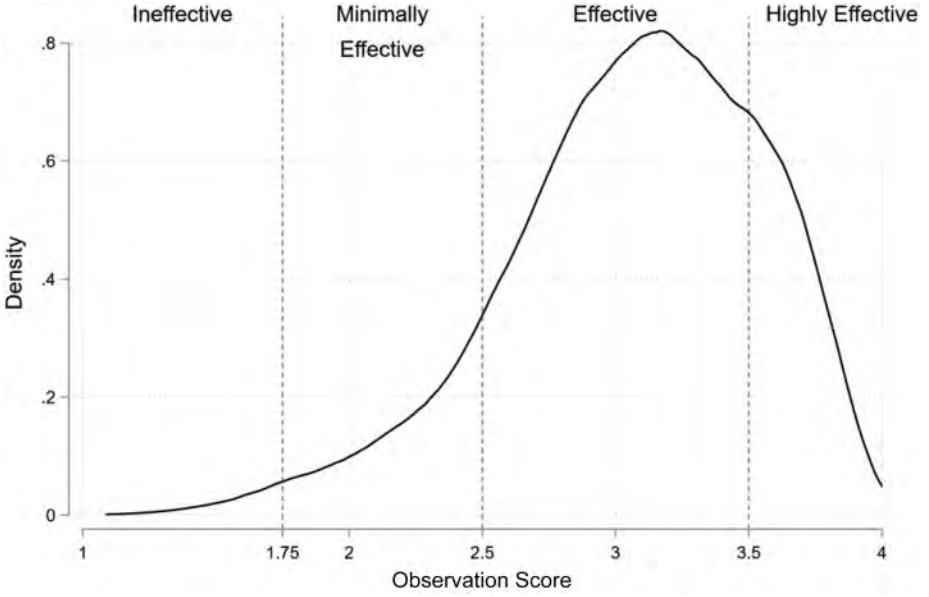
| | Observation Score on Scale of 1 to 4 | | | | |
|---|---|---|---|---|---|
| | Internal 1 | Internal 2 | Internal 3 | External 1 | External 2 |
| 2009–10 | | | | | |
| Mean | 3.113 | 3.149 | 3.200 | 2.963 | 2.993 |
| Standard deviation | 0.622 | 0.644 | 0.651 | 0.597 | 0.602 |
| 5th percentile | 1.907 | 1.815 | 1.833 | 1.852 | 1.815 |
| Median | 3.185 | 3.259 | 3.333 | 3.074 | 3.074 |
| 95th percentile | 3.944 | 3.963 | 4.000 | 3.796 | 3.815 |
| 2010–11 | | | | | |
| Mean | 2.979 | 3.081 | 3.157 | 2.876 | 3.001 |
| Standard deviation | 0.619 | 0.591 | 0.595 | 0.646 | 0.583 |
| 5th percentile | 1.780 | 1.890 | 2.000 | 1.670 | 1.890 |
| Median | 3.000 | 3.110 | 3.220 | 3.000 | 3.000 |
| 95th percentile | 3.880 | 3.890 | 4.000 | 3.780 | 3.780 |
| 2011–12 | | | | | |
| Mean | 3.160 | 3.113 | 3.201 | 2.998 | 2.959 |
| Standard deviation | 0.551 | 0.564 | 0.530 | 0.574 | 0.563 |
| 5th percentile | 2.111 | 2.000 | 2.222 | 2.000 | 1.889 |
| Median | 3.222 | 3.222 | 3.250 | 3.000 | 3.000 |
| 95th percentile | 3.889 | 3.889 | 3.889 | 3.778 | 3.750 |
| Total | | | | | |
| Mean | 3.085 | 3.115 | 3.186 | 2.946 | 2.985 |
| Standard deviation | 0.603 | 0.602 | 0.596 | 0.608 | 0.584 |
| 5th percentile | 1.890 | 1.890 | 2.000 | 1.780 | 1.880 |
| Median | 3.125 | 3.220 | 3.250 | 3.000 | 3.000 |
| 95th percentile | 3.889 | 3.890 | 3.963 | 3.780 | 3.780 |

cutoffs of Ineffective, Minimally Effective, Effective, and Highly Effective drawn. A score between 1.0 and 1.75 is rated Ineffective, between 1.75 and 2.5 is Minimally Effective, between 2.5 and 3.5 is Effective, and above 3.5 is Highly Effective. Across all evaluations and years, only 1.3 percent of teachers received an Ineffective rating, 11.1 percent received Minimally Effective, 67.5 percent received Effective, and 20.2 percent received Highly Effective.

### Calculating Evaluation Probability

A key contribution of our analysis is identifying how teachers may prepare for a pending classroom observation. To calculate the daily likelihood of an in-class observation, we start by identifying which of the five annual evaluations a teacher may receive on a given date, based on the district-provided windows. As shown in figure 1, the first principal evaluation occurs between mid-September and 1 December, the second between 1 December and 1 March, and the third between 1 March and the end of classes. The first external evaluation occurs between mid-September and 1 February, and the second is conducted before the end of classes.

Because we know when each classroom evaluation was conducted, we are able to calculate how many teachers at each school have yet to receive an evaluation during a particular window. Assuming these remaining teachers are drawn at random for the




*Notes:* Kernel density estimate of the distribution of overall in-class observation scores across all three years and all observations. Scores range between 1 and 4. Overall, 1.3 percent of evaluations received an Ineffective rating, 11.1 percent received Minimally Effective, 67.5 percent received Effective, and 20.2 percent received Highly Effective.

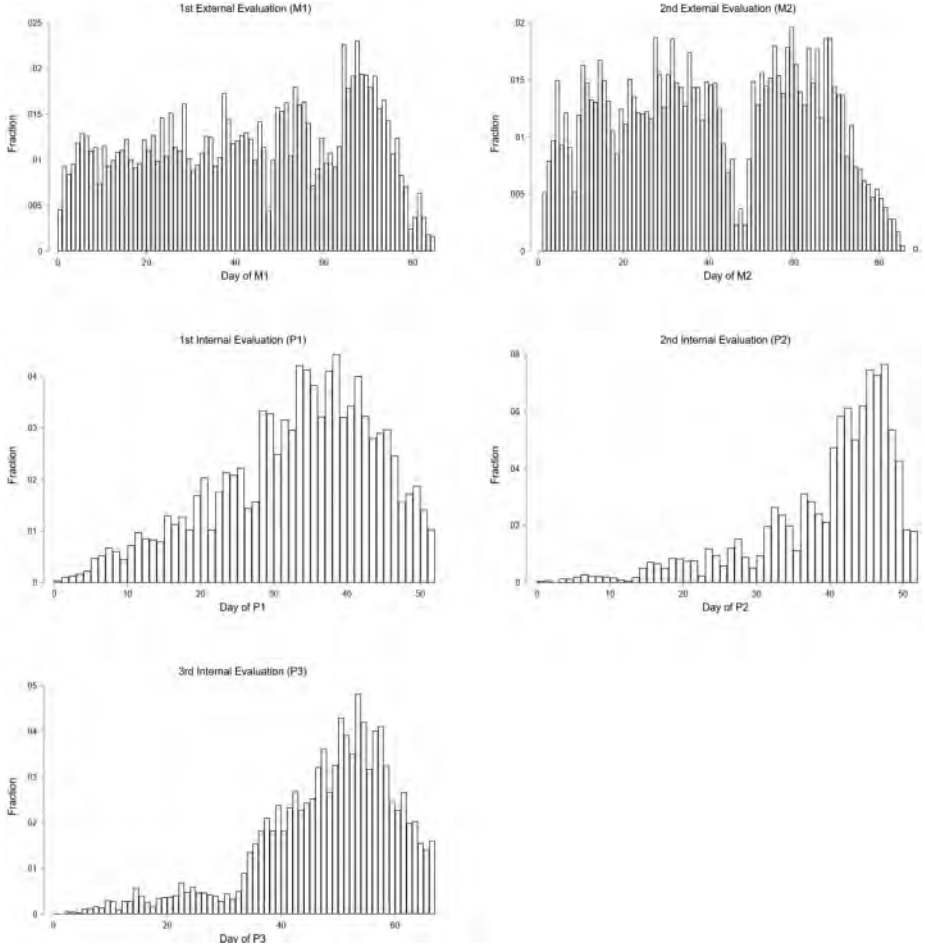
**Figure 2.** Density Plot of Overall Observation Scores

next day's evaluations, we can calculate the expected probability of an evaluation by using the average number of evaluations conducted daily as a fraction of the number of possible teachers. To be more formal, let $k$ be an evaluation indicator, where $k$ is $P1$, $P2$, or $P3$ for the principal evaluations and $M1$ or $M2$ for the evaluations conducted by external evaluators. Then let a teacher's estimate of the number of evaluations to be conducted on day $t$ at school $s$ be $\hat{N}_{ts}^{k}$. If $n_{ts}^{k}$ is the number of teachers who still need evaluation $k$ on day $t$ at school $s$, then each remaining teacher's probability of being evaluated is

$$p_{ts}^{k} = \frac{\hat{N}_{ts}^{k}}{n_{ts}^{k}}.$$

We can determine how many teachers remain to be evaluated, $n_{ts}^{k}$, but estimating how many evaluations a teacher expects to be conducted, $\hat{N}_{ts}^{k}$, requires assumptions about a teacher's knowledge of when evaluators will conduct more evaluations. If a teacher knew exactly how many evaluations would be conducted on every day, then we could simply use the observed number of evaluations each day as the teacher's estimate: $\hat{N}_{ts}^{k} = N_{ts}^{k}$. This is a strong assumption that is unlikely to be true, especially if evaluations are not evenly distributed within a window.

Principals tend to bunch their evaluations near the last third of the observation window, which means the expected number of evaluations changes over time. In the beginning of a window, teachers expect that principals will conduct few evaluations, but toward the end of the window, teachers expect more each day. On the other hand,




*Notes:* Days are measured as instruction days, which excludes in-service days, weekends, and holidays. External evaluations, M1 and M2, are distributed uniformly across the window. Internal evaluations—P1, P2 and P3—are often clustered near the end of each window.

**Figure 3.** Histograms of the Timing of Evaluations within Evaluation Windows

external evaluators distribute their evaluations more evenly, so the expected number of evaluations remains constant. Figure 3 shows the overall distribution of evaluations across each window. While the external evaluators maintain a fairly uniform distribution, principals are very often conducting evaluations in the last third of the available time. The dip in evaluations in *M2* around day 45 is a result of student testing days in April.

Instead of assuming that teachers know exactly how many evaluations will be conducted on each day, we assume they are broadly aware of the distribution of evaluations across the semester. That is, we allow for a teacher to know that the number of evaluations conducted by a principal will increase toward the end of the window. We also allow for a teacher to notice an increase in evaluations over the past few days. We can approximate this information by estimating the distribution of evaluations

with a kernel density. The kernel smoothing approximates changes in the trend of daily evaluations that we expect teachers notice. Our results are not sensitive to this assumption.[5]

For many days in the year, a teacher has the possibility of either a principal evaluation or an external evaluation (or both). The two events are independent and in rare cases both occur on the same day for a single teacher. To determine the probability of any evaluation, we use the sum of their individual probabilities. For example, if a teacher has not yet had either $P1$ or $M1$ evaluations, her probability of *any* evaluation the next day is $p_{ts} = p_{ts}^{P1} + p_{ts}^{M1} - p_{ts}^{P1} \cdot p_{ts}^{M1}$, but if she had already received her $P1$ evaluation, her probability is just $p_{ts} = p_{ts}^{M1}$.[6] We then use $p_{ts}$ in our specification.

Given an estimate of the probability of any evaluation for each day in each school, we know the probability of an evaluation on the day in which a teacher was, in fact, evaluated. We use $P^k$ in capital letters without the subscript $t$ to indicate the probability of any evaluation on the day when evaluation $k$ occurred. For a teacher receiving evaluation $P1$ on day $t$ at school $s$, the treatment variable is $P^{P1} = p_{ts}^{P1} + p_{ts}^{M1} - p_{ts}^{P1} \cdot p_{ts}^{M1}$, assuming she has not yet received her $M1$ evaluation.

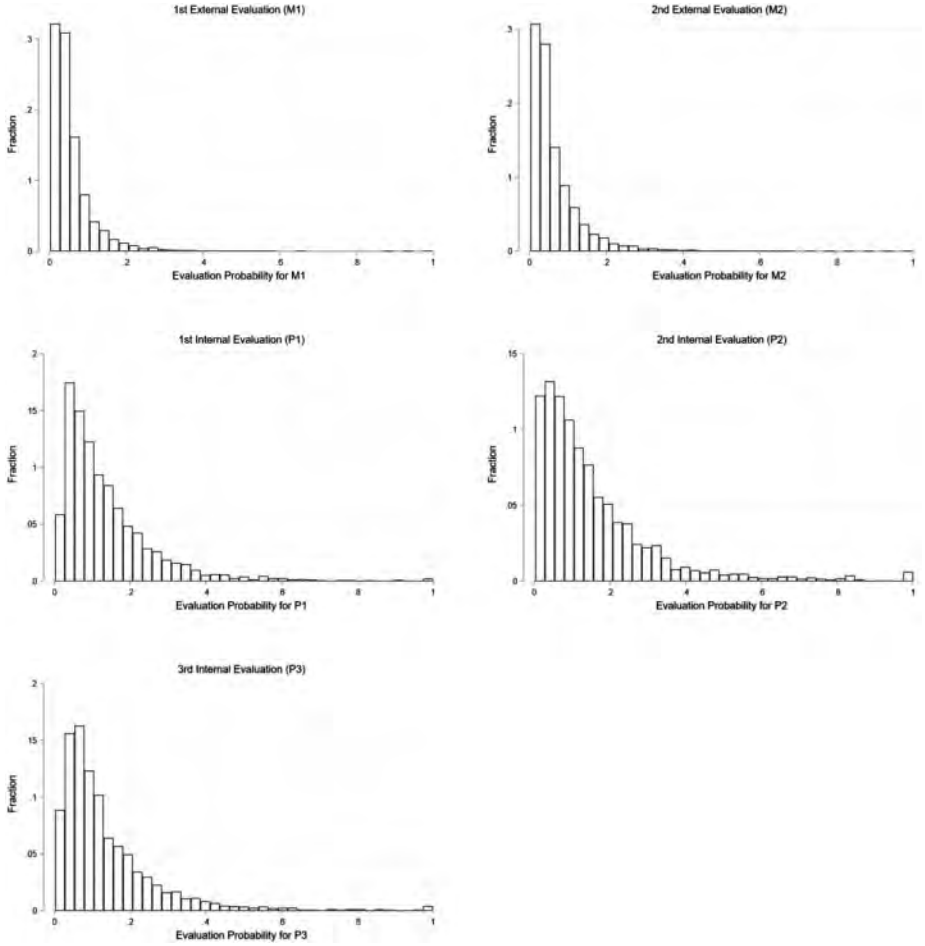
Figure 4 illustrates the distribution of evaluation probability for each of the five evaluations. Because principal evaluations often occur toward the end of the window, more teachers experience higher levels of evaluation probability for these observations than for those conducted by external evaluations. All probabilities are capped at one. The spike at a probability of one for principal evaluations is a result of several teachers being evaluated on the last day of the window. Because these teachers were certain they would be evaluated, their evaluation probability is one. The treatment distributions illustrate that most teachers received their evaluations on days with evaluation probability at or below 10–15 percent.

To understand the effect sizes, we have included a summary of each treatment variable, evaluation probabilities, for each year in table 3. The average evaluation probability for principals is larger than for external evaluators, which is expected given the mass of principal evaluations at the end of each window. We include the minimum probability and the 95th percentile, since probabilities are bounded at 1. The median treatment for principal evaluations is around 10 percent, meaning the median teacher had a 10 percent chance of being evaluated on the day of her principal evaluation. For a visual representation of treatment distribution, figure 5 shows the evaluation probability for each evaluation pooled across all three years.

#### Econometric Specification

The outcome of interest is how teachers perform on their in-class evaluations. We want to assess how a teacher responds to the increased probability of an evaluation, as well

5. To test sensitivity, we instead estimate the probability of an evaluation assuming teachers expect there to be no trend in evaluation timing. This is equivalent to assuming that teachers expect evaluations to be evenly distributed across the evaluation window. Our results are qualitatively unchanged (and are in fact stronger). However, this specification fails to allow teachers realistic foresight about upcoming evaluations, and so we have opted for the kernel estimate.
6. These additive probabilities are capped at one, though the cap was rarely needed (it applied to 0.38 percent of all observations).




*Notes:* Histograms of treatment variable, probability of an evaluation on the day of each evaluation. For internal evaluations (P1, P2, P3), the probability is usually higher because these evaluations occur later in their assigned window.

**Figure 4.** Histograms of Evaluation Probability for Each Evaluation

as identify systematic improvements she may make as the school year progresses and as she completes more evaluations. Because the analysis takes place within the school year, we attempt to control for year- and classroom-specific characteristics. To do so, we use the first principal evaluation as a control for the subsequent evaluations.[7] Because $P1$ is announced, it is not affected by timing in the way subsequent evaluations are. This evaluation also represents a baseline measure for a teacher's ability under the best circumstances. For example, in estimating the effect of evaluation probability $P^{P2}$ on the second principal evaluation score $Y^{P2}$, we use the scores from the first principal evaluation, $Y^{P1}$, as a control.

A main identifying concern is that the probability of an evaluation and the timing within the school year will be correlated, and therefore our analysis is capturing the

7. In the 2009–10 school-year, we also add the first external evaluation since it was announced.

**Table 3.** Treatment Summary by Year

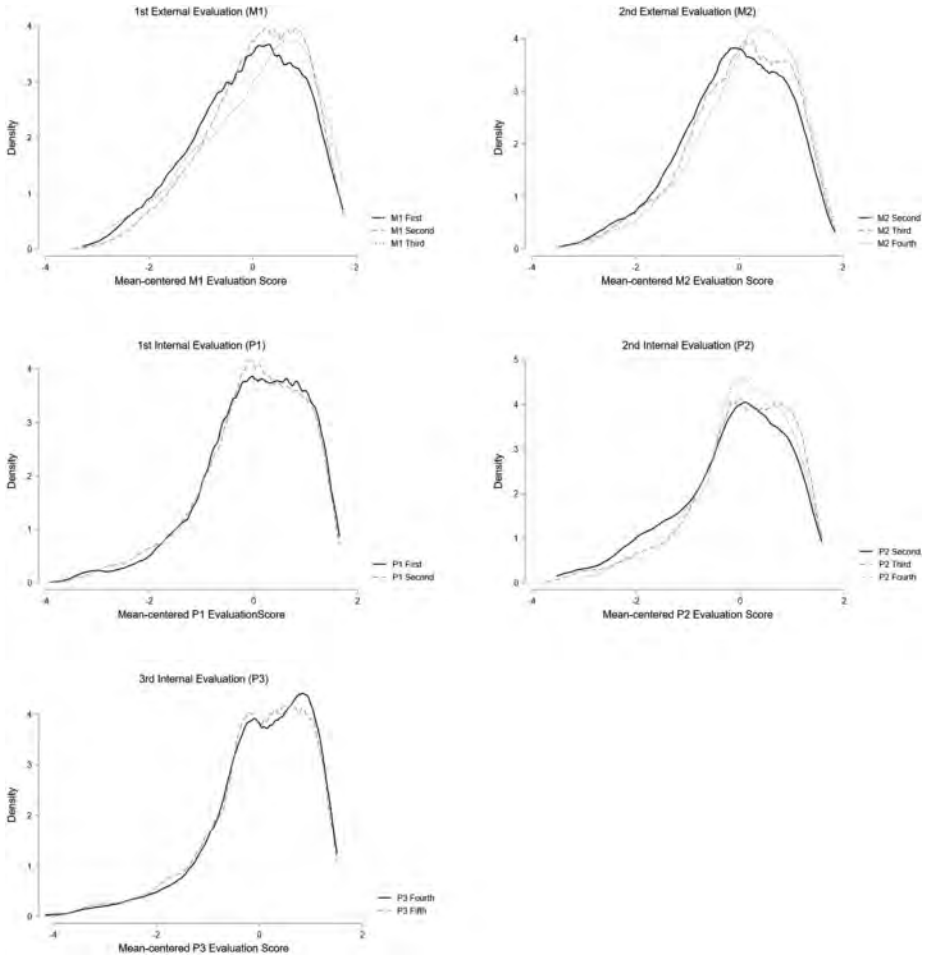
| | Probability of Any Evaluation on Day of Evaluation | | | | |
|---|---|---|---|---|---|
| | Internal 1 | Internal 2 | Internal 3 | External 1 | External 2 |
| 2009–10 | | | | | |
| Mean | 0.156 | 0.117 | 0.139 | 0.064 | 0.068 |
| Standard deviation | 0.140 | 0.114 | 0.141 | 0.079 | 0.077 |
| Min | 0.004 | 0.003 | 0.002 | 0.002 | 0.002 |
| Median | 0.113 | 0.087 | 0.096 | 0.043 | 0.044 |
| 95th percentile | 0.419 | 0.325 | 0.398 | 0.179 | 0.203 |
| 2010–11 | | | | | |
| Mean | 0.143 | 0.186 | 0.129 | 0.059 | 0.065 |
| Standard deviation | 0.127 | 0.189 | 0.127 | 0.061 | 0.068 |
| Min | 0.005 | 0.003 | 0.001 | 0.003 | 0.003 |
| Median | 0.108 | 0.124 | 0.089 | 0.040 | 0.041 |
| 95th percentile | 0.368 | 0.619 | 0.348 | 0.173 | 0.196 |
| 2011–12 | | | | | |
| Mean | 0.128 | 0.167 | 0.156 | 0.052 | 0.070 |
| Standard deviation | 0.126 | 0.159 | 0.157 | 0.049 | 0.076 |
| Min | 0.002 | 0.003 | 0.001 | 0.002 | 0.004 |
| Median | 0.092 | 0.125 | 0.108 | 0.039 | 0.045 |
| 95th percentile | 0.341 | 0.477 | 0.438 | 0.147 | 0.209 |
| Total | | | | | |
| Mean | 0.142 | 0.158 | 0.141 | 0.058 | 0.068 |
| Standard deviation | 0.132 | 0.161 | 0.142 | 0.064 | 0.074 |
| Min | 0.002 | 0.003 | 0.001 | 0.002 | 0.002 |
| Median | 0.103 | 0.110 | 0.097 | 0.040 | 0.043 |
| 95th percentile | 0.377 | 0.482 | 0.404 | 0.166 | 0.201 |

*Notes:* Treatment is the probability of being evaluated by either a principal or master educator on the day of an evaluation. Treatment levels are higher for principal evaluations because these are clustered in the last third of the evaluation window.

effects of receiving an evaluation later rather than the effect of evaluation probability. One method for assessing this possibility is to conduct the same analysis on the announced evaluations as the unannounced evaluations. To this end, we also estimate the effect of evaluation probability for announced evaluations.

The outcome variable of interest is the standardized evaluation score, $Y_{ij}^{k}$, for teacher $i$ in year $j$ on evaluation $k$, where $k$ is $P1$, $P2$, or $P3$ for internal (principal) evaluations and $M1$ or $M2$ for external evaluations. Evaluation scores are standardized within year. We control for school-level characteristics using school fixed effects $\phi_s$.

One complication in estimating treatment effects on teachers is how to appropriately control for experience (see, e.g., Taylor and Tyler 2012). It is common in the literature to use a quadratic form, with experience capped at fifteen or twenty years, or to use experience-level fixed-effects for each year of experience. Our results are unaffected by the specification of experience, but, given the richness of our data, we have opted to use experience level fixed-effects. We use $X_{it}$ to be a vector of experience-level indicators and $N_{ij}^{k}$ is the teacher's score on announced evaluations. Let $P_{ij}^{k}$ be the probability of any evaluation on the day of evaluation $k$, $T_{ik}^{k}$ be the day on which the evaluation was




*Notes:* Depicted are density estimates for each evaluation by its order for each teacher. Because evaluation windows between internal and external evaluators overlap, a teacher's first external evaluation can be her first evaluation overall, or it can be her second (if she has completed her first internal evaluation), or even her third evaluation if she has also completed her second internal evaluation. The score distributions appear to improve with each successive evaluation, except for the third internal evaluation.

**Figure 5.** Density Plots of Evaluation Scores Based on Evaluation Order

conducted, and $I_{ij}^{k,o}$ be an indicator of evaluation $k$'s order with $o \in \{1, 2, 3, 4, 5\}$. We then estimate a teacher's evaluation score with:

$$Y_{ijs}^{k} = \beta_{\circ} + \phi_s + \beta_{P^k} P_{ij}^{k} + \beta_X X_{ij} + \beta_{T^k} T^k + \sum_{o=1}^{5} \beta_{k,o} I_{ij}^{k,o} + \beta_N N_{ij} + \varepsilon_{ij}.$$

The errors are clustered at the school-by-year level. Because our outcome variable is standardized at the year level, the measured effects of $P^k$ are in units of standard deviations.

The coefficients of interest are the effect of evaluation timing, $\beta_{T^k}$, the effect of an evaluation's order, $\beta_{k,o}$, and the effect of an evaluation's probability, $\beta_{P^k}$. If the timing of an evaluation is not targeted at specific teachers, which we argue is true, then

the effects of evaluation timing, order, and probability will be separately and causally identified. This is possible because of variation between schools when evaluations are finished, and because of the overlapping evaluation windows. To depict why this is the case, consider two teachers A and B at different schools who receive their second principal evaluation on the exact same day. At teacher A's school, the principal had already conducted most of the evaluations, and so teacher A's probability was high relative to teacher B. Then teacher A provides a counterfactual for teacher B, providing an estimate of the effect of evaluation probability while holding the timing of evaluation constant.

One limitation of our data is that we are unable to reliably identify a teacher's grade and subject. Normally, we would seek to control for these potential differences in evaluation scores, but doing so is not possible. As a result, we conduct the analysis across all grades and subjects. Although this is a limitation, the evaluation rubric is designed with the explicit intention to be grade- and content-agnostic.

#### Treatment Exogeneity

Our key identifying assumption is that the timing of evaluations is independent of teacher characteristics that would affect their evaluation score, conditional on observable characteristics. Principals may want to conduct evaluations of their lowest-performing teachers first in order to provide feedback earlier in the year, which would bias our results upward since evaluation probability is low in the beginning of each window. If evaluators target weaker teachers early using information we cannot observe, our identification assumption is invalid.

To test for treatment selection bias, we regress characteristics that are observed by principals and district evaluators on our treatment variables, $P^k$. We estimate the following regression on the probability of each evaluation for each observable characteristic $X_{ij}$ for teacher $i$ in year $j$ at school $s$ :

$$P^k_{ijs} = \beta_0 + \phi_{sj} + \beta X_{ijs} + \varepsilon_{ij}. \quad (2)$$

Because there may be school-by-year systematic differences in observation timing, $\phi_{sj}$ is a school-by-year fixed effect for school $s$ in year $j$. The observable characteristics we consider are teacher value-added scores in reading and math in the previous year, an indicator for first-year teachers, the final IMPACT rating a teacher received the previous year, and the final evaluation rating a teacher received in the previous year.

Table 4 shows the results of the exogeneity checks specified by equation 2, with the treatment variables ($P^k_{ijs}$) across the columns and the observable characteristics evaluators may use to target teachers ($X_{ij}$) in each row. The cells are the coefficient $\beta$ in equation 2. The statistical significance shown has not been adjusted for multiple hypothesis testing.

The most important conclusion from table 4 is that any potentially significant characteristics have effects in a direction that would bias our results downward. For example, if principals target their second evaluation toward teachers with a Highly Effective IMPACT rating from the previous year, then Highly Effective teachers will have lower evaluation probability, reducing any positive effect we observe from evaluation probability. No evaluations are timed relative to whether a teacher is under the threat of

**Table 4.** Checks for Treatment Exogeneity

| | Evaluation Probability on Day of Evaluation | | | | |
|---|---|---|---|---|---|
| | Internal 1 | Internal 2 | Internal 3 | External 1 | External 2 |
| Previous reading VA | 0.0043 | 0.0110 | 0.0020 | 0.0007 | −0.0022 |
| | (0.006) | (0.008) | (0.010) | (0.004) | (0.004) |
| Previous math VA | −0.0044 | −0.0008 | −0.0138 | −0.0023 | 0.0003 |
| | (0.007) | (0.009) | (0.009) | (0.003) | (0.004) |
| First-year | −0.0006 | 0.0077 | −0.0022 | 0.0023 | −0.0002 |
| | (0.004) | (0.006) | (0.005) | (0.002) | (0.002) |
| IMPACT ME last year | 0.0040 | −0.0011 | 0.0002 | −0.0016 | 0.0002 |
| | (0.006) | (0.007) | (0.006) | (0.002) | (0.003) |
| IMPACT HE last year | −0.0050 | −0.0152** | −0.0058 | 0.0007 | −0.0026 |
| | (0.005) | (0.007) | (0.006) | (0.002) | (0.004) |
| Eval ME last year | 0.0065 | 0.0058 | 0.0070 | 0.0004 | 0.0048 |
| | (0.006) | (0.007) | (0.006) | (0.002) | (0.004) |
| Eval HE last year | 0.0026 | 0.0039 | 0.0001 | 0.0003 | 0.0025 |
| | (0.003) | (0.005) | (0.004) | (0.002) | (0.002) |

*Notes*: Each cell represents the estimated correlation between the row label (observable characteristic) and the column label (treatment variable). Standard errors are in parentheses. All variables except first-year are centered around the school mean to control for school fixed effects. Errors are clustered at the school-by-year level. Significance levels are not adjusted for multiple hypothesis testing. VA = value added; ME = Minimally Effective; HE = Highly Effective.

**Significant at the 5 percent level.

dismissal, nor are they correlated with a teacher's evaluation score being Minimally Effective or Highly Effective in the previous year.

Our exogeneity tests cannot be exhaustive because we are concerned with characteristics observed by evaluators but not observed in the data. However, our checks support our identifying assumption by showing that on a variety of observable characteristics known to correlate with teacher quality, evaluators are not systematically targeting weaker teachers early in the window.

## 5. RESULTS

Our main results are presented in table 5 for announced evaluations and table 6 for unannounced evaluations. For each evaluation we show the effect of increasing the likelihood of an evaluation ($P^k$) on the score for that evaluation, measured in standard deviations. We also show the effect of having an evaluation occur later in the evaluation window ($T^k$) and the effect that evaluation order has on teacher performance. Importantly, table 5 only includes our measure of the probability of an evaluation as a placebo test; because these evaluations are announced in advance, teachers actually know exactly when the evaluation is to occur. Also note that in the first year, 2009–10, the first external evaluation ($M1$) was announced.

The coefficients on probability represent the effect of an increase in probability by one full unit, which is the difference between a zero percent likely evaluation and a 100 percent likely evaluation. For example, the interpretation of the coefficient in column 2, row 1 of table 6 is that a teacher who is certain she will be evaluated improves her score by 0.309 standard deviations on $P2$ over a teacher who does not expect to be

**Table 5.** Effect of Evaluation Probability on Announced Evaluations

| | Announced | |
|---|---|---|
| | Internal Evaluation 1 All Years Score in Standard Deviations | External Evaluation 1 2010 Score in Standard Deviations |
| Evaluation Probability ($P^k$) | 0.146 | 0.162 |
| | (0.107) | (0.352) |
| Day within window ($T^k$) | −0.001 | −0.001 |
| | (0.002) | (0.002) |
| Evaluation is first | Reference | Reference |
| Evaluation is second | 0.114*** | 0.092 |
| | (0.020) | (0.066) |
| Evaluation is third | | 0.173 |
| | | (0.136) |
| *N* | 9,476 | 3,031 |

*Notes:* As a placebo test, this table shows estimates of how a teacher's in-class evaluation score improves as an announced evaluation becomes more likely. We estimate evaluation probability using an approximation of how many teachers will be evaluated at each school on each day divided by the number of teachers remaining to be evaluated. Standard errors are shown in parentheses. Errors are clustered at the school-by-year level. These results confirm our expectation that the probability of an evaluation should not affect an announced evaluation score.

***Significant at the 1 percent level.

evaluated.[8] Compared with a teacher who has no expectation of an evaluation, a teacher with an evaluation likelihood of 0.48, which is the 95th percentile, improves her second principal evaluation score by 0.15 standard deviations ($0.31 \times 0.482 = 0.15$). This is an improvement of about 0.09 points on the TLF scale.[9] To put this effect size in perspective, 7.5 percent of all teachers were within 0.09 points of one of the three rating thresholds on their internal evaluations. For the second external evaluation ($M2$), the estimate is 0.15 standard deviations or 0.09 points on the TLF scale, a difference that would have meant a change in that evaluation's for 7.9 percent of all teachers. In all, 17 percent of teachers had at least one evaluation score that was near enough to the threshold such that changes in the likelihood of observation could have changed their rating between Ineffective, Minimally Effective, Effective, and Highly Effective.

Our results show that evaluation probability has a statistically significant and positive effect on teachers' evaluation scores for all unannounced evaluations. In general, these effects do not constitute particularly large improvements in evaluation score, but one in six teachers was near enough to a threshold on an evaluation such that changes in likelihood could have led to a different rating outcome. Qualitatively, our results demonstrate that teachers prepare for their evaluations as they become more likely.

We consider how the order of an evaluation affects a teacher's performance. We find that teachers consistently make substantive evaluation improvements as they experience more evaluations. On average, our results show that a teacher improves her score between 0.04 and 0.15 standard deviations with each evaluation. To support this empirical result, figure 5 depicts density plots of evaluation scores by their order. For

8. Although we have assumed linear effects, our results are not qualitatively different when using a log specification.
9. From table 2, one standard deviation is 0.6 points on the TLF scale, making the calculation $0.15 \times 0.6 = 0.09$.

**Table 6.** Effect of Evaluation Probability on Unannounced Evaluations

| | Unannounced | | | |
|---|---|---|---|---|
| | External Evaluation 1 2011–2012 Score in Std Devs | Internal Evaluation 2 All Years Score in Std Devs | External Evaluation 2 All Years Score in Std Devs | Internal Evaluation 3 All Years Score in Std Devs |
| Evaluation Probability ($P^k$) | 0.406* | 0.309*** | 0.738*** | 0.173** |
| | (0.240) | (0.078) | (0.169) | (0.072) |
| Day within window ($T^k$) | −0.002* | 0.000 | 0.002* | 0.000 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Evaluation is 1st | Reference | | | |
| Evaluation is 2nd | 0.065 | Reference | | |
| | (0.046) | | | |
| Evaluation is 3rd | 0.187* | 0.152** | Reference | |
| | (0.100) | (0.062) | | |
| Evaluation is 4th | | 0.188*** | 0.116*** | Reference |
| | | (0.069) | (0.040) | |
| Evaluation is 5th | | | 0.162*** | 0.025 |
| | | | (0.052) | (0.022) |
| *N* | 6,439 | 8,667 | 9,070 | 9,141 |

*Notes:* This table shows estimates of how a teacher's in-class evaluation score improves as an unannounced evaluation becomes more likely. We estimate evaluation probability using an approximation of how many teachers will be evaluated at each school on each day divided by the number of teachers remaining to be evaluated. Standard errors are shown in parentheses. Errors are clustered at the school-by-year level. Results are arranged in the approximate order in which evaluations are conducted each year. Std Devs = standard deviations.

***Significant at the 1 percent level; **significant at the 5 percent level; *significant at the 10 percent level.

most evaluations, there appears to be an upward shift in scores from the lower tails as the order increases. The largest exception appears to be the last internal evaluation, for which the order does not appear to have any meaningful impact on a teacher's score. This is confirmed in column 4, row 7 of table 6.

Lastly, our results do not support the notion that having more time with a class improves a teacher's evaluation score. We find little evidence that teachers make improvements as the year progresses that are not the result of experiencing more evaluations or other preparations the teacher may make as her evaluation becomes likely. Although in some cases the estimated improvements from having more time with a class is statistically significant at the 10 percent level, its size is considerably small. This result is by no means robust to most other specifications, and as a result, we consider the statistical significance to be spurious.

A subtle yet substantial finding from our analysis is that improvements across evaluations are persistent. Across the first three columns of table 6, as evaluation order increases, we never observe a decrease in the effect of evaluation order. That is, for the second internal evaluation, a teacher does better if it is her third evaluation than had it been her second, and she improves even more if it is her fourth evaluation. We interpret this as evidence that improvements are cumulative.

### Robustness Checks

Our analysis attempts to identify within-year modifications to teacher practice. An inherent complication of this approach is that there are seasonal events that may affect a teacher's ability to perform well on an evaluation. For example, a teacher may engage in

**Table 7.** Sample Robustness Checks for Evaluation Probability

| | Unannounced Evaluations (Eval) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | External Eval 1 2011–12 | Difference | Internal Eval 2 All Years | Difference | External Eval 2 All Years | Difference | Internal Eval 3 All Years | Difference |
| Full sample | 0.406* | | 0.309*** | | 0.738*** | | 0.173** | |
| | (0.240) | | (0.078) | | (0.169) | | (0.072) | |
| Drop bottom 5 percent of treatment | 0.461* | 0.055 | 0.288*** | −0.021 | 0.553*** | −0.185 | 0.075 | −0.098 |
| | (0.239) | | (0.080) | | (0.161) | | (0.075) | |
| Drop bottom 10 percent of treatment | 0.509** | 0.103 | 0.277*** | −0.032 | 0.535*** | −0.203 | 0.072 | −0.101 |
| | (0.238) | | (0.083) | | (0.163) | | (0.075) | |
| Drop top 5 percent of treatment | 0.460 | 0.054 | 0.318*** | 0.009 | 0.799*** | 0.061 | 0.342** | 0.169 |
| | (0.500) | | (0.109) | | (0.296) | | (0.141) | |
| Drop top 10 percent of treatment | 0.768 | 0.362 | 0.418*** | 0.109 | 0.528 | −0.21 | 0.298* | 0.125 |
| | (0.648) | | (0.151) | | (0.397) | | (0.169) | |
| Drop top 5 percent and bottom 5 percent | 0.455 | 0.049 | 0.319*** | 0.01 | 0.717** | −0.021 | 0.271* | 0.098 |
| | (0.508) | | (0.116) | | (0.309) | | (0.145) | |

*Notes:* This table shows estimates for the effect of the likelihood of an in-class observation on teacher evaluation scores by different samples. To assess whether extremely low-probability or high-probability evaluation days are driving our main results, we drop observations based on their treatment size. The "Difference" column calculates the difference between the Full Sample and the subsample. Overall, these results suggest that there are nonlinearities in how evaluation probability affects teacher scores, though changing the treated sample does not qualitatively alter our results. Errors are clustered at the school-by-year level.

***Significant at the 1 percent level; **significant at the 5 percent level; *significant at the 10 percent level.

activities to celebrate Thanksgiving or Halloween. It would be unreasonable to expect her to score equally well during some of these seasonal events. Indeed, looking again at the first panel of figure 3, there is a noticeable dip in the frequency of district-led evaluations around the time of Thanksgiving. While these natural dips in the probability of an evaluation are factored into our model, it is possible that being surprised on such a low-probability day may disproportionately skew our results.

To test for possibly outsized negative effects on low-probability evaluation days, we calculate our main specification and drop observations with the lowest evaluation probability. For completeness, we also look at dropping the highest evaluation probability observations. Table 7 shows the effect of evaluation probability on a teacher's evaluation score after dropping observations in the bottom 5 and 10 percent of treatment, the top 5 and 10 percent of treatment, and then both the top 5 percent and bottom 5 percent simultaneously. We also calculate the difference between the full sample and the subsample, though none of these differences is statistically significant.

Overall, table 7 confirms our main qualitative result. Excluding different treatment subgroups does not significantly alter our findings. An additional takeaway, however, is the potential nonlinearity of how evaluation probability affects a teacher's evaluation score. As we drop the upper end of treated teachers, the effect of evaluation probability increases in many cases. This suggests that these upper-end treatments were pulling the estimated effect down, and hence the marginal effect of evaluation probability is decreasing with more treatment. Similarly, when we drop the lowest treatment group, the linear estimate of probability's effect on evaluation score often decreases. Although this evidence is not conclusive, it would suggest the benefits of a possible evaluation are diminishing as teachers are exposed to that possibility for prolonged periods of time.

**Table 8.** Subgroup Analysis for Effect of Evaluation Probability on Announced Evaluations

| | Announced | | | | | | |
|---|---|---|---|---|---|---|---|
| | Experience | | Previous Year Quality | | | Incentive Group | |
| Internal Evaluation 1 | First-Year | Veteran | Minimally Effective | Effective | Highly Effective | Tested | Non-Tested |
| Evaluation Probability ($P^k$) | 0.282 | 0.135 | 0.512 | 0.189 | 0.120 | 0.369 | 0.144 |
| | (0.263) | (0.107) | (0.492) | (0.140) | (0.151) | (0.236) | (0.132) |
| Day within evaluation window | 0.001 | −0.001 | −0.002 | 0.001 | 0.002 | 0.000 | −0.002 |
| | (0.003) | (0.002) | (0.005) | (0.002) | (0.002) | (0.003) | (0.002) |
| Evaluation is 2nd | 0.104* | 0.114*** | 0.166* | 0.078** | 0.121*** | 0.076* | 0.122*** |
| | (0.057) | (0.020) | (0.097) | (0.033) | (0.033) | (0.045) | (0.025) |
| *N* | 897 | 8,579 | 470 | 3,569 | 2,366 | 1,345 | 5,553 |

*Notes:* This table provides the same estimates as in table 5 broken out by subgroup. Errors are clustered at the school-by-year level.

***Significant at the 1 percent level; **significant at the 5 percent level; *significant at the 10 percent level.

### Heterogeneous Effects

To assess the heterogeneous effects of evaluation probability and order, we reestimate our model on specific subgroups, the results of which are in tables 8 and 9. In particular, we examine differences between first-year teachers and veteran teachers, previous year teacher quality (Minimally Effective, Effective, or Highly Effective), and the incentive group for a teacher. Incentive groups are based on which grades and subjects have testable material. Tested subjects are reading and math in grades 4 through 8, and nontested subjects and grades are all remaining teachers. The IMPACT incentive structure for teachers in tested subjects and grades is different because individual teacher value added makes up 50 percent of their overall IMPACT score, reducing the overall importance of in-class evaluations.

One stark result for announced evaluations is that regardless of experience, teacher quality, or incentive group, teachers make meaningful improvements if their first principal evaluation is the second evaluation of the year. As shown in the bottom row of table 8, teachers improve their principal evaluation score by anywhere between 0.076 and 0.122 standard deviations if it is their second evaluation of the year instead of their first. None of the differences between groups is statistically significant, suggesting these improvements are largely unrelated to a teacher's experience, quality, or incentive group. The first row of each panel shows evaluation probability has no statistically significant effect on teachers of any subgroup for announced evaluations, as expected.

When comparing first-year teachers to veteran teachers, the most interesting contrast is the differences in how they improve with successive evaluations. Veteran teachers appear to make greater improvements on their second external evaluation when it occurs after their second or third principal evaluation relative to first-year teachers (see the last two rows of the third panel in table 9). There do not appear to be any consistent differences in how veteran teachers and first-year teachers respond to the increased likelihood of an evaluation, but this could be due to effect differences among veteran teachers.

Looking at heterogeneous effects by teacher quality among veteran teachers suffers from small cell sizes, therefore any patterns we observe are only suggestive, not

**Table 9.** Subgroup Analysis for Effect of Evaluation Probability on Announced Evaluations

| | Unannounced | | | | | | |
|---|---|---|---|---|---|---|---|
| | Experience | | Previous Year Quality | | | Incentive Group | |
| **External Evaluation 1 (2011–12)** | **First-Year** | **Veteran** | **Minimally Effective** | **Effective** | **Highly Effective** | **Tested** | **Non-Tested** |
| Evaluation probability ($P^k$) | 1.187 | 0.362 | 1.134 | 0.224 | 0.313 | 0.310 | 0.486 |
| | (1.042) | (0.232) | (0.952) | (0.318) | (0.339) | (0.790) | (0.301) |
| Day within evaluation window | −0.005 | −0.002* | 0.000 | −0.003** | −0.003* | −0.003 | −0.002 |
| | (0.004) | (0.001) | (0.004) | (0.001) | (0.001) | (0.003) | (0.001) |
| Evaluation is 2nd | 0.201 | 0.053 | 0.047 | 0.030 | 0.064 | 0.030 | 0.102* |
| | (0.145) | (0.045) | (0.160) | (0.059) | (0.061) | (0.123) | (0.052) |
| Evaluation is 3rd | 0.160 | 0.191* | 0.920* | −0.093 | 0.406*** | 0.090 | 0.217* |
| | (0.482) | (0.098) | (0.493) | (0.134) | (0.154) | (0.333) | (0.122) |
| *N* | 554 | 5,885 | 471 | 3,582 | 2,386 | 894 | 3,722 |
| **Internal Evaluation 2** | **First-Year** | **Veteran** | **Minimally Effective** | **Effective** | **Highly Effective** | **Tested** | **Non-Tested** |
| Evaluation probability ($P^k$) | 0.180 | 0.328*** | 0.403 | 0.298*** | 0.332*** | 0.562*** | 0.298*** |
| | (0.197) | (0.076) | (0.373) | (0.095) | (0.125) | (0.143) | (0.093) |
| Day within evaluation window | −0.001 | 0.000 | −0.004 | 0.000 | 0.001 | −0.002 | 0.001 |
| | (0.003) | (0.001) | (0.007) | (0.002) | (0.002) | (0.003) | (0.001) |
| Evaluation is 3rd | 0.151 | 0.149** | 1.042** | 0.105 | 0.160 | 0.204 | 0.194** |
| | (0.137) | (0.067) | (0.435) | (0.103) | (0.115) | (0.170) | (0.081) |
| Evaluation is 4th | 0.187 | 0.188** | 1.080** | 0.118 | 0.209 | 0.251 | 0.203** |
| | (0.157) | (0.074) | (0.447) | (0.113) | (0.128) | (0.184) | (0.091) |
| *N* | 819 | 7,848 | 457 | 3,491 | 2,114 | 1,283 | 5,180 |
| **External Evaluation 2** | **First-Year** | **Veteran** | **Minimally Effective** | **Effective** | **Highly Effective** | **Tested** | **Non-Tested** |
| Evaluation probability ($P^k$) | 0.949** | 0.718*** | 1.328* | 0.673** | 1.086*** | 0.471 | 0.910*** |
| | (0.457) | (0.178) | (0.737) | (0.260) | (0.313) | (0.425) | (0.241) |
| Day within evaluation window | −0.001 | 0.002*** | 0.003 | 0.003** | 0.002 | 0.001 | 0.000 |
| | (0.002) | (0.001) | (0.003) | (0.001) | (0.001) | (0.002) | (0.001) |
| Evaluation is 4th | 0.075 | 0.119*** | 0.102 | 0.147** | 0.102 | 0.021 | 0.077 |
| | (0.135) | (0.041) | (0.153) | (0.058) | (0.076) | (0.103) | (0.049) |
| Evaluation is 5th | 0.138 | 0.165*** | 0.279 | 0.239*** | 0.138 | 0.139 | 0.104 |
| | (0.173) | (0.054) | (0.219) | (0.080) | (0.100) | (0.143) | (0.064) |
| *N* | 872 | 8,198 | 459 | 3,526 | 2,163 | 1,294 | 5,334 |
| **Internal Evaluation 3** | **First-Year** | **Veteran** | **Minimally Effective** | **Effective** | **Highly Effective** | **Tested** | **Non-Tested** |
| Evaluation probability ($P^k$) | 0.362** | 0.162** | 0.479 | 0.184* | 0.028 | −0.006 | 0.265*** |
| | (0.181) | (0.073) | (0.307) | (0.098) | (0.139) | (0.169) | (0.087) |
| Day within evaluation window | 0.001 | 0.000 | 0.005 | 0.001 | 0.003** | −0.001 | 0.000 |
| | (0.003) | (0.001) | (0.005) | (0.002) | (0.002) | (0.002) | (0.001) |
| Evaluation is 5th | 0.028 | 0.023 | 0.380*** | −0.045 | 0.093** | 0.010 | 0.015 |
| | (0.072) | (0.023) | (0.122) | (0.035) | (0.039) | (0.054) | (0.027) |
| *N* | 883 | 8,258 | 459 | 3,514 | 2,156 | 1,318 | 5,385 |

*Notes:* This table provides the same estimates as in table 6 broken out by subgroup. Errors are clustered at the school-by-year level.

***Significant at the 1 percent level; **significant at the 5 percent level; *significant at the 10 percent level.

definitive. Teachers who received an Effective rating in the previous year often have the weakest response to evaluation probability when compared with Minimally Effective and Highly Effective teachers, though these differences are not statistically significant. The third panel of table 9, which are the results for the second external evaluation, show that a Minimally Effective teacher who is certain of an upcoming evaluation will

increase her score by 1.33 standard deviations, whereas an Effective teacher will increase her score by roughly half that amount. Highly Effective teachers also appear to improve their scores more than Effective teachers, though this effect is not present for the third principal evaluation. Similarly, Effective teachers appear to make fewer improvements with successive evaluations than Minimally or Highly Effective teachers, though this pattern is not strong (it holds for all unannounced evaluations except the second external evaluation). In the last row of the last panel of table 9, teachers who were Minimally Effective in the previous year are shown to improve their third principal evaluation by 0.380 standard deviations if it is their last evaluation, while Effective teachers have an insignificant negative effect.

One possible explanation for these differences in results by teacher quality is that teachers on the high end of the performance distribution are more capable of adjusting their teaching to align with the rubric practices as an evaluation becomes likely, compared with teachers in the mid-range of practice. Similarly, teachers at the low end of the performance distribution likely have many domains in which to improve and could use the guidance provided by the observation rubric to enact a new practice. In contrast, teachers in the middle of the performance distribution have likely taken up the practices from the rubric they are able to independently enact, but do not yet have the skills to further refine their practice without a more intensive development opportunity. We have no empirical way to assess this possibility but the story is consistent with our prior beliefs that skilled teachers should be more capable of making minor teaching adjustments to improve their score.

Another possible explanation for the different effect sizes between Minimally Effective, Effective, and Highly Effective teachers could be the differences in incentives. For a teacher who was Minimally Effective in the previous year, she must improve her current year scores or be dismissed, adding considerably more gravity to her evaluations. A teacher who was Highly Effective in the previous year has very large payoffs if she is Highly Effective again. These payoffs potentially include an annual bonus of up to $25,000 and a permanent pay increase, among other rewards. Therefore, these two groups of teachers may be more attentive to the probability of evaluation as the window unfolds.

We also find that teachers of non-tested grades and subjects (those without individual value added) appear to be more responsive to evaluation probability for their third and final principal evaluation than teachers of tested grades and subjects. In the first line and last two columns of the fourth panel in table 9, teachers of tested grades and subjects have no discernible response to evaluation probability. Given that effectively all of the last principal evaluations occur after students complete their standardized tests, it is plausible that teachers of tested grades and subjects are less concerned with their evaluation scores in the post-test months, especially given that evaluations constitute only 35 or 40 percent of their overall IMPACT score, as opposed to 75 percent for teachers who do not teach these grades and subjects.

## 6. DISCUSSION AND ADDITIONAL POLICY QUESTIONS

Previous studies of high-stakes teacher evaluation show it is possible to improve teacher practice through in-class observations. Our results reveal two key mechanisms through which improvements occur. Specifically, we show that as the probability of a classroom

observation increases, a teacher's measured performance on that evaluation increases as well. Teachers are cognizant of the instructional standards delineated in the observation rubric, and they take steps to improve their teaching practice as a result of probable but unannounced high-stakes observations. We also show that teachers make lasting improvements from one evaluation to the next. Our analysis is uniquely capable of disentangling these two effects by quantifying both the accountability and information mechanisms of high-stakes classroom observations.

Importantly, the changes that teachers make in preparation for an evaluation are in turn tied to improved student outcomes in other work done in this same setting (Phipps 2018). The structure of the five evaluation windows ensures that for most teachers, the majority of school days have some possibility of an unannounced evaluation. Although some of the behavioral responses to an impending observation may be transitory, their effects on students are still cumulative and positive.

In addition to our key findings, there are two policy questions related to the use of unannounced evaluations that our results touch on. Namely, does an evaluation rubric mechanically direct teacher time and attention toward specific rubric components to the detriment of holistic teaching improvements? Second, using external observers is costlier than using principals to conduct evaluations, but do external observers provide more objective evaluations or better feedback?

#### Improvements on Specific Rubric Components

With nine separately scored components on each evaluation, our basic multitasking model would predict that teachers will shift their attention toward components easiest to improve. The language describing how each component is scored varies in specificity, where some domains provide specific examples of teacher behaviors, while others have more general and vague language. Similarly, some practices are conceivably more difficult to adjust within a few days and require consistent development over weeks and months.[10] We measure how teachers may adjust their performance on individual rubric components using the same specification but with the dependent variable changed to each of the nine Teach components. The results of this analysis are shown in table 10.

As found in other studies (Adnot 2016), these rubric components are highly correlated, making our hypotheses on observation components highly dependent. We have adjusted the significance levels using a Bonferroni correction because it is very likely that an improvement in one rubric component will also lead to an improvement in another.[11] Our main purpose is to identify whether there is a single rubric component that dominates teacher improvements when an evaluation becomes likely.

We find that Teach 3, Teach 6, and Teach 8, which are broken out in table 10, are each significant in at least two evaluations. The fact that Teach 8 is consistently an area of improvement fits with our prior expectations (see online table A.1 for a full description of how each component is evaluated). Teach 8 is meant to evaluate classroom routines,

10. See table A.1 for a complete description of the rubric elements, available in a separate online appendix that can be accessed on *Education Finance and Policy*'s Web site at https://doi.org/10.1162/edfp_a_00295. The language in table A.1 is the exact language provided to teachers and evaluators.

11. For a hypothesis threshold $\alpha$, the adjusted threshold for significance across $m$ hypotheses is $\alpha^* = \frac{\alpha}{m}$. In this case, because each evaluation has nine components and we suspect they are highly correlated, then $\alpha^* = \frac{\alpha}{9}$. See Dunnett (1955) for more detail.

**Table 10.** Effect of Evaluation Probability on Individual Evaluation Components

| | Effect of Evaluation Probability on Component Score (Standard Deviations) | | | |
|---|---|---|---|---|
| | M1 | P2 | M2 | P3 |
| Teach 1 | 0.185 | 0.151 | 0.717*** | 0.157 |
| | (0.261) | (0.107) | (0.188) | (0.094) |
| Teach 2 | 0.272 | 0.097 | 0.613** | 0.077 |
| | (0.279) | (0.070) | (0.187) | (0.082) |
| Teach 3 | 0.267 | 0.511*** | 0.593*** | 0.131 |
| | (0.254) | (0.092) | (0.171) | (0.090) |
| Teach 4 | 0.337 | 0.183 | 0.745*** | 0.087 |
| | (0.270) | (0.097) | (0.183) | (0.084) |
| Teach 5 | 0.478 | 0.063 | 0.575** | 0.112 |
| | (0.284) | (0.101) | (0.200) | (0.088) |
| Teach 6 | 0.036 | 0.312*** | 0.303 | 0.321** |
| | (0.513) | (0.086) | (0.249) | (0.112) |
| Teach 7 | 0.228 | 0.269** | 0.409 | 0.194 |
| | (0.259) | (0.095) | (0.191) | (0.085) |
| Teach 8 | 0.096 | 0.395*** | 0.600*** | 0.165 |
| | (0.262) | (0.077) | (0.159) | (0.084) |
| Teach 9 | 0.152 | 0.078 | 0.305 | 0.159 |
| | (0.237) | (0.076) | (0.174) | (0.090) |
| *N* | 6,520 | 8,776 | 9,162 | 9,228 |

*Notes:* Coefficients are the effect of increasing the likelihood of an in-class evaluation on the specific rubric component (called Teach components). For a complete description of the Teach elements, please see online table A.1. All significance levels have been adjusted using a Bonferroni correction factor. P2 and P3 are internal evaluations (administered by principal or assistant principal), while M1 and M2 are external evaluations (administered by a district employee called a master educator). Results for M1 are for the 2010–11 and 2011–12 years only. Errors are clustered at the school-by-year level.

***Significant at the 1 percent level; **significant at the 5 percent level.

procedures, and behavior management. Whereas routines and procedures in a classroom are built over time and must be in place for students to respond appropriately, it is possible to pay particular attention to this construct in planning for a given day to obtain a higher score. For example, a teacher who ordinarily allows students to work on a nonacademic project after they have completed the lesson may plan to provide students with a more academically focused activity to satisfy the "idleness" component. Similarly, a teacher may use additional patience and de-escalation techniques when dealing with inappropriate or off-task student behavior to ensure it is efficiently addressed, or even spend additional time during the week preparing a challenging student for an evaluation. A teacher may also plan to pay additional attention to giving instructions before a class transition to ensure minimal prompting, whereas in the absence of an anticipated observation the teacher may have been comfortable relying on prompts to redirect student behavior. These components would not show up as significant here if they were easily adjustable as the lesson unfolds, as the probability of evaluation would then have no influence. Instead, our results suggest that teachers are preparing with respect to these components specifically.

Although those three components stand out in a test of statistical significance, they do not represent a major substantive difference relative to other TLF components. Instead, our results show that teachers improve their teaching practice across multiple

desired dimensions in preparation for a possible evaluation, instead of prioritizing particular dimensions of practice. The effect of evaluation probability on other TLF components is statistically significant and larger in magnitude in different specifications, but there are no clearly observable patterns. We interpret this as evidence that teachers do not simply select a few practices for improving their evaluation score but rather make improvements across a variety of domains in preparation for an evaluation.

As further evidence that teacher responses to an increasingly likely evaluation are not excessively focused on a single dimension, we use a multivariate regression to test whether evaluation probability has statistically significantly different effects across all nine components. In this test, we fail to reject the null hypothesis that there is heterogeneity in effect size across Teach components. This highlights the statistically significant effects seen for Teach 3, 6, and 8 are driven by the fact that these coefficients are more precisely estimated and not necessarily larger in magnitude. Our results are consistent with those of Adnot (2016), who found that the majority of variation in evaluation scores in the IMPACT program can be explained with a single factor that encompasses all nine of the evaluation components.

#### Principals as Evaluators

Our results have secondary implications about the choice of evaluator. The literature is increasingly concerned with who should conduct evaluations, particularly in high-stakes environments. Our main results in table 6 show that principal evaluations are more inert than external evaluators. The overall effects for veteran teachers are statistically significantly different between observations for internal and external raters. While we are unable to determine why teacher scores are less responsive to evaluation probability when principals are the evaluators, we consider two possible explanations.

The broader literature shows that principals compress the distribution of evaluation scores, suggesting that they do not identify as much nuance in teaching or they are less willing to make errors with strong consequences for teachers. If this is the case, evaluation probability should have a smaller observed effect on scores when principals conduct the evaluation, as we find. As further evidence that principals compress the distribution of evaluation scores, we use Levene's test of homogeneity and confirm that there is a statistically significant difference between the variance of principal evaluations and the variance of external evaluations.

Our results also suggest a separate hypothesis: Principals may incorporate additional information about students or the teacher in their evaluation to which external evaluators do not have access. This could entail the consideration of teacher characteristics like collegiality, for example, or simply prior teaching performance. Although our approach does not allow for us to distinguish between the two hypotheses, we raise the question for future research. To the extent that these nuanced teaching improvements enhance student achievement, should evaluation systems use external evaluators to ensure that teaching practices are accurately reflected in evaluation scores?

#### Other Policy-Relevant Issues

It is costlier to use external classroom observers, so it is policy-relevant to know which evaluators provide better feedback. We can address this question by considering how

much teachers improve on subsequent evaluations after receiving feedback from internal and external evaluators. Perhaps surprisingly, our analysis shows that there is no discernible difference between internal and external evaluations in terms of how teachers improve post-evaluation. The order effects shown in the first column of table 6 show improvements made to an external evaluation as the result of experiencing more internal evaluations prior to an external evaluation. If the first external evaluation represents the second or third observation, this means that the teacher already completed her first or second internal evaluations. Similarly, for the second column, a teacher's second internal evaluation is third or fourth if it comes after her first or second external evaluation. The order effects across evaluations are hardly different (except for the last principal evaluation) and are not significantly different. As a result, there is no evidence that external observations lead to better or worse improvements on future evaluations, despite teachers' increased responsiveness to them.

A final remaining policy-relevant question is whether evaluations should be announced. Although our analysis does not capture all of the relevant dimensions of this question, such as how one policy affects teacher morale over another, it can speak to the policy effects on teacher behavior as measured by evaluations. First, our evidence shows that teachers make improvements to their practice when they suspect they will be evaluated. However, we acknowledge that DC is unique in its use of five annual observations for all teachers, and it is difficult to predict how these effects might vary given additional observations per year. Second, although teachers still appear to make improvements following an announced evaluation, they do not appear to be as large as improvements following an unannounced evaluation (see the second column, fourth row, of table 5 and the first column, fourth row, of table 6). However, these differences are slight. Taken together, our evidence suggests that policy makers considering teacher evaluation as a route to improved practice should consider the benefits of unannounced classroom observation. Of course, policy makers must also be attentive to the implementation of policy when hoping to change teacher behavior. Our work does not address teacher perceptions of unannounced evaluations, and teacher buy-in that may be required to influence subsequent teacher decision-making. DCPS intentionally included an announced observation at the beginning of each year, and policy makers should think critically about how evaluation components like this improve teacher perceptions of the program.

## 7. CONCLUSION

The rapid growth of teacher evaluation programs has progressed with little evidence on how these programs affect daily teacher practice. Whereas seminal work has shown that these programs have the potential to improve student outcomes and increase teacher ratings, no work has revealed how those outcomes are achieved. Our results demonstrate that, in addition to improving from one evaluation to the next, teachers improve their teaching practice along multiple dimensions when a classroom observation is more likely.

Our results highlight the need to ensure that evaluation encourages desired behavior via specific rubric constructs and language. To this end, DCPS has continued to revise and improve their evaluation rubric, moving to a more conceptual teaching

framework for the 2016–17 school year. Our analysis suggests that other districts should follow suit, reflecting on the ways in which rubrics for classroom observation reflect the desired teacher response. The results also caution against evaluation systems that do not use standards-based observation rubrics, which are unlikely to provide the needed guidance to change teaching practice.

The observed difference between external and internal evaluations emphasizes the need to understand why principals as observers are different than outside observers in terms of how teachers respond to the evaluator. We are unable to clearly establish the cause of the difference in evaluation between external evaluators and principals in DCPS. This is particularly salient as DCPS recently stopped using external evaluators in part to reallocate funds for a greatly expanded professional development program, employing school-based coaches to lead collaborative, content-specific learning teams.

Ultimately, the goal of the evaluation system in Washington, DC, is to improve teaching in order to improve student outcomes. Our analysis provides useful insights into how unannounced evaluations affect teacher behavior. The classroom observation component of teacher evaluation systems has the potential to improve teacher output by prioritizing behaviors in a production process known to be difficult and uncertain. Teachers cannot always know how a particular approach or style will affect students relative to another approach. A possible solution is to use structured in-class evaluations to reduce teacher uncertainty about their daily practice. In this framework, observations play an important role in guiding teaching priorities, supporting alignment to these priorities, and rewarding effective teaching practices. Our results suggest teachers will enact rubric practices for impending but unannounced classroom observations in ways that ultimately benefit students. Teacher evaluation systems without this observation component, then, may not be as effective.

#### ACKNOWLEDGMENTS

Special thanks to Sarah Turner, James Wyckoff, Julie Cohen, William Johnson, Leora Friedberg, and Amalia Miller for helpful comments and direction. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grants #R305B140026 and #R305H140002 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute of Education Sciences, the U.S. Department of Education, the United States Military Academy, the Department of the Army, or the Department of Defense.

#### 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 dissertation, University of Virginia, Charlottesville, VA.

Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. Teacher turnover, teacher quality, and student achievement in DCPS. *Educational Evaluation and Policy Analysis* 39(1): 54–76.

Briggs, Derek, Elena DiazBbilello, Andrew Maul, Michael Turner, and Charles Bibilos. 2014. Denver ProComp evaluation report: 2010-2012. Available https://www.colorado.edu/cadre/2017/08/25/denver-procomp-evaluation-report-2010-2012. Accessed 27 July 2020.

Cohen, Julie, and Dan Goldhaber. 2016. Observations on evaluating teacher performance. In *Improving teacher evaluation systems: Making the most of multiple measures*, edited by Jason Grissom and Peter Youngs, pp. 8–21. New York: Teachers College Press.

Cullen, Julie Berry, Cory Koedel, and Eric Parsons. 2016. The compositional effect of rigorous teacher evaluation on workforce quality. NBER Working Paper No. 22805.

Dee, Thomas, and Benjamin Keys. 2004. Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management* 23(3): 471–488.

Dee, Thomas, and James Wyckoff. 2015. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34(2): 1–31.

Dunnett, Charles W. 1955. A multiple comparisons procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 50(272): 1096–1121.

Fryer, Roland G. 2013. Teacher incentives and student achievement: Evidence from New York City Public Schools. *Journal of Labor Economics* 31(2): 373–407.

Grissom, Jason, and Susanna Loeb. 2017. Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low- and high-stakes environments. *Education Finance and Policy* 12(3): 369–395.

Hanushek, Eric A. 2011. The economic value of higher teacher quality. *Economics of Education Review* 30(3): 466–479.

Holmstrom, Bengt, and Paul Milgrom. 1991. Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization* 7:24–52.

Kane, Thomas, and Douglas Staiger. 2012. Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, Thomas, Eric Taylor, John Tyler, and Amy Wooten. 2010. Identifying effective classroom practices using student achievement data. *Journal of Human Resources* 6(46): 587–615.

Kraft, Matthew A., and Allison F. Gilmour. 2016. Can principals promote teacher development as evaluators? A case study of principals views and experiences. *Educational Administration Quarterly* 52(5): 711–753.

Lazear, Edward, and Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89(5): 841–864.

Papay, John P. 2012. Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review* 82(1): 123–141.

Phipps, Aaron. 2018. Incentive contracts in complex environments: Theory and evidence on effective teacher performance incentives. PhD dissertation, University of Virginia, Charlottesville, VA.

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan F. Burns, Laura S. Hamilton, and Brian Stecher. 2012. Team pay for performance: Experimental evidence from the round rock pilot project on team incentives. *Educational Evaluation and Policy Analysis* 34(4): 367–390.

Springer, Matthew G., Walker Swain, and Luis Rodriguez. 2016. Effective teacher retention bonuses: Evidence from Tennessee. *Educational Evaluation and Policy Analysis* 38(2): 199–221.

Steinberg, Matthew, and Morgaen 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, and Lauren Sartain. 2015. Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy* 10(4): 535–572.

Taylor, Eric, and John Tyler. 2012. The effect of evaluation on teacher performance. *American Economic Review* 102(7): 3628–3651.