Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 22-510

Does Monitoring Change Teacher Pedagogy and Student Outcomes?

Brian Phillips
United States Military Academy

In theory, monitoring can improve employee motivation and effort, particularly in settings lacking measurable outputs, but research assessing monitoring as a motivator is limited to laboratory settings. To address this gap, I leverage exogenous variation in the presence and intensity of teacher monitoring, in the form of unannounced in-class observations as part of D.C. Public Schools' IMPACT program. As monitoring intensifies, teachers use more individualized teaching and emphasize higher-level learning. When teachers are unmonitored, their students have lower test scores and increased suspensions. This novel evidence validates monitoring as a potential tool for enhancing teacher pedagogy and employee performance more broadly.

VERSION: August 2024

Suggested citation: Phipps, Aaron. (2024). Does Monitoring Change Teacher Pedagogy and Student Outcomes?. (EdWorkingPaper: 22-510). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7021-1x97

Does Monitoring Change Teacher Pedagogy and Student Outcomes?

Aaron Phipps*

Abstract

In theory, monitoring can improve employee motivation and effort, particularly in settings lacking measurable outputs, but research assessing monitoring as a motivator is limited to laboratory settings. To address this gap, I leverage exogenous variation in the presence and intensity of teacher monitoring, in the form of unannounced in-class observations as part of D.C. Public Schools' IMPACT program. As monitoring intensifies, teachers use more individualized teaching and emphasize higher-level learning. When teachers are unmonitored, their students have lower test scores and increased suspensions. This novel evidence validates monitoring as a potential tool for enhancing teacher pedagogy and employee performance more broadly.

JEL: J33, J41, J45, M52, M54, I21

Keywords: Labor Contracts, Job Performance, Compensation, Education Policy

*Assistant Professor of Economics, Department of Social Sciences, United States Military Academy. 607 Cullum Road, West Point, NY 10996. Email: aaron.phipps@westpoint.edu Phone: (845) 549-4697. Special thanks to Sarah Turner, William Johnson, James Wyckoff, and Leora Friedberg for their helpful comments and direction. Thanks are also due to the insightful comments from the Editor and reviewers; their comments have substantially improved the paper. Additional thanks to the many conference participants, presentation audiences, and conversation partners who engaged earlier versions of this paper. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 to the Rectors and Visitors of the University of Virginia. The opinions expressed are those of the author and do not represent views of the Institute, the U.S. Department of Education, the U.S. Department of Defense, the US Army, or the United States Military Academy.

I. Introduction

Despite research showing that incentives can significantly improve workplace productivity (e.g., Lazear, 2000), output-based pay remains rare in the U.S. (Prendergast, 1999; Lazear, 2018). This contrasts with extensive theoretical work demonstrating that tying compensation directly to measurable output is often the most efficient incentive structure (Holmstrom and Milgrom, 1991; Lazear and Rosen, 1981; Lazear, 1986). However, outputs can be hard to measure in the majority of occupations—particularly when institutional bureaucracy removes market signals about productivity (Prendergast, 2016). In these contexts, monitoring can be an effective solution when employers can recognize and articulate to workers the desired behaviors, leading many bureaucracies and government-owned industries to rely on monitoring to improve efficiency (Ferejohn and Shipan, 1990; Hammond and Knott, 1996; Gailmard, 2002). Unfortunately, despite some lab experiments suggesting that monitoring is the most effective incentive structure for improving effort and output (Nalbantian and Schotter, 1997; Camerer and Weber, 2012), there is little evidence on its effectiveness in real-world settings. Instead, most real-world studies of monitoring focus on preventing cheating, crime, and corruption, rather than increasing employee effort or output.

This study seeks to correct that imbalance by examining the impact of monitoring on teachers' behavior and output (as measured by their students' test scores) within IMPACT, an at-scale teacher incentive program in the Washington DC Public Schools (DCPS), launched in 2009. This context addresses the empirical challenge of studying the effects of monitoring in real-world settings. The ideal experiment would generate exogenous variation in monitoring with well-defined measures of behavior (inputs) and downstream outputs. The IMPACT program offers large incentives based on student performance on standardized tests and teacher pedagogy, as measured by several unannounced in-class observations. These observations occur in predefined, overlapping time windows, creating periods of unmonitored time that are exogenously determined.

Specifically, I use DCPS records from fourth and fifth grades to identify the effect of

additional unmonitored time on student outcomes. In both grades, the students (1) have a single teacher and (2) have current and previous-year standardized test scores. In summary, I find that unmonitored days in the month before standardized tests lead to a substantial decrease in student performance: on average, an additional day of unmonitored time corresponds to a reduction in student performance by about 0.01 standard deviations, roughly 3% of the effect of a single suspension day (Lacoe and Steinberg, 2019). A week of unmonitored time is equivalent to a 0.3 SD decrease in teacher value-added for math and a 0.4 SD decrease for English Language Arts (Chetty, Friedman and Rockoff, 2014). However, it is unclear how much this effect represents short-term losses (relative to monitored teachers) vs. substantive changes in long-term learning. What's more, comparing this effect size to other studies is difficult given the temporal proximity to standardized tests and the setting's within-year nature (Kraft, 2020). As a result, it is unlikely that these effects can be extended across the whole school year. Rather, they demonstrate substantial changes in teacher behavior and a remarkable pliability in student standardized test scores.

I reinforce these findings with additional robustness checks and an examination of possible underlying mechanisms. I estimate a tight null effect from a placebo treatment, by looking at the effects of unmonitored time occurring after students' standardized tests. In analyzing the possible mechanisms, I find much of the unmonitored time's effect is driven by increases in student disruptions and suspensions. As I show, these effects appear to be the result of a degradation in teacher classroom preparation and management.

Beyond contributing to the literature on monitoring, the results here provide insight for policy makers. Identifying an effective incentive design for teachers has proven difficult, yet doing so would help meet significant policy goals of recruiting, motivating, and retaining effective teachers (Dee and Wyckoff, 2015). Other papers have assessed IMPACT's overall effects as a program (Dee and Wyckoff, 2015; James and Wyckoff, 2020; Dee, James and Wyckoff, 2021), but my results do not address the equilibrium effects of IMPACT. Rather, this paper complements assessments of IMPACT and similar teacher evaluation policies by demonstrating the sizable effects monitoring can have on teacher performance and student outcomes.

II. Motivation and Related Literature

II.A Theories on Monitoring

Early monitoring theories focus on rational agents who seek to maximize their utility from shirking. These models, starting with Becker's work on crime (Becker, 1968), assume employees will cheat when the benefits outweigh the costs of being caught. Lazear (2006) extends this framework to education, exploring optimal monitoring conditions. He finds that when evaluations are infrequent or costly, as is often the case in teaching, revealing the evaluation criteria beforehand is optimal.

While rational agent models offer valuable insights, not all employees seek to exploit their employers. Nagin et al. (2002) introduce alternative theories, including the Conscience Model, where employees adopt identities incompatible with shirking (see also Akerlof, 1982). This model suggests that fostering a strong professional culture can be more effective than monitoring intensively for shirking. In education, teachers who strongly identify with their roles may exert more effort in preparing and delivering their lessons, potentially reducing the need for oversight. However, this effect might be diminished when school districts implement test-based incentives that externalize teachers' intrinsic motivation (Benabou and Tirole, 2003; Sliwka, 2007). These speculative issues have received little empirical attention.

Robust empirical literature supports monitoring's effectiveness in preventing corruption and crime, but its impact as a personnel strategy in real-world contexts remains understudied. Much of the existing evidence is limited to laboratory settings (see the review by Camerer and Weber, 2012). Some applied research has examined how monitoring reduces specific negative behaviors, such as cheating (Nagin et al., 2002) or teacher absenteeism in developing economies (Duflo, Hanna and Ryan, 2012). However, these studies focus on contexts where basic standards were not already being enforced. The crucial question that remains unanswered is whether monitoring can go beyond preventing negative behaviors to reinforce positive ones and improve overall employee performance.

II.B Teacher Incentives: Theoretical Considerations

Teacher incentive design faces unique challenges due to the complex nature of teaching and its desired outcomes. While much of the literature focuses on performance pay based on improvements in student test scores ("value-added"), teachers are expected to improve various student dimensions beyond standardized test performance, including behavioral aptitudes often termed "non-cognitive skills." These outcomes are difficult to measure and incentivize (Murnane and Cohen, 1986; Dixit, 2002), but recent work by Gilraine and Pope (2021) and Dinerstein and Opper (2022) finds that teachers' long-term impact on students is more strongly correlated with these non-cognitive outcomes than with short-term test performance. The multidimensional influence of good teaching complicates the design of effective incentive systems, as it is unclear which outcomes should be prioritized for bonus payments (Holmstrom and Milgrom, 1991). Moreover, value-added measures, though commonly used in performance pay systems, are inherently noisy, potentially muting teacher responses to such incentives (Lazear and Rosen, 1981).[1]

Beyond these measurement issues, it remains unclear whether outcomes-based incentives should be expected to have a strong effect. Teachers are often "motivated agents" who theoretically have more inelastic responses to incentives (see Dixit, 2002; Francois, 2000, for example). Many schools also approach teaching as a collaborative effort, complicating the relative weight of individual and team contributions (Imberman and Lovenheim, 2015; Fryer, 2013). Individual incentives – particularly those that are rank-based – could interfere with collaboration, while team-based incentives can create perverse incentives to free ride (Holmstrom, 1982; Kandel and Lazear, 1992).

Given these challenges, focusing solely on short-term student test performance may be less effective than incentivizing improvements in teacher pedagogy. Monitoring, particularly in the form of unannounced in-class observations, offers a potential tool to reach this goal. While schools have conducted such observations for decades, their

[1]These theoretical issues are covered more completely by Lazear and Oyer (2012) and Prendergast (1999). They also highlight how well the measured output aligns with the desired outcome (Akerlof and Kranton, 2005; Neal, 2011; Benabou, 2016), potential gaming of the outcome measure (Baker, 1992), and the use of subjective measures of employee output (Levin, 2003; MacLeod, 2003; Gibbs et al., 2004).

effectiveness as an incentive tool has been limited. The majority of existing evaluation systems lack rigor and are often non-binding, with most teachers receiving satisfactory scores regardless of performance (Weisberg et al., 2009). These shortcomings have spurred extensive research into more robust methods for measuring effective teaching (Kane and Cantrell, 2010; Kane and Staiger, 2012). Recent improvements in measuring effective teaching have made it more feasible to use in-class observations as incentives. These developments present an opportunity to transform traditional classroom observations from perfunctory exercises into tools for improving teaching quality.

II.C Teacher Incentives: Empirical Evidence

Empirical evidence on the effects of teacher-level performance incentive programs has been mixed, making it unclear which characteristics make these programs effective, though programs that have improved student performance (primarily measured by standardized test scores, but also by graduation rates, reading proficiency, school attendance, and other objective measures) often include in-class observations. The Career Ladder program in Tennessee, which awarded career advancement and bonuses based on in-class evaluations, showed improved student performance (Dee and Keys, 2004), though a causal interpretation is limited due to teacher self-selection. The DCPS IMPACT program (the setting of the present study) also improved student performance (Dee and Wyckoff, 2015), but the mechanisms likewise remain unclear. Conversely, some large performance pay programs lacking rigorous annual in-class teacher observations, like the Tennessee Project on Incentives in Teaching (Springer et al., 2012) and the Denver Professional Compensation program (Briggs et al., 2014), showed no measurable impact on student performance. Studies evaluating multiple programs simultaneously, such as the Minnesota Q-comp program (Sojourner, Mykerezi and West, 2014) and Teacher Incentive Fund programs (Speroni et al., 2020), found mixed effects and could not determine the key characteristics for success.[2]

[2]See Pham, Nguyen and Springer (2020) for a complete meta-analysis of incentive programs in the U.S. Notable programs with individual-level incentives based on student test scores are studied by Dee and Wyckoff (2015), Dee and Keys (2004), Hudson (2010), Sojourner, Mykerezi and West (2014), Speroni et al. (2020), Atteberry, Briggs and Lacour (2015), and Springer et al. (2012). Across these studies, incentives were only found to be effective in five programs, which are also the only programs to include in-class

In summary, teacher incentive programs offer insights into incentive design, particularly in government-provided services that lack market signals. However, research has yet to identify the mechanisms of effective teacher incentive design, including whether and how monitoring improves teacher performance (Bleiberg et al., 2021; Kraft and Christian, 2021).

III. Data and Empirical Approach

III.A Setting and Data Source

In the 2009-10 school year, DCPS introduced several reforms, collectively called the IMPACT program, in an attempt to address low student performance. IMPACT's structure was unchanged for the first three years of its implementation (Toch, 2018) and featured three key elements:

1. Teachers would undergo annual performance evaluations using multiple measures. These would include five standards-based classroom observations and a measure of student performance;
2. Teachers would be given feedback following their observations to support their professional development;
3. Teachers with poor performance would face sanctions – a freeze in their annual pay increases – or even dismissal. High-performing teachers would receive substantial one-off bonuses or even permanent pay increases.

IMPACT's implementation prioritized establishing well-calibrated, differentiating classroom observations. All teachers received three observations from school administrators (the principal or vice-principal) and two from external observers, called "Master Educators." Likewise, in this paper I distinguish between "internal observations" and "external observations." Each observation used a well-defined, nine-dimensional rubric

observations as part of the incentive. Notable programs with grade- or school-level bonuses – with varying levels of effectiveness – include the School-wide Bonus Program in New York (Fryer, 2013), the Dallas School Accountability and Incentive Program (DSAIP) (Ladd, 1999), the Kentucky Instructional Results Information System (KIRIS) (Koretz and Barron, 1998), the North Carolina ABC program (Vigdor, 2008), the Chicago version of TAP (Glazerman and Seifullah, 2012), and Houston's ASPIRE program (Imberman and Lovenheim, 2015).

called the "Teaching and Learning Framework" (TLF). Teachers received a score between 1 and 4 on each dimension, then the scores were averaged to construct the final TLF score. Afterwards, teachers received a follow-up meeting within ten school days to discuss their evaluation. For improved standardization, all of the observers received training and practiced calibrating scores by evaluating pre-recorded teaching samples. Unlike many other classroom observation systems in which nearly all teachers are identified as effective or better (Kraft and Gilmour, 2017), roughly a quarter of the teachers in my sample were rated as "less than effective."

In-class observations successfully differentiated quality teaching, in large part due to the policies surrounding their implementation. To improve both observation rigor and feedback quality, most of these observations were unannounced. The exception was the first internal observation, which teachers were made aware of at least a day in advance. Classroom observations were also relatively long, lasting a minimum of 20 minutes. These and other policies, however, also made an observation's timing within the day logistically complicated. For example, observers needed to avoid teaching blocks devoted to writing, physical education, music, or any other class that involved the students leaving the classroom or being instructed by a different teacher. Thus, most observations involved a reading, math, or science teaching block. This was logistically difficult given block schedules vary between teachers, even within the same grade.

External observers (Master Educators) were assigned to teachers in a manner meant to increase fairness and improve the quality of an observer's feedback. External observers were assigned teachers based on the observer's experience and subject expertise. For example, observers with experience teaching grades K through 5 would be assigned teachers in those grades. The district further sought to improve fairness by spreading external observers among schools and working to ensure teachers were not evaluated twice by the same observer within the same year. Additionally, external observers could not be assigned to teachers they had previously worked with or schools where they were previously employed. As a result of these policies, external observers had assignments across many schools and rarely observed more than two teachers at a single school, even across all grades. Similarly, teachers in the same school and grade usually were assigned different external observers.

All teachers received an IMPACT score that was used to qualify them for performance incentives. Where possible, teacher performance was primarily measured using estimates of a teacher's value-added toward student performance on standardized tests. Value-added scores require that a teacher's students have prior test scores available, but students in DCPS only complete standardized testing from grades 3 to 8, and then only in Math and English Language Arts (ELA, or Reading). This effectively limited value-added measures to grades 4 through 8. For these teachers, the IMPACT score assigned 50 percent weight to a teacher's value-added score and 45 percent weight to classroom evaluations. The remaining weight came from the teacher's score on the "Commitment to School and Community" rating, determined by the principal. Based on their overall numeric IMPACT score – which ranges between 100 and 400 – teachers received a rating of "Ineffective" (score below 175), "Minimally Effective" (between 175 and 250), "Effective" (between 250 and 350), or "Highly Effective" (greater than 350). For context, 11 percent of teachers in my sample are rated Highly Effective. The majority (65 percent) fall within the Effective range, while 24 percent are rated Minimally Effective or lower.

These scores had meaningful consequences. Highly Effective teachers received one-off bonuses ranging from $5,000 to $25,000 depending on the school, grade, and subject taught. If teachers were rated Highly Effective two years in a row, they received permanent pay increases that ranged from $6,000 per year to more than $20,000 per year.[3] If a teacher was rated Minimally Effective, she experienced a pay freeze, meaning her salary would not increase as normal that year. She would also be required to improve to Effective in the next year or be dismissed. Receiving a rating of Ineffective led to immediate dismissal. Only 1 percent of teachers received a final rating of Ineffective during the study period.

Several studies have assessed the broader effects of the IMPACT program on overall teacher performance and retention. One earlier study finds that the program's largest effects came from removing the lowest-performing teachers, with a smaller effect attributable to incentives for improvement (Dee and Wyckoff, 2015). Other work has

[3]Pay increases depended on a variety of factors, such as a teacher's current base pay, whether the school was a high-poverty school (60 percent or more of students received free or reduced-price lunch), or if she taught a high-need subject. See Dee and Wyckoff (2015) or Toch (2018) for more details.

assessed the program's equilibrium effects on teacher turnover (James and Wyckoff, 2020) and its sustainability (Dee, James and Wyckoff, 2021). These studies find that – even after ten years – the program continued to have positive effects by either removing or improving the lowest performing teachers.

III.B Description and Calculation of Unmonitored Time

Importantly for my purposes, the district created a clear policy outlining when in-class observations could occur. While most in-class observations were unannounced, they were designed to occur in multiple pre-specified time windows, illustrated in Figure 1. These windows would create exogenously determined periods of unmonitored time. As mentioned, there are two types of observations: three internal observations conducted by the school's principal or vice-principal (labeled $P1$, $P2$, and $P3$) and two external observations conducted by a Master Educator (labeled $M1$ and $M2$). The first internal observation was to occur by December 1; the second occurred after December 1 and before March 15, and the third would be after March 15. The external observations split the school year: the first occurred before February 1 and the second occurred afterward. In my analysis, the second internal ($P2$) and second external ($M2$) observations provide clean variations in monitored time.

The grey shading in Figure 1 highlights the time in which teachers may go unmonitored. Starting in February of all three years in the sample, all teachers had the possibility of an external observer appearing unannounced. Additionally, all teachers received an unannounced internal observation between December and mid-March (just before standardized exams). If *both* the external and internal observations were completed before standardized tests, teachers were guaranteed to have no further visits until the start of the next internal observation window.[4] As a result, teachers could have anywhere

[4]In theory, other windows of monitored/unmonitored time are possible. However, because of changes in which observations would be unannounced, these other windows became less consistent over time. For example, the span from December to February would define another possible window, but in some years the timing of the first external evaluation was known, making it less clear which teachers were unmonitored. The resulting sample is reduced by one-third due to this inconsistency in the treatment definition. A small window is also possible if all evaluations for a teacher are completed before standardized testing, but given its small size, there is little variation in treatment during this time. This window is further complicated by variations in when standardized tests were administered each year.

between zero and six weeks of unmonitored time, depending on the year.

I calculate each teacher's number of unmonitored days prior to standardized testing by determining which teachers had completed both their second internal observation and second external observation ($P2$ and $M2$). As the dates of each observation are known to me, I track the number of business days from the last of these two observations until the start of standardized testing. I calculate business days and account for all holidays and "in-service" days as provided on the publicly available district calendars.

One concern with using unmonitored time in this way is that teachers with more unmonitored time also have more opportunities to implement new observer feedback. This may potentially counteract the effects of being unmonitored. However, I can observe when teachers received their feedback and use this to control for its effects. While feedback must occur within ten business days, there is sufficient variation such that feedback time — measured as days from feedback until standardized testing — is not perfectly correlated with unmonitored time. Additionally, unmonitored time requires that both internal and external observations be completed, whereas feedback is measured from each individual observation. Bear in mind that the observation window is close to the standardized tests, making it unclear whether feedback should have much of an effect on teacher behavior. While I use feedback as a control, the estimates of its effects are a noisily measured zero.

III.C Econometric Specification

The main outcomes I measure are student standardized test scores in reading and math, denoted as Y_{ijts} for student i with teacher j in year t at school s. Let n_{jt} denote the number of unmonitored days teacher j experiences in year t prior to standardized testing. I estimate the following equation:

$$Y_{ijts} = W_{it}\Omega + X_{jt}\Gamma + \beta n_{jt} + \phi_s + \delta_j + \varepsilon_{ijts} \tag{1}$$

The coefficient β indicates the change in the marginal daily contribution for an unmonitored day relative to a monitored day, conditional on the controls. I would expect β to be negative if teacher j's behavior on unmonitored days is less effective than her peer teachers' behavior.

I control for school-level characteristics using school fixed effects ϕ_s. My setting also allows for teacher fixed-effects, δ_j, which helps assuage concerns about teachers being assigned certain types of students non-randomly, along with other possible confounders. The variable X_{jt} is a vector of annual teacher experience dummies (capped at 15) and pay-scale levels. The variables in W_{it} are student-specific characteristics: student i's previous scores, the leave-one-out average of the class's previous scores, free-reduced price lunch status, English Language Learner status, special education status, race, gender, and a dummy variable for students who spent less than 95% of the year with their teacher on record, which controls for the disruptions of switching schools or classes.

I assume that ε_{ijts} is conditionally independent of n_{jt}. That is, $E[n_{jt}\varepsilon_{ijts}|X_{jt}, W_{it}, \phi_s] = 0$. This amounts to assuming there are no unobservable characteristics of a teacher or her students that are correlated with a teacher's contribution to test scores and that also systematically change her number of unmonitored days. If evaluators systematically target low-quality teachers or underperforming students early in the year based on criteria that I cannot observe, then my results will be negatively biased.

I estimate Equation 1 using ordinary least squares with clustered errors at the classroom level (teacher by year). In line with the guidance by Abadie et al. (2023), standard errors are clustered at the level at which treatment is randomized, which is each teacher×classroom combination.[5]

III.D Data Sample and Summary

The sample is limited to students in the fourth and fifth grades as these are grades for which detailed student-teacher links are available, as well as grades when students typically have only one teacher. While standardized testing begins in third grade, DCPS only records detailed teacher-student combinations starting in fourth grade. For high school (grades 9-12), where teachers are not assessed based on student test scores, student-teacher

[5]This decision does not have a substantial effect on the results. When clustering at the teacher level, the p-value in the key results increases from 0.001 to 0.006 for reading and from 0.01 to 0.04 for math. Also, because external evaluators rarely have more than one or two teachers assigned to them at any given school, there is no *a priori* reason to believe that treatment assignment is correlated with the assigned Master Educator. However, as a check, when two-way clustering with Master Educator as an additional cluster, the standard errors change only slightly and the p-values are 0.009 for reading and 0.02 for math.

links are not reliably available. Grades 6-8 are excluded due to varied and complex class structures across schools, including multiple teachers per student or content-specific instructors. These arrangements are not fully observable in the data, making it unfeasible to accurately measure the treatment for these teachers and their students.

I also impose a few restrictions on which students and teachers are included in the sample. To be included, students must have a test score available from the prior year, which excludes students in their first year in the district. This allows for using prior student test scores in the specification, though the results are robust across other specifications with the full sample of students. I exclude novice teachers (those with no prior teaching experience) from the sample, though results likewise remain robust when including them.[6] This exclusion clarifies the interpretation of results as employee responses to monitoring. Novice teachers, who are still developing their teaching practices, have a limited ability to adjust their pedagogy in response to monitoring, as shown in the DCPS context (Phipps and Wiseman, 2021). Additionally, there was initial ambiguity in the program regarding unannounced observations for novice teachers, potentially confounding their treatment (Toch, 2018). This concern is supported by the data, and program policies were adjusted in the 2012-2013 school year to address this issue.

Table 1 provides context on the sample and its characteristics. The sample includes approximately 4,000 students and 220 teachers per year, with an average class size of 18 students. The student population is predominantly low-income, with nearly 70 percent of students receiving free or reduced-price lunch. Demographically, the student body is 70 percent Black, 15 percent Hispanic, and 11 percent White. The teacher population proportionally matches the Black student representation, but White teachers are overrepresented while Hispanic teachers are underrepresented. The study covers about 81 schools annually, with substantial between-school variation in student race and socioeconomic status. Table 1 also provides information on the frequency and extent of the treatment, i.e., unmonitored days. Over the three-year study, 25 percent of teachers had

[6]When including novice teachers, the estimated coefficient of the effect of an unmonitored day is -0.007 (0.0028) standard deviations in reading scores and -0.007 (0.0036) in math. The p-values for these estimates are 0.02 and 0.05, respectively.

some unmonitored time each year. For these teachers, the average unmonitored span was approximately six school days, slightly more than a week of instruction.

IV. Causal Identification and Balance

The measured effect of unmonitored time is causal under the identifying assumption that the number of unmonitored days is independent of unobserved qualities that affect a student's test scores, including the qualities of her teacher:

$$E[n_{jt}\varepsilon_{ijts}|W_{it},\, X_{jt},\, \phi_s] = 0$$

for student i with teacher j in year t and school s. This assumption is violated if administrators or external observers choose the timing of their visits based on some student or teacher quality I do not observe in the data that also correlates with student performance.

The causal interpretation of my results rests on the complexity of scheduling classroom observations, which makes systematic targeting of specific teachers or student groups improbable. This complexity stems from various factors, beginning with the diverse schedules of teachers. Observations must accommodate these schedules, avoiding conflicts with specialized activities like physical education or music instruction, which vary across teachers and grades within schools. External observers, who typically evaluate teachers across multiple schools, face additional constraints. With 80% observing only one or two teachers per school in my sample, these observers' ability to target specific teachers or student groups is limited. As expected, observers reported scheduling based on their district obligations, location, and teacher availability, rather than targeting specific classes.

Administrators, including principals and vice-principals, face similar logistical challenges in conducting internal observations. Coordinating their schedules to systematically observe certain teachers or students earlier in the year would be difficult, given the limited overlapping availability between administrators and teachers, especially within a single grade. Moreover, administrators were instructed to avoid predictable

patterns in their observations, further reducing the likelihood of systematic targeting.

Anecdotal evidence from conversations with administrators and external observers supports this view. They indicated no motivation for targeted observations and deemed the practice unrealistic given the complexity of coordinating such targeting across multiple observers or administrators. In practice, principals often conducted evaluations on an ad hoc basis as their schedules allowed. The combination of scheduling constraints, lack of coordinated efforts, and practical implementation challenges create plausibly exogenous variation that justifies my causal identification strategy.

To test this assumption empirically, I provide evidence that treatment was balanced among teachers and students. While the information available to me is not identical to that available to a principal, I have information on both students and teachers about their characteristics and prior performance that I use to check for correlations in treatment.

Table A1 in the Appendix reports the balance results for students with unmonitored days as the outcome and rows as student covariates. These specifications include school, year, and subject fixed-effects. There are no consistent patterns in sign or magnitude for previous scores, free and reduced-price lunch, English learner, or special education. The coefficients are small, showing less than 0.2 unmonitored days for any characteristic and less than 0.05 for half of the dichotomous characteristics. In most cases, the standard errors are larger than the coefficient, and none of the coefficients are significant to any degree. The F-statistics are also not significant and range between 0.76 and 0.97.

Similarly, the number of unmonitored days does not appear to correlate with any observable teacher characteristics. Table A2 in the Appendix shows these results, where the observable characteristics considered are a teacher's prior experience and her race or ethnicity. Note that the race variables for teachers provided by the school district were coded differently than for students, limited to White, Black, Hispanic, and Asian. Additional characteristics available in the last two years of the data are a teacher's lagged observation score and whether or not a teacher was ranked Minimally Effective or Highly Effective in the previous year (relative to teachers with an Effective rating). Again, the coefficients are small and none are significant. For example, an additional year of experience appears to correlate with a 0.01 increase in unmonitored days (standard error of

0.09). Being rated Minimally Effective in the previous year correlated with a decrease of -0.05 unmonitored days (standard error of 0.64). The F-statistics range between 0.32 and 0.47 and are not significant.

V. Results

V.A Effects of Unmonitored Time on Student Performance

Unmonitored time has substantial effects on student outcomes, at least in the time period leading up to standardized tests. The estimated effects of unmonitored time are visualized in Figure 2, showing that each additional monitored day alters reading scores by -0.0095 standard deviations and math scores by -0.009 standard deviations. More detailed results are shown in Table 2, where the results in Figure 2 are found in Columns 2 and 5. The 95 percent confidence interval for the p-values of 1,000 randomization inference trials is also shown in brackets. Randomization inference is conducted as an additional test of the treatment effect's significance, given that the analysis is clustered at the classroom level and may suffer from a (potentially) small sample size. These tests help assuage concerns that the setting is under-powered and the results are driven by an abnormally large "draw" of effect size. The first three columns look at standardized reading test outcomes and the last three look at math. For each set, the first column shows results for the simplest specification, which includes only year, school, teacher, and grade fixed-effects and controls for each student's previous test scores, as well as their classmates' previous scores. The second column adds teacher experience controls and student demographic controls. Finally, the third column controls for the time a teacher has after she's received her observation feedback. This controls for the potential countervailing effect that teachers with more unmonitored time are also likely to have more time to implement the feedback they received.

Progressing from left to right in Table 2, adding controls for teacher experience and student characteristics appears to reduce the effect of unmonitored time only slightly: from -0.0114 to -0.0095 standard deviations in reading and from -0.0112 to -0.0090 in math.

Looking to Columns 3 and 6, controlling for feedback increases the observed impact of unmonitored time, as expected, to -0.0101 standard deviations in reading and to -0.0103 in math. However, feedback time is somewhat correlated with unmonitored time, which makes the standard errors on unmonitored time nearly double as a result of the collinearity. The effect feedback has on student performance (not shown) is noisy and not statistically significant.

To understand the average effects, recall that about 25% of teachers experience unmonitored time each year. Among these teachers, the average length of unmonitored time is slightly more than five business days. That means the average unmonitored teacher's students perform about 5% of a standard deviation worse on both reading and math, representing a drop of roughly 0.3 SD in *teacher* value-added in math and 0.4 SD in reading (Chetty, Friedman and Rockoff, 2014). This suggests a substantial portion of the variation in teacher value-added is affected in the months just before standardized tests. These results can also be compared to the estimated effects of a suspension day: one unmonitored day reduces student test scores by roughly 3% of the estimated effect of a suspension day (Lacoe and Steinberg, 2019).

These observed effects are relatively large, but they should be interpreted carefully with regard to their context and compared cautiously to other interventions. Following guidance by Kraft (2020) on interpreting educational intervention effects, several factors distinguish this setting: the short-term nature of the changes in teacher effort, the proximity of treatment to the outcome measurement, the uniquely low-socioeconomic status sample, and the use of district-specific standard deviations. These factors can contribute to larger effect sizes compared to year-long interventions or those measured in more diverse populations. The short duration of unmonitored periods, typically less than a few weeks, aligns with the literature showing larger effects for short-term efforts. Additionally, the sample's homogeneity likely results in smaller standard deviations, potentially inflating effect sizes compared to nationally representative samples. Particularly noteworthy is the "treatment" proximity to standardized tests. This research design observes within-year pedagogical variation during critical months for standardized testing, a unique approach in the literature. This aspect makes the results most comparable to

research on student suspensions, which similarly measures short-term, proximate effects. Causal estimates of suspension effects are also large, ranging between 0.28 and 0.54 standard deviations per suspended day (Lacoe and Steinberg, 2019). These contextual factors are important to bear in mind when interpreting the magnitude of these effects.

It is also important to consider whether these effects indicate enduring changes in student performance. As Gilraine and Pope (2021) show, long-run measures of teacher value-added in their setting are only 51% correlated with immediate value-added. This also highlights the extent to which immediate test scores are malleable. The time period I observe is one in which teachers face substantial pressure to help students prepare for standardized tests. The counterfactual teachers — those who are still monitored during this time — may be exerting substantial effort to defend against a poor observation score during a high-pressure part of the school year. Principals also received performance bonuses based on their students' standardized scores, which may influence the severity of their pedagogical criticism. The degree to which pedagogy directly before testing can affect student performance may underscore the short-term effects of intensive teaching.

Overall, these results show that monitoring can materially affect employee performance. This finding indicates that the benefits of monitoring extend beyond the prevention of unethical behavior and can positively influence employee behavior. Given the intense cultural and financial pressure to increase student test scores, it seems unlikely that unmonitored teachers completely shirked their teaching responsibilities. Rather, the monitored teachers would have pressure to maintain classroom and pedagogical standards on top of test preparation. In later sections, I examine how teachers changed their pedagogy and its effects on student behaviors.

Robustness and Sensitivity

The setting provides a placebo treatment, which is one way to alleviate concerns that these results are generated by factors other than monitoring. Specifically, unmonitored time that occurs *after* students complete their standardized tests should have no effect on student performance. We can check this effect because of the overlapping windows of the final internal observation combined with the second external observation (see Figure 1).

The placebo specification is the same as in the main analysis but with unmonitored time *after* standardized tests used in lieu of unmonitored time. The results are shown in Table 3, which indeed shows no evidence that student test scores correlate with the placebo treatment. The placebo coefficient is one tenth the magnitude of my key results with an estimated effect of about ± 0.001 in reading and math, but with tighter standard errors.

The balance checks described earlier suggest that the treatment is not specifically targeted towards underperforming teachers and students. However, there might be other relevant unobservable factors that correlate both with the treatment and student performance. Using the procedure from Oster (2019), I calculate the amount of explanatory variation that must come from unobservable characteristics to explain the observed effect size. The results are shown in the Appendix, Table A3. Even under the extreme assumption that the maximum R^2 is greater than 0.95, I find that my results could not be explained by unobserved factors unless they correlated with the outcome by more than 137 percent as much as the observable characteristics. Given the availability of teacher fixed effects and previous student test scores, this seems unlikely.

My key results (Table 2) show that the specification is not sensitive to which covariates are included. However, the specification may be sensitive to how I specify and control for teacher experience. For my preferred specification, I use an indicator for years of experience capped at 15 years. As a check on the sensitivity to this choice, Table A4 of the Appendix shows that the results are not meaningfully different for smooth measures of experience or those without a cap at 15 years.

Lastly, to understand possible heterogeneous effects by grade and the extent to which they may drive my core results, Table A5 shows results broken out by grade. Interestingly, the math effects are significantly higher in fourth grade than in fifth. The fourth-grade math curriculum covers fractions and operations on fractions, while the fifth-grade math curriculum covers decimals and their operations. For reading, there is no meaningful distinction in the curriculum and there is no observable difference in the effect of unmonitored days for reading between fourth and fifth grade.

V.B Assessing the Mechanisms

Changes in pedagogy

The large effects in my results highlight the importance of understanding how teachers may be changing their pedagogical approach when unmonitored. Without data on teacher activity before and after their classroom observations, I cannot directly measure pedagogy during unmonitored time. However, variation in the daily probability of an observation provides an opportunity to estimate how teachers change their pedagogical approach when they are monitored less intensely. I do so using teacher scores on their second external observation ($M2$). The external observation is the most prominent one in the month leading up to standardized tests, and as shown both here and in Phipps and Wiseman (2021), external evaluators maintain greater rating fidelity, resulting in evaluation scores with less noise. As I will show, the probability of being observed varies most during this time frame because of how it overlaps with the end of the internal observation window.

The treatment variable for this analysis is a teacher's perceived probability of being observed on each day in the observation window. This is somewhat intuitive: teachers who have not yet been observed but are nearing the end of the observation window are more likely to anticipate an upcoming observation. If a teacher has not been evaluated by the last day of the window, she can (in theory) be certain to receive her observation on the next day. Similarly, if she knows her principal has conducted observations for all but two of her peers, her odds of being next increase. My measure of perceived probability captures both these elements.

Measuring monitoring intensity relies on assuming how well teachers understand their probability of being observed. The district provided data on the date of each in-class observation for each teacher, which I use to calculate how likely the remaining teachers are to be observed on each of the remaining days. Two factors determine my estimate of a teacher's beliefs about their probability: the number of teachers that remain to be observed at her school and how many observations a teacher expects to be conducted at her school each day. It is then possible to calculate the probability, assuming each remaining teacher has an equal probability.

Let v be an observation indicator, where v is $P2$ for the second internal observation and $M2$ for the second external observation. Then let a teacher's estimate of the number of observations to be conducted on day d at school s be $\hat{L}^v_{ds}$. If R^v_{ds} is the number of remaining teachers needing an evaluation v, then each remaining teacher's probability of being evaluated is

$$p^v_{ds} = \frac{\hat{L}^v_{ds}}{R^v_{ds}}. \tag{2}$$

The data allow me to determine the number of remaining teachers, R^v_{ds}. But estimating how many observations a teacher expects to be conducted, $\hat{L}^v_{ds}$, requires additional assumptions. I simplify by assuming teachers expect internal and external observers to conduct their observations somewhat evenly across the allotted window, which is to assume $\hat{L}^v_{ds} = \frac{N_s}{T^v}$, where N_s is the total number of teachers at a school and T^v is the number of days within the observation window for observation v. Then the probability on day d is $p_d = \frac{N/T}{R_d}$, where I've dropped the subscripts for readability. As the pool of possible teachers to be evaluated decreases (R_d), the probability of being observed increases.

As an example, consider a principal who needs to conduct 25 observations over 50 work days; a teacher would reasonably expect the principal to conduct roughly one observation every other day. If the principal fails to maintain a consistent schedule, the probability of being observed will not increase until other teachers are observed. This will mechanically occur as the time remaining in the window shortens and the principal "catches up" on completing evaluations.[7] My proposed estimate of the daily observation probability incorporates both possibilities, accounting for the less consistent observations of principals.[8] The resulting daily observation probability varies by day, by the order in which a teacher is observed relative to her peers, and — as I describe next — by whether a teacher has one or two outstanding observations on a given day. Importantly, these sources

[7]The interpretation is slightly different in the case of external observers because there are multiple observers assigned to a school. The estimate described here assumes that a teacher views the whole group of potential external observers as drawing the remaining teachers at her school at random. Since teachers do not know who their external evaluator is and evaluator assignments are orthogonal to any observable characteristics, this assumption seems reasonable.

[8]An alternative estimation could be to assume teachers are generally aware of an observer's behavior. I test this by estimating $\hat{L}^v_{ds}$ using kernel smoothing of an observer's actual observations, or simply using the *actual* number of observations on that day. However, the results do not meaningfully change.

of variation mean that even teachers evaluated on the same day can and do face different monitoring intensity, either because they are at different schools or because some may have already received one of their two outstanding observations. Figure 3 plots the probability of an observation by the day on which the observation occurred, illustrating the variation in monitoring intensity within days and across the time period.

Determining the effect of monitoring intensity uses the following specification. For a teacher j and standard S on the $M2$ observation at school s and year t,

$$S_{jts}^{M2} = \boldsymbol{X}_{jt}\Gamma - p_{jt}\mu + \sum_{\nu=P1,M1} S_{jt}^{\nu}\kappa^{\nu} + \boldsymbol{T}_{jt}\omega + \rho d_{jts} + \phi_s + \delta_t + \varepsilon_{jts} \tag{3}$$

where $\boldsymbol{X}_{jt}$ is a vector of experience indicators, ϕ_s is a school fixed-effect, and δ_t is a year fixed-effect. The term p_{jt} is measured as the probability of receiving *any* observation on the day of the teacher's $M2$ observation, and μ is the coefficient of interest. Note that if a teacher still has both her $P2$ and $M2$ observations outstanding, then the joint probability will be $p_{jt} = p^{P2} + p^{M2} - p^{P2} \times p^{M2}$.

A teacher's performance also depends on the order in which her observations occur: if the $M2$ observation is her fourth in the year, she usually does better than if it is her third. To account for this, I include the term $\boldsymbol{T}_{jt}$, a vector of indicators for if the $M2$ observation was third (before $P2$), fourth (after $P2$), or fifth (after $P3$). S_{jt}^{ν} are scores on Standard S for $\nu = P1, M1$, which are observations that must have already occurred by the time of her $M2$ observation.

Lastly, I include a time trend for the day of the observation, d_{jts}, to capture any system-wide changes in pedagogy as standardized tests approach. The results are robust to this decision, however, because the probability of being observed depends on the joint probability of overlapping observation windows. This results in a higher total probability of being observed at the *beginning* of the window as the second administrative observation ($P2$) is wrapping up (see Figure 3).

Looking at the description of the individual evaluation components (see Table A6 for complete descriptions), there are a few elements that would reflect a marked reduction in pedagogical quality. To highlight these elements in particular, I use the district's own

groupings that were enacted later. In 2016, the district consolidated its nine standards into three groups. The final and largest of these groups is "Engage students in rigorous and higher-level work," which is comprised of Standards 3, 6, 7, and 8, the standards I highlight. Standard 3, "Engage students at all learning levels in accessible and challenging work," measures whether teachers focus too much on some students at the expense of others. Standards 6 and 7 touch on probing for and building up deeper learning. Standard 7 is "Develop higher-level understanding through effective questioning," and Standard 6 captures probing for deeper understanding and building up knowledge gradually ("scaffolding"). Lastly, Standard 8 measures lesson pacing, student behavior, and idleness. All these pedagogical elements could be expected to suffer either as a result of decreased monitoring pressure and reduced teacher effort or as teachers are increasingly pressed to ensure their underperforming students meet test requirements. The pace of learning for these students may not meet their needs, and they likely have more behavioral concerns that could be exacerbated with less engaging instruction.

The results for the specification in Equation 3 are shown graphically in Figure 4, where Standards 3, 6, 7, and 8 are highlighted and the lines indicate the 95% confidence interval.[9] The coefficients are scaled such that they are a linear estimate of the difference in teacher behavior when switching from a 100 percent chance of an observation to a zero percent chance.

The results are quite large for several of the standards. For context, the average score across the nine Standards is roughly 3.1, and the average standard deviation is about 0.8. This means an average teacher moving from a 6% probability of being observed (the average) to an 11% probability (a one standard deviation increase) will receive 0.1 more points on Standard 7, which is roughly 12% of a standard deviation. Overall, the results show that standards measuring student engagement (Standard 3), responses to student understanding (Standard 6), development of higher-level learning (Standard 7), and pacing or student idleness (Standard 8) are where teachers consistently sacrifice the most when monitored less. At these times, teachers appear to shift their pedagogical priorities rather

[9]The confidence intervals shown are robust to multiple-hypothesis test adjustments, such as using sharpened q-values as in Benjamini, Krieger and Yekutieli (2006).

significantly.

It is important to note, however, that the pedagogical changes observed here still only apply to teachers who are currently monitored. That is, the results do not speak directly to how teachers change their behavior when they have absolutely no threat of an observation; they can only inform on the margin of monitoring intensity. But with that caveat in mind, it appears that monitoring affects teacher pedagogy in important ways, and when unmonitored, these changes in teaching quality have substantial effects on student performance.

These observed pedagogical changes appear as a rational response to decreased monitoring given the time and energy it takes to prepare classes. They also provide insight into which teaching elements are most costly to teachers. The decreased performance along standards relating to "engage students in rigorous and higher-level work" suggest that thoughtful, longer-run learning is more demanding. These results also suggest that "higher-level work" may have more immediate-term results than teachers believe. Indeed, Dinerstein and Opper (2022) find that teachers who are most capable at improving "untargeted outcomes" (i.e., those outcomes with no performance incentive attached) are capable of improving student test scores the most, consistent with the hypothesis that other learning and behavioral objectives are complements to test scores, not substitutes.

By neglecting student engagement at all student levels or mismanaging instructional time, teachers may open up the classroom for more disruptive behaviors. That is, unmonitored time may reduce the extent and quality of learning, but it may also invite other detriments to learning. This explanation seems plausible if these pedagogical changes are associated with serious behavioral problems, which is what I turn to next.

Suspensions

One way these pedagogical outcomes could have outsized effects on student performance is if they increased behavioral problems. Class disruptions can affect all students in the class, but suspended students are particularly likely to suffer on standardized tests (Noltemeyer, Ward and Mcloughlin, 2015; Lacoe and Steinberg, 2019). Of the measured pedagogical practices, sacrificing pacing, classroom management, and

student engagement are especially likely to affect student behavior.

To test this mechanism, I obtain suspension data for the 2011 and 2012 school years and repeat the analysis. Panel A of Table 4 shows the measured effect of unmonitored time on a student's total number of short-term suspensions (these exclude expulsions and suspensions incurred through criminal activity such as bringing a weapon to school). The specification is similar to Equation 1 but with student yearly suspensions as the outcome. Additionally, the suspension specification does not include previous year suspensions but instead uses student fixed-effects. This difference is driven by data limitations.[10]

In my preferred specification (Column 3 of Panel A), each unmonitored day leads to a statistically significant increase of 0.006 suspensions in the year. For context, students are suspended an average of 0.127 times a year, which means an unmonitored day increases suspensions by 4.5%. After a week of unmonitored time, this is an increase of 0.03 annual suspensions or a 24% increase. The subsequent columns help clarify how the coefficient evolves as I add teacher fixed effects (Column 2) and teacher experience (Column 3). Adding teacher fixed effects has no meaningful impact on the coefficient, though controlling for teacher experience does, increasing the estimated effect from 0.0051 to 0.0062, implying that teacher experience plays an important role in student suspensions. Though not shown, the coefficient for experience is negative, meaning more experienced teachers issue fewer suspensions even after controlling for student fixed effects.

To determine whether or not these additional suspensions occurred during unmonitored days, Panel B of Table 4 shows results based on the daily probability of issuing a suspension for monitored and unmonitored days. Here, each observation is a teacher$\times$day, where the treatment is whether or not that day is monitored and the outcome is whether or not the teacher issued a suspension on that day. The specification considers only days that fall within possible unmonitored days, which is roughly 30 business days. Unfortunately, accurate information regarding the date of suspension is available only for

[10]While the district maintained the necessary records of student prior scores for all years, they did not keep consistent student identifiers to link students across all three years. This was corrected in the 2011 and 2012 school years, which are also the only years for which suspension data were available. In the same vein, this means using lagged student suspensions would limit the sample to a single year, severely limiting its statistical power, which is why I opted to use fixed effects instead.

fifth grade for 2011 and 2012, which greatly reduces the sample size. However, the results are (marginally) significant and consistent with the previous analyses. On an unmonitored day, teachers are between 3.0% and 4.1% more likely to issue a suspension. Given the average teachers in this time frame have a 3.7% likelihood of issuing a suspension, teachers are effectively twice as likely on an unmonitored day.

Lacoe and Steinberg (2019) estimate that an additional suspension day reduces a student's performance between 0.28 and 0.54 standard deviations, which aligns with non-causal estimates from a meta-analysis (Noltemeyer, Ward and Mcloughlin, 2015). A back-of-the-envelope calculation would suggest that an unmonitored day can reduce an average student's performance by between 0.002 and 0.003 standard deviations through suspensions alone: (additional suspensions) $\times$ (suspension effect) $= 0.006 \times 0.54 = 0.003$. This is roughly 20 to 30% of the observed effect of an unmonitored day. Of course, these are effects estimated through the suspended student alone. Their misbehavior can also negatively affect peers (Carrell, Hoekstra and Kuka, 2018), but given the limitations of the data, it is difficult to pin down exactly how much student behavioral problems account for the effects of unmonitored time. Even still, they likely constitute a substantial portion.

VI. Discussion and Conclusion

Surprisingly little is known about how monitoring affects employee performance outside the laboratory setting. This study advances the literature by examining the impact of monitoring in a real-world context, public schools in Washington DC. My findings reveal that monitored teachers demonstrate significant changes in pedagogical practices, while their students perform substantially better and are less likely to be suspended than students of unmonitored teachers. These results have implications for education policy as well as personnel policy more broadly.

VI.A Implications for Education Policy

This study's findings provide insight into a crucial question for policymakers aiming to improve K-12 education through professional accountability. When done right, these

incentives can elevate career paths for high-quality teachers, attracting more talent to the profession and improving retention rates (Croft, Guffy and Vitale, 2018; Hoxby and Leigh, 2004). However, poorly crafted incentives can backfire. Survey data shows that teachers subject to test-based bonuses report a higher likelihood of leaving the profession, along with decreased feelings of autonomy and satisfaction with their compensation.[11] This underscores the balance policymakers must strike in designing effective accountability measures for teachers.

This study advances that agenda and contributes to the growing body of literature on teacher incentives by being the first to demonstrate the significant (causal) impact of monitoring on teacher performance, independent of feedback effects. These findings suggest that accountability and motivation are crucial drivers of enhanced teacher performance, extending beyond the benefits of improved training alone. While the feedback effects I estimate are noisy, the null effect aligns with recent research by (Kraft and Christian, 2021; Bleiberg et al., 2021).

VI.B Implications for Personnel Policy

Monitoring's success in this context is likely attributable to the qualities that distinguish IMPACT's design: a clearly designed and rigorous rubric, high-quality training for observers, and substantive consequences resulting from a teacher's performance. The observed changes in teacher pedagogy, particularly in areas of student engagement and higher-level learning, highlight how such monitoring can encourage teachers to adopt more effective instructional practices. These positive outcomes stem from IMPACT's ambitious design that credibly distinguishes teacher quality based on *inputs*, provides clear avenues for improvement, and rewards high performance. These general elements are applicable to a variety of occupations. Healthcare provides one example. Attempts to implement performance pay based on patient outcomes have generally failed to improve patient health

[11] Based on survey results from the nationally representative School and Staffing Survey, teachers with test-based bonuses included in their salary are more likely to say they would leave teaching if given the chance (20.5% versus 15.4% of teachers without test-based bonuses). Test-based bonuses also correlate with a decreased sense of autonomy (24% say they have no autonomy versus 16%) and decreases in satisfaction with salaries and pay (38% are at least somewhat satisfied as opposed to 49% in other schools). Calculations are the author's, using NCES PowerStats Version 1.0.

(Shen, 2003; Serumaga et al., 2011; Flodgren et al., 2011; Green, 2014). In contrast, performance incentives based on monitoring well-defined, accessible employee behaviors have shown more promise (Flodgren et al., 2011).

Effective monitoring requires an understanding of the production function of employees in a given job. In healthcare, the desired behaviors are well-studied and measurable. Similarly, monitoring is feasible in the K-12 education setting in no small part because of the substantial effort devoted to generating effective measures of teacher pedagogy. What's more, an ecosystem of researchers have studied (and continue to study) pedagogy. Without these rigorous assessments, monitoring is ineffective (Weisberg et al., 2009).

The potential downsides of such high-stakes monitoring systems still remain. If poorly designed, monitoring can encourage gaming or over-emphasis on a single component of the evaluation rubric (e.g., Holmstrom and Milgrom, 1991; Baker, 1992; Jacob and Levitt, 2003; Martinelli et al., 2018). Employees may also be reticent to forfeit autonomy over their selection of inputs, making such programs difficult to implement or overly restrictive (Toch, 2018). There is also a financial cost to implementing evaluations. In 2017, for instance, DCPS chose to reduce the number of evaluations due to cost concerns.

Still, this study demonstrates that thoughtful, well-implemented monitoring offers a valuable alternative in many jobs where output-based performance pay is infeasible or ineffective. While this study provides evidence that monitoring can substantially improve employee outputs, it is impossible to directly compare whether monitoring is more effective than performance pay in this setting. This leaves an important avenue for future work. Additional research is also needed to parse the effectiveness of different incentive pay programs in a variety of settings, as well as to determine whether monitoring's effects are ameliorated by market pressures that are already present in the private sector.

References

Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. 2023. "When Should You Adjust Standard Errors for Clustering?" *The Quarterly Journal of Economics*, 138(1): 1–35.

Akerlof, George A. 1982. "Labor Contracts as Partial Gift Exchange." *The Quarterly Journal of Economics*, 97(4): 543.

Akerlof, George A., and Rachel E. Kranton. 2005. "Identity and the Economics of Organizations." *Journal of Economic Perspectives*, 19(1): 9–32.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, 113(1): 151–184.

Atteberry, Allison, Derek C Briggs, and Sarah Lacour. 2015. "Year 2 Denver ProComp Evaluation Report : Teacher Retention and Variability in Bonus Pay , 2001-02 through 2013-14." Colorado Assessment Design Research and Evaluation Center, University of Colorado, Boulder.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy*, 100(3): 598–614.

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76(2): 169–217.

Benabou, Roland. 2016. "Bonus Culture: Competitive Pay, Screening, and Multitasking." *Journal of Political Economy*, 124(2): 305–370.

Benabou, Roland, and Jean Tirole. 2003. "Intrinsic and Extrinsic Motivation." *Review of Economic Studies*, 70(3): 489–520.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika*, 93(3): 491–507.

Bleiberg, Joshua, Eric Brunner, Erica Harbatkin, Matthew A. Kraft, and Matthew Springer. 2021. "The Effect of Teacher Evaluation on Achievement and Attainment: Evidence from Statewide Reforms." Annenberg Institute at Brown University.

Briggs, Derek, Elena DiazBbilello, Andrew Maul, Michael Turner, and Charles Bibilos. 2014. "Denver ProComp Evaluation Report: 2010-2012." Colorado Assessment Design Research and Evaluation Center, University of Colorado, Boulder.

Camerer, Colin F., and Roberto A. Weber. 2012. "Experimental Organizational Economics." In *The Handbook of Organizational Economics.* , ed. Robert Gibbons and John Roberts, 213–262. Princeton, NJ:Princeton University Press.

Carrell, Scott E., Mark Hoekstra, and Elira Kuka. 2018. "The Long-Run Effects of Disruptive Peers." *American Economic Review*, 108(11): 3377–3415.

Chetty, Raj, John Friedman, and Jonah Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633–2679.

Croft, Michelle, Gretchen Guffy, and Dan Vitale. 2018. "Encouraging More High School Students to Consider Teaching." American College Testing.

Dee, Thomas S., and Benjamin J. 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 S, and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management*, 34(2): 1–31.

Dee, Thomas S., Jessalynn James, and Jim Wyckoff. 2021. "Is Effective Teacher Evaluation Sustainable? Evidence from District of Columbia Public Schools." *Education Finance and Policy*, 16(2): 313–346.

Dinerstein, Michael, and Isaac M. Opper. 2022. "Screening with Multitasking: Theory and Empirical Evidence from Teacher Tenure Reform." *NBER Working Paper*, 30310.

Dixit, Avinash. 2002. "Incentives and Organizations in the Public Sector: An Interpretative Review." *Journal of Human Resources*, 37(4): 696–727.

Duflo, Esther, Rema Hanna, and Stephen P Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review*, 102(4): 1241–1278.

Ferejohn, John, and Charles Shipan. 1990. "Congressional Influence on Bureaucracy." *Journal of Law, Economics, and Organization*, 6(Special Issue): 1–20.

Flodgren, Gerd, Martin P Eccles, Sasha Shepperd, Anthony Scott, Elena Parmelli, and Fiona R Beyer. 2011. "An Overview of Reviews Evaluating the Effectiveness of Financial Incentives in Changing Healthcare Professional Behaviours and Patient Outcomes." *The Cochrane Library.*

Francois, Patrick. 2000. "'Public Service Motivation' as an Argument for Government Provision." *Journal of Public Economics*, 78(3): 275–299.

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

Gailmard, S. 2002. "Expertise, Subversion, and Bureaucratic Discretion." *Journal of Law, Economics, and Organization*, 18(2): 536–555.

Gibbs, Michael, Kenneth Merchant, Wim Van der Steded, and Mark E Vargus. 2004. "Determinants and Effects of Subjectivity in Incentives." *The Accounting Review*, 79(2): 409–436.

Gilraine, Michael, and Nolan G. Pope. 2021. "Making Teaching Last: Long-Run Value-Added." *NBER Working Paper*, 29555.

Glazerman, Steven, and Allison Seifullah. 2012. "An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years." Mathematica Policy Research, Inc.

Green, Ellen P. 2014. "Payment Systems in the Healthcare Industry: An Experimental Study of Physician Incentives." *Journal of Economic Behavior and Organization*, 106: 367–378.

Hammond, T. H., and J. H. Knott. 1996. "Who Controls the Bureaucracy?: Presidential Power, Congressional Dominance, Legal Constraints, and Bureaucratic Autonomy in a Model of Multi-Institutional Policy-Making." *Journal of Law, Economics, and Organization*, 12(1): 119–166.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." *The Bell Journal of Economics*, 13(2): 324–340.

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

Hoxby, Caroline M, and Andrew Leigh. 2004. "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States." *American Economic Review*, 94(2): 236–240.

Hudson, Sally. 2010. "The Effects of Performance-based Teacher Pay on Student Achievement." *SIEPR Discussion Papers*, 94305(09): 1–49.

Imberman, Scott A, and Michael F Lovenheim. 2015. "Incentive Strength and Teacher Productivity: Evidence from a Group-Based Teacher Incentive Pay System." *Review of Economics and Statistics*, 97(2): 364–386.

Jacob, Brian A, and Steven D Levitt. 2003. "Rotten apples: An investigation of the prevalence and predictors of teacher cheating." *The Quarterly Journal of Economics*, 118(3): 843–877.

James, Jessalynn, and James H. Wyckoff. 2020. "Teacher Evaluation and Teacher Turnover in Equilibrium: Evidence From DC Public Schools." *AERA Open*, 6(2): 2332858420932235.

Kandel, Eugene, and Edward P Lazear. 1992. "Peer Pressure and Partnerships." *Journal of Political Economy*, 100(4): 801–817.

Kane, Thomas J, and Douglas O Staiger. 2012. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Bill and Melinda Gates Foundation, Met Project, Seattle.

Kane, Thomas J, and Steven Cantrell. 2010. "Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project." Bill and Melinda Gates Foundation, MET Project, Seattle.

Koretz, Daniel, and Sheila Barron. 1998. "The Validity of Gains in Scores on the Kentucky Instructional Results Information System." RAND.

Kraft, M., and A. Christian. 2021. "Can Teacher Evaluation Systems Produce High-Quality Feedback? An Administrator Training Field Experiment (EdWorkingPaper: 19-62)." Annenberg Institute at Brown University.

Kraft, Matthew A. 2020. "Interpreting Effect Sizes of Education Interventions." *Educational Researcher*, 49(4): 241–253.

Kraft, Matthew A., and Allison F. Gilmour. 2017. "Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness." *Educational Researcher*, 46(5): 234–249.

Lacoe, Johanna, and Matthew P. Steinberg. 2019. "Do Suspensions Affect Student Outcomes?" *Educational Evaluation and Policy Analysis*, 41(1): 34–62.

Ladd, Helen F. 1999. "The Dallas School Accountability and Incentive Program: An Evaluation of its Impacts on Student Outcomes." *Economics of Education Review*, 18(1): 1–16.

Lazear, Edward P. 1986. "Salaries and Piece Rates." *The Journal of Business*, 59(3): 405–431.

Lazear, Edward P. 2000. "Performance Pay and Productivity." *American Economic Review*, 90(5): 1346–1361.

Lazear, Edward P. 2006. "Speeding, Terrorism, and Teaching to the Test." *Quarterly Journal of Economics*, 121(3): 1029–1062.

Lazear, Edward P. 2018. "Compensation and Incentives in the Workplace." *The Journal of Economic Perspectives*, 32(3): 195–214.

Lazear, Edward P, and Paul Oyer. 2012. "Personnel Economics." *The Handbook of Organizational Economics*, 21(4): 479–517.

Lazear, Edward P, and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *Journal of Political Economy*, 89(5): 841–864.

Levin, Jonathan. 2003. "Relational Incentive Contracts." *American Economic Review*, 93(3): 835–857.

MacLeod, W Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review*, 93(1): 216–240.

Martinelli, César, Susan W. Parker, Ana Cristina Pérez-Gea, and Rodimiro Rodrigo. 2018. "Cheating and Incentives: Learning from a Policy Experiment." *American Economic Journal: Economic Policy*, 10(1): 298–325.

Murnane, Richard, and David Cohen. 1986. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and Few Survive." *Harvard Educational Review*, 56(1): 1–17.

Nagin, Daniel S, James B Rebitzer, Seth Sanders, and Lowell J Tayler. 2002. "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment." *American Economic Review*, 92(4): 850–873.

Nalbantian, Haig R, and Andrew Schotter. 1997. "Productivity Under Group Incentives: An Experimental Study." *American Economic Review*, 87(3): 314–341.

Neal, Derek. 2011. "The Design of Performance Pay in Education." In *Handbook of the Economics of Education*. Vol. 4, , ed. Eric Hanushek, 499–548. Elsevier Science.

Noltemeyer, Amity L., Rose Marie Ward, and Caven Mcloughlin. 2015. "Relationship Between School Suspension and Student Outcomes: A Meta-Analysis." *School Psychology Review*, 44(2): 224–240.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business and Economic Statistics*, 37(2): 187–204.

Pham, Lam D., Tuan D. Nguyen, and Matthew G. Springer. 2020. "Teacher Merit Pay: A Meta-Analysis." *American Educational Research Journal*, 58(3): 527–566.

Phipps, Aaron R., and Emily A. Wiseman. 2021. "Enacting the Rubric: Teacher Improvements in Windows of High-Stakes Observation." *Education Finance and Policy*, 16(2): 283–312.

Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 37(1): 7–63.

Prendergast, Canice. 2016. "Bureaucratic Responses." *Journal of Labor Economics*, 34(S2).

Serumaga, Brian, Dennis Ross-Degnan, Anthony J Avery, Rachel A Elliott, Sumit R Majumdar, Fang Zhang, and Stephen B Soumerai. 2011. "Effect of Pay for Performance on the Management and Outcomes of Hypertension in the United Kingdom: Interrupted Time Series Study." *British Medical Journal*, 342: d108.

Shen, Yu Chu. 2003. "The Effect of Financial Pressure on the Quality of Care in Hospitals." *Journal of Health Economics*, 22(2): 243–269.

Sliwka, Dirk. 2007. "Trust as a Signal of a Social Norm and the Hidden Costs of Incentive Schemes." *American Economic Review*, 97(3): 999–1012.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2014. "Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota." *Journal of Human Resources*, 49(4): 945–981.

Speroni, Cecilia, Alison Wellington, Paul Burkander, Hanley Chiang, Mariesa Herrmann, and Kristin Hallgren. 2020. "Do Educator Performance Incentives Help Students? Evidence from the Teacher Incentive Fund National Evaluation." *Journal of Labor Economics*, 38(3).

Springer, Matthew G., John F. Pane, Vi-Nhuan Le, Daniel F. McCaffrey, Susan Freeman Burns, Laura S. Hamilton, and Brian Stecher. 2012. "Team Pay for Performance." *Educational Evaluation and Policy Analysis*, 34(4): 367–390.

Stepner, Michael. 2013. "BINSCATTER: Stata module to generate binned scatterplots." *Statistical Software Components*, S457709.

Toch, Thomas. 2018. "A Policymaker's Playbook." FutureEd.

Vigdor, Jacob L. 2008. "Teacher Salary Bonuses in North Carolina." *Vanderbilt Peabody College Working Papers*, 2008-3.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness." The New Teacher Project.

Tables

Table 1
Summary Statistics by Year

Students	12,259	
Grade		
4th Grade	6,239	50.9%
5th Grade	6,020	49.1%
Male		
Female	6,262	51.1%
Male	5,997	48.9%
Race		
Black	8,533	69.8%
Hispanic	1,860	15.2%
White	1,355	11.1%
Other	472	3.9%
English Language Learner		
No	11,399	93.0%
Yes	860	7.0%
Special Education		
No	10,762	87.8%
Yes	1,497	12.2%
Receiving Free and Reduced-Price Lunch		
No	4,026	32.8%
Yes	8,233	67.2%
Teachers	681	
Average Years of Experience (std. dev.)	11.25	(7.99)
Has Unmonitored Time		
No	517	75.9%
Yes	164	24.1%
Average Days Unmonitored if Unmonitored (std. dev.)	5.86	(3.99)

Notes: Sample includes all 4th and 5th grade students – and their teachers – with a prior reading and math score whose teacher has at least one year of prior experience teaching and appears at least twice.

Table 2
Effect of Unmonitored Time on Student Test Outcomes
(Outcome: Standard Deviations on Standardized Tests)

	Reading			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Unmonitored Days	-0.0114***	-0.0095***	-0.0101^{+}	-0.0112**	-0.0090**	-0.0103^{+}
	(0.0030)	(0.0028)	(0.0053)	(0.0038)	(0.0035)	(0.0060)
	[0.0022 0.0130]	[0.0130 0.0319]	[0.0084 0.0246]	[0.0287 0.0541]	[0.0630 0.0975]	[0.0321 0.0586]
Teacher/School/Grade FEs	X	X	X	X	X	X
Previous Student Scores	X	X	X	X	X	X
Teacher Experience		X	X		X	X
Student Demographics		X	X		X	X
Feedback Time			X			X
Observations	12820	12820	12820	12305	12305	12305

Significance indicators: $+$ 0.1, $*$ 0.05, $**$ 0.01, $***$ 0.001

Notes: This table shows the key results for student reading and math outcomes using the specification in Equation 1. All standard errors are clustered at the teacher×year level. The 95% confidence intervals for p-values from 1,000 Randomization Inference trials are shown in brackets. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Previous student test scores include reading and math scores for both the individual student's test scores as well as the leave-out mean of their classmates' previous scores. Student demographics include gender, race, and indicators for English as a second language, special education status, and a free/reduced price lunch. Feedback Time refers to the number of business days from when a teacher receives feedback on each evaluation up to the date of standardized testing. All specifications include year, school, teacher, and grade fixed-effects.

Table 3
Placebo Test for Unmonitored Time
(Outcome: Standard Deviations on Standardized Tests)

	Reading			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
Unmonitored Placebo	0.0007	0.0008	0.0009	-0.0014	-0.0013	-0.0012
	(0.0009)	(0.0009)	(0.0009)	(0.0011)	(0.0010)	(0.0010)
Teacher/School/Grade FEs	X	X	X	X	X	X
Previous Student Scores	X	X	X	X	X	X
Teacher Experience		X	X		X	X
Student Demographics		X	X		X	X
Feedback Time			X			X
Observations	12820	12820	12820	12305	12305	12305

Significance indicators: + 0.1, ∗ 0.05, ∗∗ 0.01, ∗ ∗ ∗ 0.001

Notes: The results shown estimate the effect of a placebo on student test scores. The placebo is unmonitored days that occur *after* students have completed their standardized tests. All errors are clustered at the teacher×year level. The placebo's effect is more precisely estimated and is effectively zero. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Previous student test scores include reading and math scores for both the individual student's test scores as well as the leave-out mean of their classmates' previous scores. Student demographics include gender, race, and indicators for English as a second language, special education status, and a free/reduced price lunch. Feedback Time refers to the number of business days from when a teacher receives feedback on each evaluation up to the date of standardized testing. All specifications include year, school, teacher, and grade fixed-effects.

Table 4
Effect of Unmonitored Time on Short-Term Student Suspensions

Panel A: Total Annual Short-Term Suspensions Issued by a Teacher

	(1)	(2)	(3)
Unmonitored Days	0.00506^{+}	0.00505^{+}	0.00616*
	(0.00303)	(0.00258)	(0.00249)
School and Grade FEs	X	X	X
Student FE	X	X	X
Teacher FE		X	X
Experience			X
Average Suspensions	0.127	0.127	0.127
Teachers	255	255	255

Panel B: Daily Probability of Issuing Short-Term Suspension

	(1)	(2)	(3)
Unmonitored Day	0.0300^{+}	0.0370^{+}	0.0414^{+}
	(0.0174)	(0.0222)	(0.0215)
Teacher FE	X	X	X
School FE	X	X	X
Date FE		X	X
Teacher Experience			X
Average Daily Suspension Rate	0.0370	0.0370	0.0370
Teachers	123	123	123

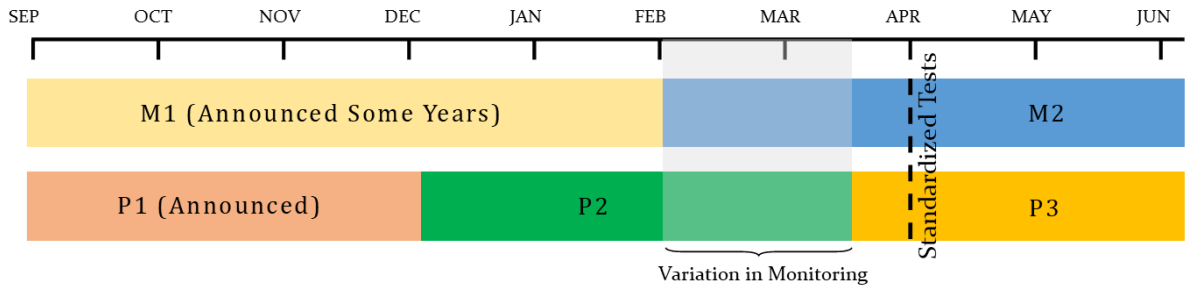
Significance indicators: $+$ 0.1, $*$ 0.05, $**$ 0.01

Panel A: Results demonstrate the effect of unmonitored time on a teacher's total short-terms suspension rates in a year. The specification is $Y_{ijts} = X_{jt}\Gamma + \beta n_{jt} + \eta_i + \delta_j + \nu_t + \phi_s + \varepsilon_{ijts}$ for student i, teacher j, year t, school s, and unmonitored days n_{jt}. X_{jt} is teacher experience and experience squared, and δ_j, η_i, ν_t, and ϕ_s are fixed effects. The coefficient reported is β, which measures the change in the average suspensions for a student for an unmonitored day relative to a monitored day. Errors are clustered at the teacher$\times$year level. As with other results, the sample includes students with a previous year's test score and a teacher with one or more years of experience. However, suspension data is only available for 2011 and 2012.

Panel B: Estimated change in daily probability of a teacher issuing a suspension for each unmonitored day. The sample is limited to days in which it is possible for teachers to have unmonitored time (roughly a 30 day window). The specification is $S_{jdts} = X_{jt}\Gamma + \beta D_{jd} + \delta_j + \nu_{td} + \phi_s + \varepsilon_{jdts}$ for teacher j, day d, year t, and school s. The outcome S_{jdts} is an indicator for whether or not teacher j issues at least one suspension on day d. Treatment is D_{jd}, which is a dummy variable with a value of 1 if day d is unmonitored. The coefficient reported is β, which indicates the average marginal change in a teacher's daily propensity to issue a suspension. X_{jt} is teacher experience and experience squared, ϕ_s is a school fixed effect, and ν_{td} is a date fixed-effect. Errors are clustered at the teacher$\times$year level. The sample includes teachers with one or more years of experience. However, reliable daily Suspension data is only available for fifth grade, which results in a smaller sample size than Panel A.

Figures

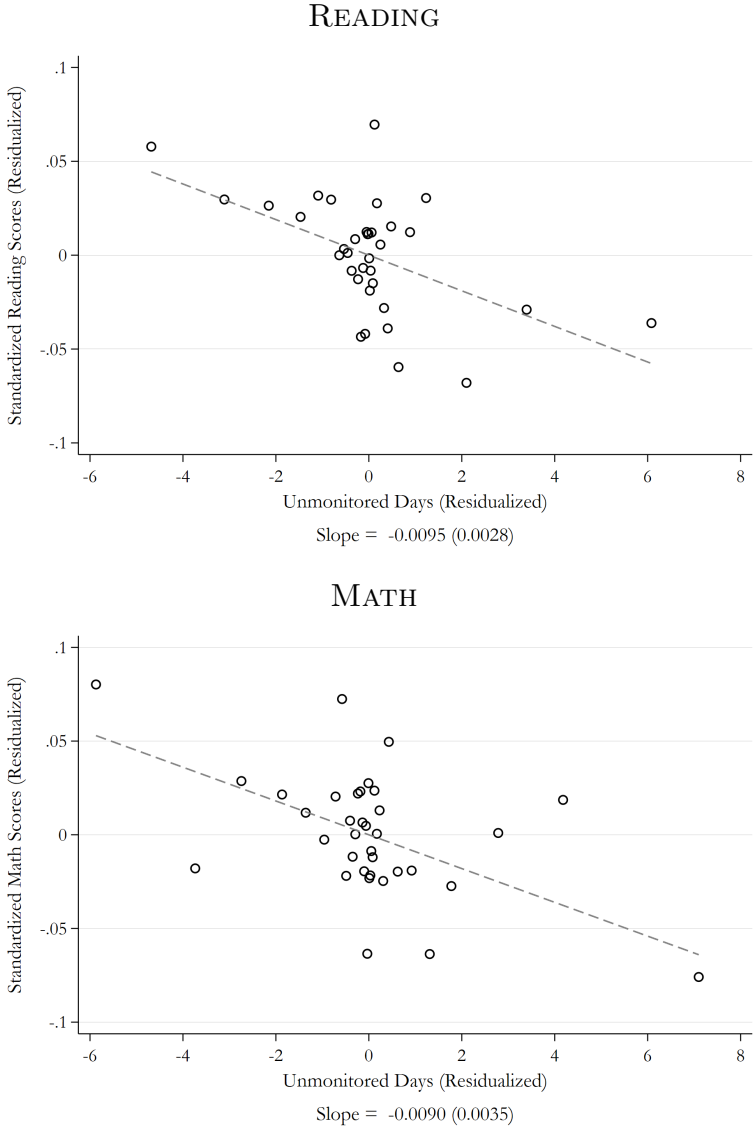
Figure 1
Depiction of each evaluation window in DCPS



Note: Color boxes indicate the time windows of the five in-class observations each teacher receives. $M1$ and $M2$ indicate the first and second External Observations, conducted by a "Master Educator." $P1$, $P2$, and $P3$ indicate the three Internal Observations, conducted either by the school principal or vice principal. One evaluation must occur within each window. Most evaluations are unannounced, meaning the teacher does not know when it happens until the observer comes to her classroom. For announced evaluations (indicated in the Figure), teachers are informed no later than the day before their evaluation. The gray shaded area indicates the time in which teachers can have variation in monitoring. Because evaluation windows in DCPS overlap, unmonitored time is defined as days in which there is no possibility of an unannounced evaluation from either evaluator. While standardized tests are shown to begin in April, they occurred in late March in some years, though never before the end of the $P2$ window.

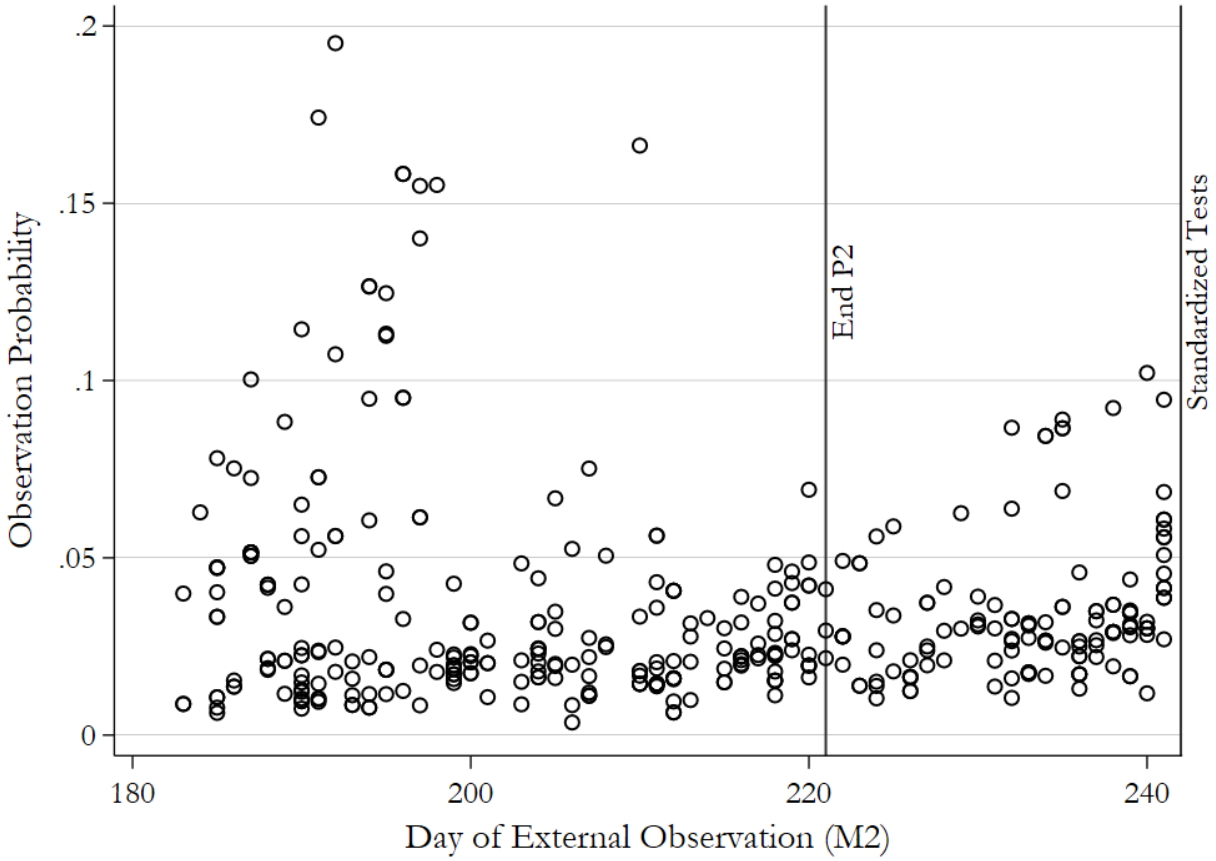
Figure 2

Relationship between teacher unmonitored time and student performance



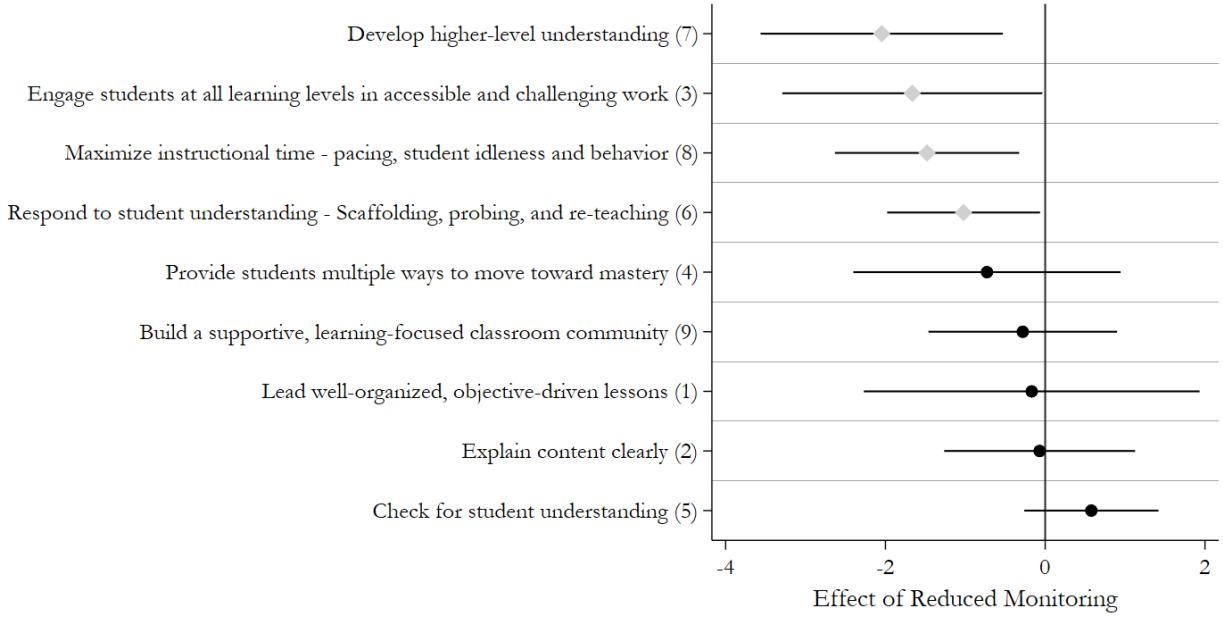
Note: Plots show the relationship between a student's standardized test scores and her teacher's number of unmonitored days. The points are residualized using the covariates of Columns 2 and 5 of Table 2 (student's previous test scores, the leave-out mean of their classmates' previous scores, teacher experience, basic student demographic variables, teacher fixed-effects, school fixed-effects, and year fixed-effects). Data are then binned into 30 quantiles. The standard errors — shown in parentheses — are clustered at the classroom level (teacher × year). Made in part using `binscatter` in Stata (Stepner, 2013).

Figure 3
Variation in Observation Probability by Date of Observation



Note: This figure plots the total probability of receiving *any* observation on the day that a teacher received her second External Observation ($M2$). If a teacher still has both her Internal Observation ($P2$) and her External Observation ($M2$) outstanding, then the joint probability will be $p = p^{P2} + p^{M2} - p^{P2} \times p^{M2}$, which explains the higher daily probabilities in the earlier days of the time period shown. Note that none of the teachers experienced a 100% probability, which occurs largely because the teachers in my sample are not all the teachers at each school, but the probabilities are calculated as shown in Equation 2 using all teachers.

Figure 4
Effect of Decreasing Monitoring on Evaluation Scores



Note: This figure reports the estimated effect of reducing monitoring on teacher pedagogical standards. The outcome is a teacher's score for each Standard, rated on a scale of 1 to 4. Lines indicate 95% confidence intervals. For context, the average score for each element is roughly 3.1, and the average standard deviation is 0.80. Coefficients are standardized such that they represent the effect of going from a 100% chance of an observation to zero. An average teacher moving from a 6% probability of being observed (the average) to an 11% probability (a one standard deviation increase) will receive 0.1 more points on Standard 7, which is roughly 12% of a standard deviation. The results are robust to multiple hypothesis adjustments, such as sharpened q-values (Benjamini, Krieger and Yekutieli, 2006). See Table A6 for a full description of the Standards. The specification is as follows: For a teacher j and standard S on the $M2$ evaluation at school s and year t, I estimate $S_{jts}^{M2} = \boldsymbol{X}_{jt}\Gamma - p_{jt}\mu + \sum_{\nu=P1,M1} S_{jt}^{\nu}\kappa^{\nu} + \boldsymbol{T}_{jt}\omega + \rho d_{jts} + \phi_s + \delta_t + \varepsilon_{jts}$. Here $\boldsymbol{X}_{jt}$ is a vector of experience indicators, ϕ_s is a school fixed-effect, and δ_t is a year fixed-effect. A teacher's performance also depends on the order in which her evaluations occur: if the $M2$ evaluation is her fourth of the year, she usually does better than if it is her third. To account for this, I include the term $\boldsymbol{T}_{jt}$, a vector of indicators for if the $M2$ evaluation was third (before $P2$), fourth (after $P2$), or fifth (after $P3$). S_{iy}^{q} are scores on Standard S for $\nu = P1, M1$, which are evaluations that must have already occurred by the time of her $M2$ evaluation. The term p_{jt} is measured as the probability of receiving *any* evaluation on the day of the teacher's $M2$ evaluation, and μ is the coefficient of interest. Note that if a teacher still has both her P2 and M2 evaluations outstanding, then the joint probability will be $p_{jt} = p^{P2} + p^{M2} - p^{P2} \times p^{M2}$. Lastly, I include a time trend for the day of the evaluation, d_{jts}.

Appendix

Table A1
Balance Check across Students
(Outcome: Unmonitored Days)

	(1)	(2)	(3)
Male	0.0443	0.0360	0.0435
	(0.0397)	(0.0396)	(0.0395)
Black	0.0242	-0.0129	0.0275
	(0.0813)	(0.0844)	(0.0842)
Hispanic	0.223	0.160	0.186
	(0.154)	(0.155)	(0.158)
Other Race	0.0619	0.0350	0.0412
	(0.114)	(0.117)	(0.117)
English Learner		0.0847	0.140
		(0.0847)	(0.0894)
Special Education		0.118	0.179
		(0.141)	(0.143)
Free or Reduced Price Lunch		0.154	0.166
		(0.102)	(0.103)
Previous Reading			0.0530
			(0.0466)
Previous Math			0.0262
			(0.0528)
School FE	X	X	X
Year FE	X	X	X
Subject FE	X	X	X
Observations	16119	16119	16119
F	0.657	0.821	0.968
p	0.685	0.597	0.474

Significance indicators: + 0.1, ∗ 0.05, ∗∗ 0.01, ∗ ∗ ∗ 0.001
Notes: This table reports the results of regressions aimed at checking if treatment is systematically targeted at certain students based on observable characteristics. Errors are clustered at the teacher-year level. The outcome is number of unmonitored days for a student's teacher. There do not appear to be any statistically significant coefficients. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis. Because treatment occurs at the classroom level, all student-teacher combinations are preserved across both subjects, which is why observations are greater than in Table 1.

Table A2
Balance Check across Teachers
(Outcome: Unmonitored Days)

	(1)	(2)	(3)	(4)
Teacher Experience	-0.0264	-0.0232	0.00999	0.00869
	(0.0625)	(0.0618)	(0.0933)	(0.0946)
Teacher Experience Squared	0.00137	0.00132	0.000142	0.000192
	(0.00219)	(0.00217)	(0.00328)	(0.00332)
Black		-0.0572	-0.0477	-0.0410
		(0.307)	(0.416)	(0.424)
Hispanic		-0.156	2.822	2.783
		(0.997)	(2.163)	(2.193)
Asian		1.070	0.821	0.804
		(1.165)	(1.132)	(1.146)
Lagged Evaluation Score			-0.298	-0.332
			(0.488)	(0.509)
Highly Effective (t-1)				0.123
				(0.496)
Minimally Effective (t-1)				-0.0497
				(0.635)
School FE	X	X	X	X
Year FE	X	X	X	X
Observations	678	678	399	399
F	0.393	0.316	0.470	0.372
p	0.675	0.903	0.830	0.935

Significance indicators: + 0.1, ∗ 0.05, ∗∗ 0.01, ∗ ∗ ∗ 0.001

Notes: This table shows results for regressions checking if treatment is systematically targeted at certain teachers based on observable characteristics. Errors are clustered at the teacher-year level. The outcome is number of unmonitored days for a teacher. There do not appear to be any statistically significant coefficients. The F-Test is a joint hypothesis test of whether all coefficients are zero (not including the fixed-effects). These too appear to fail to reject the null hypothesis. Columns 4 and 5 show results including previous performance as part of the IMPACT program, which limits the sample to the last two years in the sample.

Table A3
Estimated Required Proportional Selection on Unobservables

R-Squared	Math	Reading
0.70		27.46
0.75	7.54	6.81
0.80	3.54	3.89
0.85	2.31	2.72
0.90	1.72	2.09
0.95	1.37	1.70

These are estimates of the required selection on unobservables necessary to account for the effect sizes observed (Altonji, Elder and Taber, 2005). For an assumed maximum R^2 value (Column 1), results show the required amount of selection on unobservables to achieve the observed coefficients as a fraction of how much observables explain the outcome. The recommendation from Altonji, Elder and Taber (2005) is that if 100% or more selection on unobservables is required then the results are considered robust. Values are calculated using `psaCalc` (Oster, 2019).

Table A4
Results under Different Specifications of Teacher Experience
(Outcome: Standard Deviations on Standardized Tests)

	Reading				Math			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Unmonitored Days	-0.0103***	-0.0094**	-0.0103***	-0.0122***	-0.0085*	-0.0085*	-0.0085*	-0.0065*
	(0.0029)	(0.0030)	(0.0028)	(0.0035)	(0.0037)	(0.0038)	(0.0037)	(0.0031)
Previous Student Scores	X	X	X	X	X	X	X	X
Teacher Experience	X	X	X	X	X	X	X	X
Student Demographics	X	X	X	X	X	X	X	X
Observations	12820	12820	12820	12820	12305	12305	12305	12305

Significance indicators: + 0.1, ∗ 0.05, ∗∗ 0.01, ∗ ∗ ∗ 0.001

Notes: This table demonstrates the key results for student math outcomes where each column uses a different measurement of teacher experience. All standard errors are clustered at the teacher-year level. Column (1): Experience is a continuous variable capped at 15 years; both linear and squared terms included. Column (2): Experience is a categorical variable capped at 15 years. Column (3): Same as Column (1) but not capped. Column (4): Same as Column (2) but not capped.

Table A5
Results by Grade
(Outcome: Standard Deviations on Standardized Tests)

	Reading			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
4th Grade × Unmonitored Days	-0.0096^{+}	-0.0098^{+}	-0.0107	-0.0169***	-0.0143***	-0.0157*
	(0.0055)	(0.0054)	(0.0070)	(0.0044)	(0.0042)	(0.0070)
	[0.0833]	[0.0673]	[0.1307]	[0.0001]	[0.0006]	[0.0254]
5th Grade × Unmonitored Days	-0.0127***	-0.0093***	-0.0097^{+}	-0.0070	-0.0051	-0.0065
	(0.0030)	(0.0025)	(0.0051)	(0.0049)	(0.0045)	(0.0062)
	[0.0000]	[0.0003]	[0.0571]	[0.1599]	[0.2597]	[0.2904]
Previous Student Scores	X	X	X	X	X	X
Teacher Experience		X	X		X	X
Student Demographics		X	X		X	X
Feedback Time			X			X
Observations	12820	12820	12820	12305	12305	12305

Significance indicators: + 0.1, ∗ 0.05, ∗∗ 0.01, ∗ ∗ ∗ 0.001

Notes: This table demonstrates the key results for student math outcomes broken out by grade. All errors are clustered at the teacher-year level. Sample includes students with a previous year's test score and a teacher with one or more years of experience. Effects of unmonitored time are more concentrated in fourth grade than in fifth. The fourth grade math curriculum covers fractions and operations on fractions, while the fifth grade math curriculum covers decimals and their operations.

Table A6

Description of the Components of the Teaching and Learning Framework

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 1 *Lead well-organized, objective-driven lessons*	*Lesson Organization* The lesson is well-organized: All parts of the lesson are connected to each other and aligned to the objective, and each part significantly moves all students toward mastery of the objective. *Lesson Objective* The objective of the lesson is clear to students and conveys what students are learning and what they will be able to do as a result of the lesson. Students also can authentically explain what they are learning and doing beyond simply repeating the stated or posted objective. *Objective Importance* Students understand the importance of the objective. Students also can authentically explain why what they are learning and doing is important, beyond simply repeating the teachers' explanation.
Teach 2 *Explain content clearly*	*Clear, Coherent Delivery* Explanations of content are clear and coherent, and they build student understanding of content. The teacher might provide explanations through direct verbal or written delivery, modeling or demonstrations, think-alouds, visuals, or questioning. Explanations of content also are delivered in as direct and efficient a manner as possible. *Academic Language* The teacher gives clear, precise definitions and uses a broad vocabulary that includes specific academic language and words that may be unfamiliar to students when it is appropriate to do so. Students also demonstrate through their verbal or written responses that they are internalizing academic vocabulary. *Emphasize Key Points* The teacher emphasizes key points when necessary, such that students understand the main ideas of the content. Students also can authentically explain the main ideas of the content beyond simply repeating back the teacher's explanations. *Student Understanding* Students show that they understand the explanations. When appropriate, concepts also are explained in a way that actively and effectively involves students in the learning process. For example, students have opportunities to explain concepts to each other. *Connections* The teacher makes connections with students' prior knowledge, students' experiences and interests, other content areas, or current events to effectively build student understanding of content.

Table A6
(Continued)

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 3 *Engage students at all learning levels in accessible and challenging work*	*Accessibility* The teacher makes the lesson accessible to all students. There is evidence that the teacher knows each student's level and ensures that the lesson meets all students where they are. *Challenge* The teacher makes the lesson challenging to all students. There is evidence that the teacher knows each student's level and ensures that the lesson pushes all students forward from where they are. *Balance* There is an appropriate balance between teacher-directed and student-centered learning during the lesson, such that students have adequate opportunities to meaningfully practice, apply, and demonstrate what they are learning.
Teach 4 *Provide students multiple ways to move toward mastery*	*Multiple Ways Toward Mastery* The teacher provides students multiple ways to engage with content, and all ways move students toward mastery of lesson content. During the lesson, students are also developing deep understanding of the content. *Appropriateness for Students* The ways the teacher provides include learning styles or modalities that are appropriate to students' needs; all students respond positively and are actively involved in the work.
Teach 5 *Check for student understanding*	*Key Moments* The teacher checks for understanding of content at all key moments. *Accurate Pulse* The teacher always gets an accurate "pulse" at key moments by using one or more checks that gather information about the depth of understanding for a range of students, when appropriate.

Table A6
(Continued)

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 6 *Respond to student understanding*	*Scaffolding* When students demonstrate misunderstandings or partial understandings, the teacher always uses effective scaffolding techniques that enable students to construct their own understandings, when appropriate. *Re-Teaching* The teacher always re-teaches effectively when appropriate, such as in cases in which most of the class demonstrates a misunderstanding or an individual student demonstrates a significant misunderstanding. The teacher also anticipates common misunderstandings (e.g., by offering a misunderstanding as a correct answer to see how students respond) or recognizes a student response as a common misunderstanding and shares it with the class to lead all students to a more complete understanding. *Probing* The teacher always probes students' correct responses, when appropriate, to ensure student understanding.
Teach 7 *Develop higher-level understanding through effective questioning*	*Questions and Tasks* The teacher asks questions that push all students' thinking; when appropriate, the teacher also poses tasks that are increasingly complex that develop all students' higher-level understanding. *Support* After posing a question or task, the teacher always uses appropriate strategies to ensure that students move toward higher-level understanding. *Meaningful Response* Almost all students answer questions of complete complex tasks with meaningful responses that demonstrate movement toward higher-level understanding, showing that they are accustomed to being asked these kinds of questions.

Table A6
(Continued)

STANDARD	DESCRIPTION OF HIGHLY EFFECTIVE TEACHING
Teach 8 *Maximize instructional time*	*Routines, Procedures, and Transitions* Routines, procedures, and transitions are orderly, efficient, and systematic with minimal prompting from the teacher' students know their responsibilities and some students share responsibility for leading the operations and routines in the classroom. *Student Idleness* Students always have something meaningful to do. Lesson pacing is also student-directed or individualized, when appropriate. *Lesson Pacing* The teacher spends an appropriate amount of time on each part of the lesson. *Student Behavior* Inappropriate or off-task student behavior never interrupts or delays the lesson, either because no such behavior occurs or because when such behavior occurs the teacher efficiently addresses it.
Teach 9 *Build a supportive, learning-focused classroom community*	*Investment* Students are invested in their work and value academic success. Students are also invested in the success of their peers. For example, students can be seen helping each other or showing interest in other students' work without prompting from the teacher. *Risk-Taking* The classroom environment is safe for students, such that students are willing to take on challenges and risk failure. For example, students are eager to ask questions, feel comfortable asking the teacher for help, feel comfortable engaging in constructive feedback with their classmates, and do not respond negatively when a peer answers a question incorrectly. *Respect* Students are always respectful of the teacher and their peers. For example, students listen and do not interrupt when their peers ask or answer questions. *Reinforcement* The teacher meaningfully reinforces positive behavior and good academic work, when appropriate. Students also give unsolicited praise or encouragement to their peers, when appropriate. *Rapport* The teacher has a positive rapport with students, as demonstrated by displays of positive affect, evidence of relationship building, and expressions of interest in students' thoughts and opinions. There is also evidence that the teacher has strong, individualized relationships with some students in the class.