NBER WORKING PAPER SERIES

INFORMATION AND EMPLOYEE EVALUATION: EVIDENCE FROM A RANDOMIZED INTERVENTION IN PUBLIC SCHOOLS

Jonah E. Rockoff
Douglas O. Staiger
Thomas J. Kane
Eric S. Taylor

Working Paper 16240
http://www.nber.org/papers/w16240

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2010

We thank seminar participants at Harvard University (Economics Department, Kennedy School, and School of Education), Columbia University (Business School and Teachers College), UC Davis, University of Oregon, London School of Economics, the Research Institute of Industrial Economics (IFN), and the University of Tel Aviv for many helpful comments and suggestions. Financial support was provided by the Fund for Public Schools. All opinions expressed herein represent those of the authors and not necessarily those of the New York City Department of Education. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools

Jonah E. Rockoff, Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor

NBER Working Paper No. 16240
July 2010
JEL No. D83,D86,H75,I21,J45

ABSTRACT

The evidence that productivity varies greatly across teachers has given rise to the idea that student achievement data should be included in performance evaluation, despite limited empirical evidence on subjective evaluation or the use of objective performance measures in U.S. public schools. In this paper, we examine the results of a randomized pilot program in which school principals were provided with estimates of the performance of individual teachers in raising their students' test scores in math and English. Our analysis establishes several facts consistent with a simple Bayesian learning model of employee evaluation in the presence of imperfect information. First, objective teacher performance estimates based on student data and principals' prior beliefs are positively correlated, and the strength of this relationship rises with the precision of the objective estimates and the precision of subjective priors. Second, principals who are provided with objective performance data incorporate this information into their posterior beliefs, and do so to a greater extent when the data are more precise and when their priors are less precise. Moreover, after the provision of performance data, the probability of job separation rises for teachers with low performance estimates, and, in line with this change in attrition patterns, student achievement exhibits small improvements the following year. These results suggest that objective performance data provides useful information to principals in constructing employee evaluations and using these evaluations to improve productivity.

Jonah E. Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

Thomas J. Kane
Harvard Graduate School of Education
Gutman Library, Room 455
Appian Way
Cambridge, MA 02138
and NBER
kaneto@gse.harvard.edu

Douglas O. Staiger
Dartmouth College
Department of Economics
HB6106, 301 Rockefeller Hall
Hanover, NH 03755-3514
and NBER
douglas.staiger@dartmouth.edu

Eric S. Taylor
Harvard Graduate School of Education
50 Church St., 4th Floor
Cambridge, MA 02138
eric_taylor@gse.harvard.edu

Economists interested in educational production have increasingly focused on issues surrounding teachers. There is considerable evidence that productivity varies greatly across teachers (Rockoff (2004), Rivkin et al. (2005), Aaronson et al. (2007)), and two recent papers (Gordon et al. (2006), Kane and Staiger (2008)) argue in favor of using estimates of a teacher's "value-added"—the teacher's contribution to changes in student achievement, as measured by standardized tests—to evaluate teachers.[1] Motivated in part by this research, the federal government is encouraging states and school districts to use student achievement growth in measures of teacher effectiveness and to implement policies to "recruit, develop, reward, and retain effective teachers" as part of the incentives built into its $4.3 billion Race to the Top Fund.

How should estimates of teacher performance based on student outcomes be used to evaluate teachers? Some important insights are provided by the vast body of research in economics on performance evaluation, incentives in employment contracts, and personnel decisions (see reviews by Prendergast (1999), Gibbons (1998, 2005)). First, teachers are likely to be "motivated agents" who work for "idealistic reasons or because they enjoy working with children" (Dixit (2002)) and may not require high-powered incentives (see also Besley and Ghatak (2005)). Thus, it is reasonable to believe that disparities in teacher effectiveness may have more to do with persistent variation in skills than exertion of effort.[2]

Second, teaching is a multi-dimensional task with multiple goals. In this setting, basing employee evaluations on simple objective performance measures may lead to dysfunctional behaviors (Holmstrom and Milgrom (1991)), like cheating on standardized tests (Jacob and

[1] There is still considerable discussion surrounding the validity of the assumptions underlying value-added estimates (Todd and Wolpin (2003), Rothstein (2009)). While our paper does not focus on these concerns, we refer the reader to the analyses and discussions in Harris and Sass (2006), Goldhaber and Hansen (2009), Koedel and Betts (2009), and Staiger and Rockoff (2010).

[2] This view may not be entirely appropriate in describing public schools in developing countries, where teachers often do not show up to work or are found not to be teaching when independently observed (Chaudhury et al. 2006).

Levitt (2003)).[3] In order to avoid this problem, principals can link personnel decisions to holistic, subjective evaluations that capture a teacher's overall contribution to the school.

Third, even when outcomes for the employee are determined by holistic evaluations, the presence of an objective measure of performance that is verifiable to a third party can be useful (Baker, Gibbons, and Murphy (1994)). Such a measure may be valuable to principals who possess imperfect information on employee performance. Moreover, even if principals can easily observe performance, objective measures allow them to dismiss unproductive workers without gaining bad reputations and assure workers they will be rewarded if they perform well.

There is relatively little empirical work by economists on subjective employee evaluation, and much of it is descriptive. For example, studies have documented the prevalence (or lack thereof) of subjective and objective performance evaluation across occupations (e.g., Baker et al. (1988), Murphy and Oyer (2001)), managers' reluctance to differentiate among employees (e.g., Medoff and Abraham (1980), Murphy (1992)), and managerial bias in the evaluation process (see Prendergast and Topel (1993)). In the context of public schools, there is some work showing that school principals' opinions of teacher effectiveness are correlated with estimates based on student test scores (Murnane (1975), Jacob and Lefgren (2008), Harris and Sass (2008)), but principals, like private sector managers, are reluctant to give teachers poor evaluations (Weisberg et al. (2009)) or officially dismiss them (Jacob (2007, 2010)).

In this paper, we examine how managers develop and use subjective evaluations, and how these processes are affected by the presence of objective performance data. We do so in the context of a pilot program conducted by the New York City Department of Education (hereafter

[3] Teachers are also highly unionized in the U.S. and oppose linking pay directly to objective performance measures, providing few examples of how such contracts work in practice (see Ballou (2001), Turner and Goodman (2009)). For research on teacher pay-for-performance contracts outside of the U.S., see Lavy (2002, 2009), Muralidharan and Sundaraman (2008).

the DOE) during the school year 2007-2008. School principals were randomly selected from a group of volunteers to receive performance measures (i.e., estimates of "value-added") for teachers at their schools, as well as training on the methodology used to construct the estimates. The remaining volunteer principals serve as a control group, allowing us to draw causal inferences regarding the impact of distributing the objective performance data.

We begin by laying out a basic model of employee evaluation in which principals use imperfect information to learn about teacher effectiveness, i.e., a Bayesian learning model along the lines of Jovanovic (1979). This provides us with several testable predictions for the relationship between teacher performance data and principals' prior and posterior beliefs regarding teacher effectiveness, all of which are borne out empirically. First, there is a strong relationship between value-added and principals' baseline evaluations of teacher effectiveness, consistent with the earlier work cited above. Furthermore, as the model would predict, this relationship is stronger when value-added estimates and principals' priors are relatively more precise. Second, principals who are provided with objective performance data incorporate this information into their posterior beliefs, and do so to a greater extent when the data are more precise and when their priors are less precise.

We also investigate several additional potential effects of providing objective performance data, though these are outside the scope of our simple theoretical framework. First, we find no evidence that principals responded to the receipt of objective performance data by spending less time collecting information through classroom observation. Second, we find that the objective performance estimates changed patterns of teacher turnover—teachers with low performance estimates were more likely to exit their schools after this information was provided to principals. These differences in turnover patterns imply small improvements in the

productivity of the teachers in the treatment schools, and we find changes in achievement in line with these expectations, particularly with respect to math test scores. These findings contribute to our understanding of how managers incorporate information on worker performance into evaluation and personnel decisions. They also suggest that the provision of objective performance data on teachers to school principals may be a useful tool for the improvement of school quality.

In Section 2 we describe the pilot, compare treatment and control groups, present descriptive statistics from baseline and follow-up surveys, and describe other sources of data. In Section 3, we discuss our basic theoretical framework to guide our analysis of the impact of the treatment. We examine the relationship between value-added estimates and principals' prior beliefs regarding teacher performance in Section 4, and we estimate impacts of the provision of performance data on a variety of outcomes in Sections 5 and 6. Section 7 concludes.

2. The Teacher Data Initiative

The goals of the DOE pilot program were to develop the internal capacity to estimate teacher value-added, design and disseminate reports to principals, and train principals to understand the methodology and the reports. Several factors motivated the program. First, DOE officials believed that many principals had limited capacity to access and analyze data, had a small scope of comparisons within their schools, and may have lacked training in teacher evaluation or even relied on personal biases in their evaluations. Thus, value-added estimates could provide them with new and potentially useful information. In addition, it was felt that principals would have local knowledge regarding student assignment, and that this could help them interpret estimates of teacher value-added in the context of any peculiar matching of students and teachers. Third, recent changes to their management and accountability systems

had given principals considerable decision-making power and responsibility, and the value-added reports were seen as a means of empowering principals without imposing choices upon them.[4]

It is unclear whether providing objective teacher performance information to principals has any impact on their beliefs or actions. Principals' subjective evaluations may already account for all of the meaningful variation in value-added estimates, or principals might place little faith in their accuracy and give them little or no weight when forming opinions or making decisions. In order to understand the impact of this initiative on principals, teachers, and students, the DOE piloted the program in a randomized control trial.[5]

In early summer 2007, the DOE identified principals from the school year 2006-2007 who were expected to be principal at the same school in the coming year and whose school contained any of the grades 4 to 8.[6] These principals received an e-mail with basic information about the initiative, a link to a web site with background on value-added, an invitation to attend one of three presentations on the initiative conducted at different locations in the city during the summer, and a link to respond if they were interested in participating in the pilot program.

Of the roughly 1,000 principals who were sent an invitation, 335 principals expressed interest in becoming part of the program and were sent a baseline survey on August 8, 2007.

---

[4] Principals in the DOE had received greater power to allocate financial resources within their schools and purchase services as they saw fit, rather than receive services through central administration. Also, under the DOE accountability system, principals could earn bonuses of up to $25,000 and could be removed for poor performance.

[5] Most of the logistical work for the pilot program was conducted by the Battelle Memorial Institute, under contract with the DOE. Battelle performed survey data collection, the provision of professional development to principals, estimation of teacher value-added and preparation of the value-added reports. We do not discuss the value-added estimation methodology here, but details were provided in a guide to participating principals and available upon request from the authors. The methodology uses linear regression to predict student test scores based on prior information, and averages residuals at the teacher level. Because the standardized tests are taken prior to the end of the school year, teachers' value-added estimates are based partially on test score performance of students in the following year. This partial weighting on next year's performance is not done for teachers of 8th grade (due to a lack of 9th grade test data) or teachers observed for the first time in the most recent year of data.

[6] The DOE also excluded middle schools with known problems in the data linking teachers to students. Schools not serving grades 4 to 8 were excluded because students in New York are only tested annually in grades 3 through 8, and the methodology used to estimate value-added relied on controls for prior test scores.

They were told that a randomly selected subset of principals who completed the survey by September 21st would be provided value-added reports and training. 223 principals completed the survey, 112 were selected into the treatment group, and the remaining 111 schools constitute the control group. Selection was done by assigning a random number to each school, sorting by number within K-8, Middle, and Elementary schools, and selecting the first 12, 27, and 73 schools in each group, respectively.

We compare the average characteristics of treatment and control principals (schools) at baseline and test for statistically significant differences (Table 1, left side). The treatment and control groups are very similar with regard to enrollment, principals' characteristics (work experience and demographics), students' characteristics (poverty, demographics, and participation in special programs), and teachers' opinions of school environment from a citywide survey from spring 2007. Thus, the randomization successfully created comparable groups.

Treatment principals were invited to attend a three hour training session to learn about the methodology for estimating value-added and to receive reports on their teachers. Sessions were held over several evenings in December 2007. The first two hours focused on the statistics behind value-added estimation, a walk-through of a sample value-added report, and discussion of uses of information on value-added. Principals then received their teachers' reports, and the remaining hour was devoted to answering principals' questions. Of the treatment principals, 71 attended a session in person, while 24 participated in an online session (similar to a conference call, but with a presentation viewed via computer), and 1 viewed a session on video.[7] The DOE did not distribute value-added reports to 16 treatment principals who did not attend or view a training session, but these principals are included as part of the treatment group in our analysis.

[7] The principals who attended the live/online sessions completed a short survey instrument to provide feedback to the DOE. 95 percent of principals attending reported that the session was a valuable use of their time, and over 80 percent reported that they understood the 'teacher value-added' metric and could understand the reports.

A sample value-added data report is included as Appendix Figure 1. Value-added measures were based only on students teachers had taught at their current schools, and were calculated separately by grade level. Multiple reports were distributed for 32 percent of middle school teachers and 14 percent of elementary school teachers who had taught multiple grades. A report contained four different value-added measures in each subject (math and/or English). Each teacher's performance was compared to teachers citywide and to teachers with similar levels of experience working in classrooms with similar student composition; for each type of comparison, value-added was measured based on up to three years of prior data and on just the prior year. A 95 percent confidence interval was reported for all estimates based on the estimated variance of the value-added measure.[8]

A follow-up survey was sent to treatment and control principals in late May 2008, to be completed by mid-July. Two treatment principals and one control principal asked to be removed from the study and were not sent the follow-up survey. All other principals were sent the survey, including those in the treatment group that did not attend professional development and did not receive value-added reports. Of the 110 treatment principals invited to take the follow-up survey, 84 began the survey, 81 completed the teacher evaluations, and 79 completed all survey questions; of the 110 control principals invited, 94 began the survey, 93 completed the teacher evaluations, and 91 completed all survey questions. The difference in response rates between the groups is partly driven by treatment principals who did not receive reports; only 5 of these 16 principals completed the follow-up survey. Nevertheless, one might be concerned with comparisons of only those treatment and control principals responding to the follow-up survey.

[8] The report also presents value-added estimates specific to student subgroups (e.g., English Language Learners, Special Education students, students who scored in the bottom third of the school's distribution in the prior year). In our analysis, we restrict our attention to the value-added estimates based on all students.

To address this concern, we limit our sample to follow-up survey respondents and compare treatment and control principals on the same characteristics as the baseline sample (Table 1, right side). Again, we find no significant differences between the two groups. We have also made these comparisons for treatment and control principals who completed the entire follow-up survey—some started the survey but did not finish—and, similarly, find no significant differences. Although we can only test for differences in observables, these results support the idea that treatment and control principals who responded to the follow-up survey were comparable. Of course, for our analyses of non-survey outcomes (e.g., teacher turnover, student test scores), we include all participating schools, regardless of survey response.

In order to participate, principals had to volunteer and complete the baseline survey, and only about one quarter of eligible principals did so.[9] While participation may have been based on exogenous factors, like the principal's availability to attend one of the information sessions conducted in summer 2007, it is possible that principals that did not volunteer may have been unwilling to use student data to evaluate teachers, or were so steeped in data analysis that the reports would have been superfluous. Conversely, reluctant principals may have been unfamiliar with data analysis and had the most potential to learn and (possibly) change their views. If participating principals were significantly different from non-participants, our findings might have limited external validity. Using observable characteristics, we find little evidence that participating principals (schools) were substantially different from those serving students in grades 4-8 citywide (Table 2). We find a few statistically significant differences, but even these appear small. For example, compared with the citywide population, sample principals had slightly more teaching experience (7.2 vs. 6.2 years) and were slightly more likely to be female

[9] This level of participation is not unusually low. For example, roughly one in five eligible schools volunteered to participate in Tennessee's well-known class size reduction experiment, Project STAR.

(79 vs. 73 percent). Moreover, we find no significant differences in their teachers' views on issues such as frequency of classroom observations, the feedback teachers receive from the principal, how much the principal prioritizes teaching, or the use of student data in instruction.

Several additional factors may be important in interpreting the effects we find for the value-added reports in this setting. First, in the fall of 2007, the DOE launched an accountability system which, for elementary and middle schools, was built around student performance on math and English exams taken in grades 4-8 (see Rockoff and Turner, forthcoming). Principals could earn bonuses of up to $25,000 for doing well and could be removed for doing poorly under this system, and thus had incentives to focus on teacher performance in these grades and subjects. On the other hand, the broader political climate may have not been conducive to principals' use of the value-added reports. The teachers' union was briefed on the pilot study in the summer of 2007, did not support it, and filed a formal grievance on the matter at the start of the school year. The DOE, partly in response, advised treatment principals when they received the reports that they were *not* to be used for formal teacher evaluation during the pilot year. Though the identities of participating principals were held confidential, the pilot's existence was well known (e.g., it made the front page of the New York Times in January, 2008) and at the request of teachers unions, the New York State legislature amended state law in April 2008 so that teachers could "not be granted or denied tenure based on student performance data."

Despite union opposition, the DOE expanded dissemination of teacher data reports to all elementary and middle schools in December 2008. Treatment principals received an updated set of reports, while control principals received reports for the first time. This should not affect the interpretation of the vast majority of our analysis, since our follow-up survey occurred before the expanded dissemination of teacher data reports and, at that time, no one knew if the program

would be continued or expanded. However, we will analyze the impact of the pilot on test scores in the spring of 2009. By this time both treatment and control principals had received value-added reports, though control school principals had received them only a few months prior and likely had little time to act on this information in ways that would impact spring test scores.

2.1 Baseline and Follow-up Surveys

First and foremost, the baseline and follow-up surveys asked principals to evaluate the performance of their teachers. Specifically, principals were given a list of the teachers in their schools who (based on DOE records) taught math and/or English to students in grades four through eight. Principals were asked to confirm that each teacher had indeed taught in these areas (nearly all teachers were confirmed), and then asked to evaluate each teacher "overall," and then specifically "in terms of raising student achievement" in math or English (or both for teachers who taught both subjects). Principals were directed to compare each teacher to all "teachers [they] have known who taught the same grade/subject," not just to teachers within their school or with similar levels of experience. Evaluations were made on a six point scale: Exceptional (top 5%), Very Good (76-95th percentile), Good (51-75th), Fair (26-50th), Poor (6-25th percentile), or Very Poor (bottom 5%).

Descriptive statistics on teachers at baseline are presented in Table 3 (left side). On a 1-6 scale, the mean overall evaluation was 4.3, suggesting that principals were more generous with their evaluations than indicated by the anchor percentages at each point on the scale. Indeed, over three quarters of the teachers received an overall evaluation indicating above median performance (Good, Very Good, or Exceptional), and subject specific evaluations in math and

English were only slightly lower on average.[10] Nevertheless, the standard deviation of evaluations of about 1.1 indicates significant variation in principals' priors. In the official evaluation system for teachers in the DOE, by contrast, less than 2 percent of teachers are evaluated as "Unsatisfactory" per year, and the rest are deemed "Satisfactory."

Notably, the correlation between subject specific evaluations in math and English is quite high (0.91) and the correlations between subject specific and overall evaluations was only slightly lower (0.87 for math and 0.88 for English). Nevertheless, the imperfect agreement among these evaluations suggests room for nuance in how principals evaluate performance. We explore this issue further in our empirical analysis (Section 4).

In addition to evaluations, principals were asked to provide the number of formal classroom observations and total classroom observations they had made of each teacher during the prior school year.[11] This allows us to examine how principals allocate their observational time across teachers and whether this allocation is correlated with value-added measures. We also can examine whether the value-added reports affected principals' time allocation. On average, principals reported doing 2.2 formal observations and 6.4 total observations of each teacher during the prior school year.[12] Principals reported no formal observations for only 3 percent of teachers, and no total observations for just 0.4 percent of teachers, but there was considerable variance across principals in the frequency of observation. 20 percent of principals reported doing no more than four total observations of *any* teacher in their school, while 15 percent reported making "more than ten" total observations of *every* teacher in their school.

[10] The percentages given an overall evaluation in each category (from Exceptional to Very Poor) were 13.7, 33.6, 30.2, 17.2, 3.8, and 1.6, respectively.

[11] Formal observations are part of the official DOE teacher evaluation system. Untenured teachers in elementary and middle schools in the DOE must be formally observed at least twice per year. Tenured teachers must be formally observed once, or can be evaluated via a "performance option" which entails setting goals at the start of the school year and submitting a report to the principal at the end of the year on how those goals were met.

[12] To calculate these averages, we coded teachers with "more than ten" observations as having a value of 11. We also use this coding when examining the relationship between value-added measures and observations (Table 6).

The second part of the baseline survey asked principals about measuring teacher effectiveness, their use of student test score data, and other issues related to teachers. For example, they were asked about how they assess teachers (outside of classroom observation), their views of the potential benefits and problems with measuring teacher performance using student test scores, and their ability to attract and retain high quality teachers in their schools.

Descriptive statistics on principals' responses to the second part of the baseline survey are provided in Appendix Table A1.[13] It is worth noting that 80 percent of treatment and control principals stated that they "regularly compare differences in average student growth in test scores for different teachers" and over three quarters of each group said that "student performance on state tests" was one of the top two factors (beyond classroom observation) they considered when assessing the overall effectiveness of their teachers. Principals' biggest concern "about using average student growth in test scores to inform evaluations of teachers and schools" was that "teachers affect important student outcomes, such as behavior, self-esteem, and intellectual curiosity, in ways that cannot be measured with standardized tests." Thus, principals in both groups were already using student test scores to decide which of their teachers were least and most effective, though they value other teaching skills that may be missed by standardized tests.

It is also noteworthy that more than three quarters of principals "strongly agree" with the statement "I know who the more and less effective teachers are in my school," and over 80 percent "strongly agree" or "agree" with the statement "I am able to retain the most effective teachers in my school." In contrast, less than half the principals "strongly agree" or "agree" with the statement "I am able to dismiss ineffective teachers in my school," and only a quarter did so with the statement "anyone can learn to be an effective teacher." Thus, it appears that these

[13] Appendix Table A1 also tests for differences in treatment and control principals' responses to the baseline survey. Average responses across groups were statistically different at the ten percent level on just 1 of 33 items.

principals have fairly strong prior beliefs on which teachers are effective and ineffective, they are concerned about the continued presence of ineffective teachers in their schools, and they do not believe that additional training can lead all of their teachers to become effective.

In the follow-up survey, teacher evaluations were followed by questions about the evaluation process and the importance of various issues when using students' standardized test scores to assess individual teachers. Treatment principals were also asked about their confidence that the value-added calculations addressed these issues, their opinions on the usefulness of the reports, and whether they shared reports with administrators and/or teachers in their school.

Descriptive statistics on principals' responses to the second part of the follow-up survey are provided in Appendix Tables A2 and A3. When asked for the top four factors (other than observation) influencing their evaluation of teachers, the most frequent item (cited by 96 percent of control principals and 93 percent of treatment principals) was student performance on state standardized tests. Thus, it is fairly certain that these principals would have used test outcomes to evaluate teacher effectiveness even without value-added reports. However, a higher share of treatment principals (46 vs. 31 percent) marked state standardized tests as the "top factor" used for evaluation other than observation. Treatment principals were also more likely to claim to have used average scores and average score growth on state tests to evaluate or compare teachers, and 55 percent of treatment principals said they used the value-added reports to evaluate or compare teachers.[14] This provides an initial indication that the treatment changed the set of information principals incorporated into their evaluations.

Finally, treatment principals' confidence in whether the value-added methodology controlled for various factors largely accords with reality (see Table A3). For example,

[14] If we limit the follow-up survey sample to those treatment principals who actually received value-added reports, 66 percent reported using them to evaluate or compare teachers.

principals were confident that the methodology accounted for factors like teaching experience, prior test scores, and class size (all of which were control variables) and expressed little confidence that the methodology accounted for factors like the presence of a classroom aide or whether a teacher's students received outside help (neither of which were control variables).[15]

2.2 Other Data Sources

In addition to the surveys, we use data from the value-added reports, which exist for all schools in the city (even though only treatment principals received them). In order to be familiar to principals, value-added measures were reported in "proficiency rating units," a scale based on state examinations and used by the DOE in its school accountability system. However, we normalize them (at the city level) to have a mean of zero and standard deviation of one for purposes of our evaluation. Average teacher value-added estimates at baseline for treatment and control groups were close to zero (see Table 3, left side), suggesting participating schools was similar to other schools in the DOE on average. The variance of value-added estimates was higher for those based on one year of data than for those using up to three, and higher for estimates that do not adjust for years of teaching experience relative to those that do. Variation in value-added was also smaller for English than math, consistent with studies in New York and elsewhere (see Hanushek and Rivkin (2010) for a review). In addition to value-added estimates and confidence intervals, these data contain a categorical variable for years of teaching experience. Roughly half of the teachers had at least five years of experience, while about one third had less than three years of experience.

[15] Still, it is interesting that around 25 percent of treatment principals did not express confidence that the methodology accounted for issues such as teaching experience and prior test scores. Additionally, a small fraction of principals (between 5 and 10 percent) were confident that the methodology accounted for factors which were not controlled for, such as whether a teacher had a personal issue or whether students were distracted on the day of the test by construction noise.

Human resources records provide us with information on whether a teacher switched to another school within the DOE or left the DOE's teacher workforce. Roughly 89 percent of the teachers in the baseline survey still worked in the DOE in the school year 2007-2008, and 85 percent were teaching in the same school. Roughly 82 percent still worked in the DOE in the school year 2008-2009, and 75 percent were teaching in the same school.[16]

Finally, we have student level data on test scores for the school year 2008-2009 and demographics (i.e., gender, race/ethnicity, free lunch receipt, English language learner, and special education status). These data also contain links to the students' math and English teachers. Thus, we can determine whether a teacher is still providing instruction in the same grades and/or subject, ask whether the treatment had an impact on student achievement, and test whether value-added estimates made in the summer of 2007 had predictive power for students' learning gains on tests taken in 2009. Only 58 percent of teachers in our baseline sample were teaching math or English to students in grades 4 to 8 in the same school in the school year 2008-2009.

3. Theoretical Framework: The Bayesian Learning Model

We use a simple Bayesian learning model to consider the principal's evaluation problem. Principals accumulate information regarding the effectiveness of their teachers and use this information to construct their beliefs. At time t, the principal has formed a prior belief regarding the true effectiveness of each teacher, assumed to be normal distributed (Equation 1). The mean of the prior, μ_0, is the expected value and, empirically, is akin to the evaluation provided by the principal in our baseline survey. The parameter h_0 is the precision of the prior (i.e., the inverse

[16] Turnover in treatment schools was slightly higher than in control schools (about four percentage points higher for each observed event, see Table 3). These differences, in particular those occurring before random assignment, may be due to differences in observed teacher effectiveness. We return to this point later in the paper.

of its variance) and depends on how much information the principal has accumulated, with more information leading to higher precision (lower variance).

$$(1)\ \ \mu \mid t \sim N\left(\mu_0, \frac{1}{h_0}\right)$$

Between time t and $t+1$, the principal's expectation of true teacher effectiveness will change based on the accumulation of new information (e.g., through classroom observation). Thus, all principals (including those in our control group) may change their evaluations over time. Equation 2 describes the posterior expectation of teacher effectiveness for treatment and control principals, where ε is the information routinely accumulated between periods and V is the value-added estimate provided to treatment principals, i.e., an imperfect signal of teacher effectiveness to which the principal would otherwise not have had access.[17]

$$(2)\ \ E(\mu \mid V, t+1) = u_1 = \begin{cases} \text{If treatment}: \dfrac{h_0\mu_0 + h_\varepsilon\varepsilon + h_V V}{h_0 + h_\varepsilon + h_V} \\ \text{If control}: \dfrac{h_0\mu_0 + h_\varepsilon\varepsilon}{h_0 + h_\varepsilon} \end{cases}, \varepsilon \sim N\left(\mu, \frac{1}{h_\varepsilon}\right), V \sim N\left(\mu, \frac{1}{h_V}\right)$$

It is straightforward to show that this simple model yields two intuitive predictions for our empirical work:

Prediction 1: Value-added estimates (V) should be correlated with principals' prior beliefs (μ_0), since both are signals of teacher effectiveness. This correlation should be stronger when value-added estimates and prior beliefs are more precise (i.e., greater values of h_V and h_0).

Prediction 2: Conditional on principals' prior beliefs (μ_0), treatment principals' posterior beliefs regarding teacher effectiveness should, relative to control principals, place more weight on value-added estimates and less weight on prior beliefs. Treatment principals should place more weight on value-added estimates with greater precision (h_V), and less weight on value-added estimates when their priors are more precise. Principals' posterior beliefs in the control group may also be conditionally correlated with value-added estimates, to the extent that value added is correlated with the new information gathered between surveys by all principals.

[17] Note we have assumed that the error components of the principal's prior belief and the value-added estimate are independent. Relaxing this assumption does not affect the qualitative implications of the model, but an increase the correlation of these error components essentially reduces the extent to which value-added provides new information.

We use principals' baseline evaluations to measure prior beliefs (μ_0), principals' follow-up evaluations to measure posterior beliefs (μ_1), estimates from the value-added reports to measure V, and the confidence intervals given in the value-added reports to measure h_V. Unfortunately, we do not have a readily available measure of the precision of principals' prior beliefs (h_0). We therefore proxy for h_0 in our analysis using data on the number of years during which the principal had observed the teacher.[18]

The model does not give clear predictions for the actions that treatment principals might take in reaction to the value-added information they receive. However, there are several reasonable hypotheses that we can test with the available data. First, the provision of value-added information may crowd out information gathering by treatment principals. For simplicity we did not model information collection as a choice, but if classroom observation is costly, then treatment principals could respond to the provision of value-added by observing teachers less frequently—the marginal benefit of an additional classroom observation on the principal's posterior belief will be smaller when the principal has more precise information from other sources. Second, it is reasonable to believe that providing value-added information may create a stronger relationship between value-added estimates and teacher turnover, particularly through changes in principals' posterior beliefs. Also, along the lines of Baker et al. (1994), value-added estimates could also lower the cost to principals of acting on existing beliefs, by providing independent and verifiable confirmation of their priors. Finally, if value-added information causes changes in turnover or resource allocation that improve overall educational quality (e.g.,

[18] While we lack complete data on teachers' work histories, we know the number of years each teacher has been working in the value-added subjects/grades in his/her current school and the number of years the principal has been working in the current school. To measure the number of years during which the principal had likely observed the teacher, we take the minimum of these two variables. We do know teachers' total years of experience, but this is likely to overestimate experience within a school for a large fraction of teachers.

through selective retention or additional training), we might expect to see student achievement rise in treatment schools, relative to control schools.

4. Value-added Estimates and Principals' Priors

To test Prediction 1 from our framework, we measure the relationship between value-added estimates and principals' prior beliefs using a series of linear regressions. First, we estimate a series of specifications where a baseline evaluation given to teacher *i* (R_i) is regressed on a teacher's value-added estimate (V_i), as shown in Equation 3.

$$(3)\quad R_i = \alpha + \beta V_i + \varepsilon_i$$

For easy interpretation, principal evaluations and value-added estimates are normalized to have a mean of zero and standard deviation of one. In specifications that pool across math and English, we average the subject-specific value-added estimates and principal evaluations for teachers of both subjects. Standard errors are clustered by school.

As expected, principals' pre-experimental evaluations of a teacher's overall performance are significantly higher for teachers with higher value-added estimates (Table 4, Panel A). We estimate similar effect sizes (0.21 to 0.23) using multi-year estimates that compare teachers to their peers (i.e., those with similar experience and similar classrooms of students), single year estimates that compare teachers to peers, and multi-year estimates that compare teachers citywide (Columns 1 to 3). We then run "horse races" between different value-added estimates to test the relative strength of their relationship to principals' evaluations. Multi-year estimates dominate estimates based on only the past year of student performance (Column 4), and multi-year estimates using the peer comparison are also stronger predictors of evaluations than those based on citywide comparisons. Though both are statistically significant when included together in the regression (Column 5), the conditional relationship between citywide value-added

and principals' priors disappears if we control for teacher experience (Column 6).[19] In the remainder of our analysis, we measure value-added using the multi-year peer comparison estimate; the conditional correlation between this measure and principals' priors is robust to controlling for principal fixed effects (Column 7).

We find very similar results when we replace the principal's overall evaluation of the teacher with the evaluation of the teacher's ability to raise student achievement in math and/or English (Table 4, Panel B). Though all these results are qualitatively similar, the point estimates on value-added are slightly larger. For example, when controlling for principals fixed effects and experience (Columns 7 and 14), the value-added coefficient is 0.25 for evaluations of overall performance and 0.27 for evaluations of performance in raising achievement. This suggests that principals may distinguish between performance in raising student achievement from other aspects of the job, and we explore this notion further below.[20]

We also find the coefficients on value-added are all positive and significant if we run these baseline regressions splitting the sample across teachers of math, English, or both math and English (Table 5). The effect size for teachers of only math is roughly 0.38 for both evaluations of overall performance and raising student achievement, noticeably larger than our pooled estimates. Also, for teachers of only English, the effect size for overall performance evaluations (0.16) is noticeably smaller than our baseline estimates and somewhat smaller than the estimate when we consider evaluations for raising student achievement (0.22). This provides some suggestive evidence that principals' views of teachers' overall contributions are more aligned

[19] Coefficient estimates on experience indicators are not reported but are available upon request. Conditional on value-added, principals' baseline evaluations were lowest for teachers who just completed their first year, and highest for teachers with three to nine years of experience, while teachers with only a few years of experience or ten or more years of experience tend to be rated in the middle.

[20] While we pool treatment and control groups for power, these baseline findings are not substantially or statistically different if we limit the sample to one group. For example, in the specification shown in Column 14, the value-added coefficient is 0.277 for the treatment group only (standard error 0.029, 1,324 observations) and 0.263 for the control group only (standard error 0.044, 1,184 observations).

with student achievement on statewide standardized exams for math than for English in later grades when students receive instruction in these subjects from different teachers. However, among teachers of both subjects (who generally teach lower grade levels), the effect sizes were larger for value-added estimates in English than for math, either when estimated separately (Columns 3, 4, 8, and 9) or together in the same regression (Columns 5 and 10). Thus, on the whole, principals' prior beliefs were not clearly tied more strongly to one of the two subject areas.

Principals' evaluations of teachers' "overall" performance and performance in "raising student achievement" are very highly but imperfectly correlated (0.87). To explore the distinctions between these different evaluations and their relationship with value-added, we first regress the principals' evaluation of overall performance on value-added while controlling for a principal's evaluation of a teacher's ability to raise student achievement. Here, the coefficient on value-added is very close to zero and insignificant (Table 6, Column 1). In contrast, if we regress the principals' evaluation of a teacher's ability to raise student achievement on value-added while controlling for a principal's evaluation of overall performance, the coefficient on value-added is positive (0.054) and highly significant (Table 6, Column 2). Thus, principals do indeed distinguish between a teacher's performance on improving standardized test outcomes and other aspects of job performance, even though these performances are highly correlated.

For teachers of both math and English, the principal's evaluation of performance in raising math achievement is very highly but imperfectly correlated (0.91) with the evaluation for raising English. This motivates us to explore whether principal evaluations that are specific to math and English capture important subject specific elements of teacher effectiveness (as opposed to general teaching skills) and whether these are related to value-added. To do so, we

restrict our sample to teachers who provide both math and English instruction and estimate the specifications shown in Equations 4a and 4b; where *k* and *j* subscripts denote different subjects.

$$\text{(4a)}\quad R_{ij} = \alpha + \pi R_{ik} + \beta^{jj} V_{ij} + \varepsilon_{ij}$$

$$\text{(4b)}\quad R_{ij} = \alpha + \pi R_{ik} + \beta^{jk} V_{ik} + \varepsilon_{ij}$$

If both the principals evaluations (*R*) and value-added (*V*) truly measure subject-specific teaching performance, then our estimate of β^{jj} should be positive and significant while our estimate of β^{jk} should be close to zero. In other words, conditional on the principal's evaluation in subject *k*, the principal's evaluation in subject *j* should be related to value-added in subject *j* and not related to value-added in subject *k*.

This is precisely what we find (Table 6, Columns 3 to 6). For math evaluations, the coefficient on math value-added is positive and statistically significant (0.028), but the coefficient on English value-added is small and insignificant (-.007). Likewise, when English evaluation is the dependent variable the coefficient on English value-added is positive and statistically significant (0.049), but the coefficient on math value-added is small (0.017) and insignificant. Thus, despite being highly correlated across subjects, both subjective evaluations and objective performance estimates are sensitive to teachers doing well in a particular subject.

Having established a consistent baseline relationship between value-added and principals' performance evaluations, we proceed to test the additional elements of Prediction 1. First, the relationship between principals' priors and value-added estimates should be greater when the estimates have greater precision. To test this prediction, we add an interaction between the value-added estimate and a measure of its precision (h_V)—the inverse of the confidence

interval provided to principals on the value-added report.[21] Consistent with the Bayesian learning model, value-added estimates are more strongly correlated with principals' priors when those estimates have tighter confidence intervals (Table 7). Our estimates of the interaction of value-added and precision are positive and highly significant, and very similar regardless of whether we use the principal's overall evaluation (Columns 1 to 3) or evaluation of the teacher's ability to raise student achievement (Columns 4 to 6), or whether we control for teacher experience and school fixed effects (Columns 2 and 5). One concern with this specification is that teachers for whom more years of data are used to generate their value-added estimates will tend to have smaller confidence intervals, and, even conditional on total teaching experience, these teachers are more effective. However, controlling for the number of years of data used in the teacher's value-added estimate has little impact on our estimates (Columns 3 and 6).[22]

To measure the strength of the principals' priors, we adjust our basic regression specification in several ways. First, we place the value-added estimate as the dependent variable and the principal's evaluation as an independent variable. We also interact the evaluation with the years of experience the principal has, and limit the sample to teachers with three years of value-added, i.e., teachers for whom we are sure they have worked in the school for at least three years. More experienced principals' evaluations are based on more information than those made by principals with less experience, and we would therefore expect a positive coefficient on the interaction term. In other words, when it comes to predicting the value-added of experienced teachers, new principals should not be as accurate as experienced principals.

[21] We find very similar results to those described here if we use an interaction with the inverse of the variance of the value-added estimate, which is technically the correct measure of precision. However, we use the inverse of the confidence interval later in our analysis to test Prediction 2 because the confidence interval is what was actually provided to principals in the value-added report. We therefore also use it here for consistency.

[22] Another concern may be that the most precise estimates are for teachers of only math, and in Table 5 we showed that value-added was a stronger predictor of principals' evaluations for these teachers. However, we find positive and significant interactions of value-added estimates with their precision in regressions that separately examine teachers by subject areas taught.

The empirical results are in line with this expectation and the Bayesian learning model. Both the main effect of principals' evaluations and the interaction with principal experience are positive and statistically significant (Table 8). Using principals' overall evaluations, we find a main effect of 0.151 and an interaction of 0.011 (Column 1). If we use evaluations of teachers' ability to raise student achievement, we get slightly larger point estimates (0.158 and 0.015, Column 4). One potential issue with this specification is that very experienced teachers may have taught other subjects or grade levels, or may simply have variation in their performance that is not captured by the available value-added estimate (based on three years of data) but would have been observed by a very experienced principal. We therefore might expect to find a larger coefficient on the interaction of the principal's evaluation and principal experience when we limit the sample to teachers with fewer years of experience. This is indeed the case, with interaction terms growing as we remove teachers with 10 or more years of experience (Columns 2 and 5) and, teachers with 5 or more years of experience (Columns 3 and 6).[23] Thus, this element of Prediction 1 is well borne out by the data, though it is unfortunate that we did not solicit more direct measures of the precision of principals' prior beliefs.

Finally, we examine how the frequency of principals' observations is related to teachers' value-added estimates. To do so, we estimate regressions of the form shown in Equation 3, but with observation frequency as the dependent variable.[24] We find that principals more frequently observe teachers with lower value-added estimates (Table 9, Columns 1 and 4). There are several reasons why observational frequency and value-added may be related in this way. First,

[23] We also estimated a specification that interacted principals' evaluations with an indicator for whether the principal strongly agreed with the statement "I know who the most effective teachers are in my school" in the baseline survey. While this interaction is positive, as we would predict, it was small and statistically insignificant. This lack of power may not be surprising, given that three-quarters of the principals in our sample strongly agreed with the statement about the extent of their knowledge.

[24] There are a few teachers for whom the principal reported only the number of formal observations and left the question on total observations blank. The results are insensitive to dropping these teachers from our regression of formal observation frequency.

if one of the main goals of observation is to identify ineffective teachers, principals may spend more time observing teachers they believe are performing poorly in order to gain a more precise evaluation. Second, observation may also be used to provide constructive criticism to help teachers improve, so that, again, observation is skewed towards low performing teachers.

We estimate a few additional specifications predicting observation frequency. First, we find that the inclusion of experience has little impact on the coefficient on value-added estimates, though the coefficients on teacher experience (not reported but available upon request) show that principals reduce the frequency of their observations as teachers gain more experience (Columns 2 and 5). Adding principal fixed effects (Columns 3 and 6) also has little impact on this finding, though standard errors decrease substantially in the regression of total observations. Finally, the value-added coefficients shrink considerably if we include the principal's overall evaluation as a control variable (Columns 4 and 8). Thus, the relationship between observation frequency and value-added is mediated by principals' prior beliefs on teacher effectiveness.

5. The Impact of Information on Employee Evaluation

Our primary prediction for the impact of information is that principals should place more weight on value-added estimates and less weight on prior beliefs. To test this, we estimate regression of posterior evaluations on teacher value-added and prior evaluations (Equation 5).

$$(5)\ R_{it+1} = \alpha + \lambda R_{it} + \beta V_{it} + \varepsilon_{it}$$

The evaluation given to teacher *i* at time *t+1* (R_{it+1}) is specified as a function of the teacher's prior evaluation (R_{it}), the value-added estimate (V_{it}) and a disturbance term (ε_{it}). We estimate regressions for treatment and control groups separately and compare their coefficients.

The results are in line with our prediction that providing value-added estimates to principals had a significant impact on the formation of posterior beliefs. We find a highly

significant positive effect of value-added on post-experimental evaluations for the treatment group (0.123) and a small and insignificant effect (0.017) for the control group (Table 10, Column Group 1). When we include principal fixed effects, the coefficient on value-added estimates for the control group rises slightly (0.038) and becomes marginally significant (p-value 0.14), but the coefficient for the treatment group rises by a similar amount (to 0.149) and the difference between the coefficients remains statistically significant (Column Group 2). The coefficients on prior evaluation are both positive and significant for both groups. While the estimates for the treatment group is smaller (e.g., 0.793 vs. 0.824 in Column Group 1) the differences across the two groups are not significant and insensitive to controlling teacher experience and school fixed effects.

The Bayesian learning model predicts that principals would place relatively more weight on value-added reports that were relatively more precise and less weight on value-added estimates for the teachers for whom they had a relatively precise prior. To test this, we interact both value-added and the principal's prior evaluation with (a) our measure of the value-added estimate's precision and (b) the number of years the teacher had been under the principal's supervision.[25] As predicted, for the treatment group we find a significant positive interaction of value-added with precision and a significant negative interaction of value-added with the number of years the principal has supervised the teacher. Also, as predicted, we find a negative and marginally significant (p-value 0.11) interaction of the principal's prior evaluation with the precision of value-added and a significant positive interaction of the prior evaluation with the number of years the principal has supervised the teacher. In contrast, the interaction coefficients

[25] This is based on the minimum of the number of years the principal has been at the school and the number of years of data from the current school used to construct the teacher's value-added estimate. Unfortunately we do not have information on teaching experience within the school, and our measure of total DOE experience is likely to misclassify many experienced teachers who have changed schools.

for the control group are all much closer to zero, never even marginally significant, and sometimes of a different sign. These results are also robust to the addition of teacher experience and school fixed effects.

Thus, our findings are quite consistent with the predictions of the simple learning model. Principals who receive performance data on their teachers use this information in updating their priors. They put more weight on the new data and less on their priors when the data is more precise (i.e., greater values of h_V), and less weight on new data and more on their priors when their priors are more precise (i.e., greater values of h_0).

As in our examination of principal's priors, we also examine the influence of value-added on principals' posterior evaluations separately for teachers of math, English, and both math and English. Here we find consistent evidence that the value-added estimates in math were more influential than those for English (Table 11). Specifically, we find positive significant effects of math value-added on principals' posterior evaluations for teachers of math and for teachers of math and English for the treatment group, while the control group estimates are significantly smaller and not distinguishable from zero (Column Groups 1, 3, and 5). Meanwhile, we find no significant effects of English value-added estimates on principals' posterior beliefs, and the coefficient estimates in the treatment and control groups are very similar (Column Groups 2, 4, and 6). Why principals were more influenced by the value-added reports in their evaluation of math teaching is unclear. It is possible that the timing of the English exam—given in January, as opposed to math which is given in April—increased principals' concerns about the ability of the value-added methodology to accurately measure an individual teacher's contribution to student achievement. It may also be that principals were more confident in their ability to gauge the quality of English instruction, and therefore put less weight on the value-added estimates.

6. Information Acquisition, Worker Turnover, and Productivity

The results presented in Section 5 establish the impact of information provision on principals' subjective evaluations of work performance. However, it is unclear how changes in principals' beliefs regarding teacher effectiveness translated into changes in personnel decisions or the quality of education provided at the school. In this section, we ask whether providing information on employee performance causes principals to gather less information via classroom observation, changes patterns of turnover, or raises student achievement.

6.1 Information Crowd-out

Principals have many duties besides teacher evaluation, and a policy that provides them with data on teacher performance may crowd out time and energy spent on observing teachers in the classroom. In the context of the pilot we study, we can look for evidence of crowd-out in the short run, i.e., value-added reports were received in December and the follow-up survey asked principals about formal and total observations of each teacher during the pilot year. However, if we estimate regressions of formal or total observations (either levels or changes from the baseline survey) on an indicator for treatment, we find no evidence of crowd-out (Table 12). In other regressions, not reported but available upon request, we found no significant interaction between treatment status and value-added, or between treatment status and the principal's baseline evaluation.

6.2 Employee Turnover

Given that value-added information did affect performance evaluations, one of the channels through which we might expect to see change occur is through selective retention of

teachers. We therefore examine how the propensity of teachers to exit their schools after the pilot year was related to their value-added estimates, and whether this relationship differed between treatment and control schools. We use two regression specifications, shown by Equations 6a and 6b:

$$(6a)\ E_{it+1} = \beta V_{it} + \varepsilon_{it+1}$$
$$(6b)\ E_{it+1} = \gamma V_{it} + \lambda R_{it} + \zeta_{it+1}$$

E_{it+1} is an indicator for whether teacher *i* is no longer employed in the same school at time *t+1*, V_{it} is the teacher's value-added estimate at baseline, R_{it} is the principal's evaluation at baseline, and ε_{it+1} and ζ_{it+1} are disturbance terms. Since our dependent variable is binary, we present results using both linear regression and logit specifications, though our results are quite similar across the two estimation methods. As with our analysis of principals' post-experimental evaluations, we run regressions separately for treatment and control groups and then test for differences between the groups. Again, standard errors are clustered by school.

We find clear evidence that providing the value-added reports did indeed cause teachers with lower value-added estimates to be more likely to exit treatment schools (Table 13). The coefficient on value-added is statistically significant and negative in the treatment group, and we can reject the equality of the treatment and control coefficients at the 11 percent level for OLS and the 12 percent level for the logit regression (Column Groups 1 and 4). When we include the principal's prior evaluation of the teacher (Column Groups 2 and 5), the value-added coefficients remain significant and negative for the treatment group and we can reject equality with the control group at the 6 percent level in both types of regressions. The coefficient on the principal's prior evaluation is negative for both groups, but is significantly larger for the control group. Thus, as with our analysis of posterior evaluations, we find treatment principals putting

more weight on new information and less weight on their prior beliefs. These results are robust to including teacher experience and school fixed effects (Column Groups 3 and 6).[26]

While we recognize that employment is not a subject specific outcome, we also separately analyze the impact of value-added on employment by subject due to our results on principals' posterior beliefs, which were stronger in math than English. Similarly, we find that the negative impact of value-added on turnover was driven by value-added in math, not English (Table 14). This provides further evidence that the provision of performance data was used by principals in forming beliefs about job performance and continued employment.

6.3 Student Achievement

There are two channels through which the provision of objective performance data could have raised productivity in treatment schools. First, if the value-added estimates are valid predictors of how teachers will perform in the future, the results on turnover suggest that we might see a slight increase in student achievement, particularly in math, due to selective retention. Second, principals may use the information to provide low performing teachers with additional resources and training, though we have no data by which to assess this channel.

To estimate the overall effect of the treatment on student achievement, we estimate a student-level regression of achievement gains (i.e., 2009 score minus 2008 score) on an indicator for being in the treatment group. We allow for random effects at the school and teacher level to account for the nested structure of the data. Our estimate of the treatment effect in math (Column 1 of Table 15) is positive, small (0.024 student level standard deviations) and

[26] As a further check, we examined the probability that a teacher exited the school before the start of the pilot (i.e., between the school years 2006-2007 and 2007-2008). These "placebo tests" showed a very similar and insignificant relationship between value-added and exiting the school for treatment and control schools, while the coefficients on principals' pre-existing beliefs regarding teacher effectiveness were negative, significant, and very similar for the two groups. This supports the notion that principals make personnel decisions based on their subjective evaluations of job performance, and did this similarly in treatment and control schools prior to the start of the pilot.

marginally significant (p-value 0.16). This estimate is insensitive to controlling for teacher experience (Column 2) or student level and grade level covariates (Column 3).[27]

It is important to note that, among the students in treatment and control schools for whom 2008-2009 test score gains were available, fewer than half were taught by a teacher who was evaluated in the pilot's baseline survey, due to turnover and reassignment (to non-tested grades or subjects) within the school. When we limit our sample to students taught by teachers in the pilot study, the point estimate for the treatment group rises to 0.04 and has a p-value of 0.10 (Column 4). Thus, achievement gains produced by retained teachers were higher in the treatment schools than in the control schools. When we include principals' overall evaluations of teachers at baseline as a covariate (Column 5) we find it has a positive significant coefficient (0.074) and that the treatment effect estimate rises to 0.057 (p-value 0.04). In other words, among teachers from the pilot who were still in the value-added grades and subjects, those in the treated group were originally rated worse, on average, than those in the control group. This is consistent with the notion that positive signals from the value-added caused treatment principals to keep some teachers of whom they held a low opinion, and that negative signals from the value-added caused them to dismiss some teachers of whom they had a (relatively) high opinion. In further support of this notion, when we control for the value-added estimates themselves— which also have a positive and significant coefficient— the treatment coefficient shrinks back to 0.044 (p-value 0.09). That is, conditional on the principals baseline opinion, treatment school teachers still providing math instruction in these grades had higher value-added than control school teachers.

[27] Student characteristics include: prior test score, prior test score interacted with grade level, prior test score in the other subject (e.g.., reading when predicting math gains), student gender, racial/ethnic subgroup, English language learner status, special education status, and eligibility for free or reduced price lunch. Grade level covariates include grade fixed effects interacted with the grade configuration of the school (e.g., grade 6 student in middle schools).

Given the results of earlier sections, it is not surprising that we find no evidence of significant changes in student achievement in English (Columns 7 to 12). The point estimates on the treatment effect are never even marginally significant and change signs. However, the coefficients on principals' baseline evaluations and teachers' value-added estimates are both positive and highly significant. This suggests that both measures are, as we posited in our conceptual framework, imperfect but useful estimates of teacher effectiveness. [28]

7. Conclusion

In this paper, we take advantage of a unique experiment conducted in the New York City public schools to learn about how managers evaluate employee job performance and how objective performance data influences this process. We frame our analysis in the context of a Bayesian learning model; school principals learn about the effectiveness of their teachers using imperfect information, and can be aided by the provision of "value-added" measures that estimate teacher performance using standardized achievement test scores.

Our empirical analysis presents several facts that are consistent with this model. First, value-added and principals' prior beliefs about teacher effectiveness are positively correlated. Second, this relationship is stronger when value-added measures are more precisely estimated or when the principal has supervised the teacher for a longer period of time—our proxy for the precision of principals' prior beliefs. Third, principals change their evaluations of teachers in response to information on value-added, a fact we are able to document due to the randomized selection of principals who received information. Fourth, the impact of providing information is greater when value-added estimates are more precise and smaller when principals have already

[28] Note that if we estimate these coefficients separately for treatment and control schools, they are always positive and significant in both subjects, and are not statistically distinguishable across the two groups.

supervised their teachers for a greater number of years. We also find that teachers with lower value-added estimates are more likely to exit teaching in the school after the principal receives this information, and we find small, marginally significant improvements in test scores that are consistent with these changes in selective retention.

Overall, our results consistently show that objective job performance data can be useful to managers in forming subjective evaluations of employee job performance and raising workforce productivity. However, in the context of public schools, it is also interesting to consider how the privacy of this information affects its usefulness as a policy tool. In the pilot program we study, the value-added reports constituted private information for the principal who received them, and many of the teachers with low value-added estimates that exited treatment schools found another teaching job within the school district. Boyd et al. (2007), also studying schools in New York City, find that low value-added teachers who transfer schools continue to perform poorly, suggesting that private information on teacher performance contributes to the continued employment of poor performing teachers, a process often described as the "dance of the lemons" by education professionals (see Ravitch (2007)). If the value-added reports contain information useful to principals about their current teachers, we speculate they might provide useful information to principals considering applicants who previously worked in other schools.

References

Aaronson, D., Barrow, L. & Sander, W. (2007) "Teachers and Student Achievement in the Chicago Public High Schools," Journal of Labor Economics, 25(1): 95-135.

Baker, G.P., Jensen, M.C., and Murphy, K.J. (1988) "Compensation and Incentives: Practice vs. Theory," Journal of Finance, 43(3): 593-616.

Baker, G.P., Gibbons, R. and Murphy, K.J. (1994) "Subjective Performance Measures in Optimal Incentive Contracts," Quarterly Journal of Economics, 109 (4): 1125-1156.

Ballou, D. (2001) "Pay for Performance in Public and Private Schools," Economics of Education Review 20(1): 51-61.

Besley, T. and Ghatak, M. "Competition and Incentives with Motivated Agents," American Economic Review, 95(3): 616-636.

Boyd, D., Grossman, P., Lankford, H., Loeb, S. & Wyckoff, J. (2007) "Who Leaves? Teacher Attrition and Student Achievement," Unpublished Working Paper.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K. and Rogers, F.H. (2006) "Missing in Action: Teacher and Health Worker Absence in Developing Countries," Journal of Economic Perspectives, 20(1): 91-116.

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

Gibbons, R. (1998) "Incentives in Organizations," Journal of Economic Perspectives, 12(4): 115-132.

Gibbons, R. (2005) "Incentives Between Firms (and Within)," Management Science, 51(1): 2-17

Gordon, R., Kane, T., & Staiger, D. (2006) The Hamilton Project: Identifying Effective Teachers Using Performance on the Job. Washington, DC: The Brookings Institution.

Harris, D.N. and Sass, T.R. (2008) "What Makes for a Good Teacher and Who Can Tell?" Unpublished Manuscript.

Hanushek, E.A. and Rivkin S.G. (2010) "Generalizations about Using Value-Added Measures of Teacher Quality," American Economic Review, Papers and Proceedings 100(2): 267-271.

Holmstrom, B. and Milgrom, P. (1991) "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," Journal of Law, Economics, and Organization, 7(Sp): 24-52.

Jacob, B.A. (2007) "The Demand Side of the Teacher Labor Market," Unpublished Manuscript, University of Michigan.

Jacob, B.A. (2010) "Do Principals Fire the Worst Teachers?" NBER Working Paper 15715.

Jacob, B.A., and Lefgren, L.J. (2008) "Principals as Agents: Subjective Performance Measurement in Education" Journal of Labor Economics 26(1): 101-136.

Jacob, B.A. and Levitt, S.D. (2003) "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," Quarterly Journal of Economics, 118(3): 843-877.

Jovanovic, B. (1979) "Job Matching and the Theory of Turnover," Journal of Political Economy, 87(5): 972-990.

Kane, T.J. and Staiger, D.O. (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation" NBER Working Paper #14607.

Lavy, V. (2002) "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," Journal of Political Economy, 110(6): 1286-1317.

Lavy, V. (2009) "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," American Economic Review, 99(5): 1979–2011.

Medoff, J.L. and Abraham, K.G. (1980) "Experience, Performance, and Earnings," Quarterly Journal of Economics, 95(4): 703-736.

Muralidharan, K., and Sundararaman, V. (2008) "Teacher Incentives in Developing Countries: Experimental Evidence from India," Unpublished Working Paper.

Murnane, Richard J. (1975) *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Balinger.

Murphy, K.J. (1992) "Performance Measurement and Appraisal: Motivating Managers to Identify and Reward Performance," in W.J.J. Burns (Ed.), Performance Measurement, Evaluation, and Incentives, Boston, MA: Harvard Business School Press pp. 37-62.

Murphy, K.J. and Oyer, P. (2001) "Discretion in Executive Incentive Contracts: Theory and Evidence," Unpublished Manuscript.

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

Prendergast, C. and Topel, R. (1993) "Discretion and Bias in Performance Evaluation," European Economic Review, 37(1): 355-365.

Ravitch, D. (2007) Edspeak: A Glossary of Education Terms, Phases, Buzzwords, Jargon. Alexandria, VA: ASCD.

Rivkin, S.G., Hanushek, E. A. & Kain, J. (2005) "Teachers, Schools, and Academic Achievement," Econometrica, 73(2): 417–458.

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

Rockoff, J.E. and Turner, L.J. (2008) "Short Run Impacts of Accountability on School Quality," NBER Working Paper #14564.

Rothstein, J. (2009) "Student Sorting and Bias in Value-added Estimation: Selection on Observables and Unobservables," Unpublished Manuscript, Princeton University.

Staiger, D.O. and Rockoff, J.E. (2010) "Searching for Effective Teachers with Imperfect Information," Journal of Economic Perspectives, Summer 2010.

Todd, P.E. and Wolpin, K.I. (2007) "On the Specification and Estimation of the Production Function for Cognitive Achievement," Economic Journal, 113(1): 3-33.

Turner, L.J. and S. Goodman (2009) "Group Incentives for Teachers: The Impact of the NYC School-Wide Bonus Program on Educational Outcomes." Columbia University Department of Economics Discussion Paper 0910-05.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009) *The Widget Effect*. Brooklyn, NY: The New Teacher Project.

Table 1: Comparison of Treatment and Control Groups at Baseline and Follow-up Surveys

	Baseline (112 Treatment, 111 Control)				*Followup (84 Treatment, 94 Control)*			
	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C
Total Enrollment	705	717.8	12.3	0.79	715	724.9	9.7	0.85
Principal Characteristics								
Years of Experience as Principal (in School)	3.3	3.2	-0.1	0.81	3.2	3.2	0.0	0.98
Years of Experience as Assistant Principal	2.4	2.7	0.3	0.40	2.4	2.8	0.3	0.43
Years of Experience as Teacher	6.8	7.6	0.8	0.36	6.6	7.8	1.2	0.18
Years of Experience in School (Any Position)	4.4	5.1	0.7	0.31	4.4	4.8	0.5	0.53
Principal Age	48.0	48.8	0.8	0.50	48.2	49.9	1.7	0.16
Principal is Black or Hispanic	48.6%	41.9%	-6.7%	0.32	47.9%	44.1%	-3.8%	0.61
Principal is Female	81.1%	77.7%	-3.4%	0.53	80.9%	77.4%	-3.5%	0.57
Student Characteristics (Grades 4-8)								
On Free Lunch	87%	85.4%	-1.2%	0.64	85.7%	85.0%	-0.7%	0.81
English Language Learners	15%	13.3%	-1.4%	0.30	14.7%	13.0%	-1.7%	0.30
In Special Education	10%	9.8%	0.3%	0.79	9.2%	10.2%	1.0%	0.40
Black	72%	72.8%	0.5%	0.91	71.9%	72.9%	1.0%	0.83
Hispanic	29%	31.8%	2.5%	0.52	30.2%	31.0%	0.8%	0.85
School Environment (Teacher Survey, Spring 2007)								
The Principal…								
Visits Classrooms to Observe the Quality of Teaching	0.071	-0.03	-0.101	0.44	0.091	-0.026	-0.117	0.44
Gives Me Regular and Helpful Feedback	-0.047	-0.069	-0.022	0.86	-0.026	-0.101	-0.075	0.59
Places a High Priority on the Quality of Teaching	-0.027	0.004	0.031	0.80	-0.031	0.034	0.065	0.64
Teachers in this School…								
Use Student Data to Improve Instructional Decisions	0.096	0.077	-0.019	0.87	0.093	0.058	-0.035	0.80
Receive Training in the Use of Student Data	0.036	0.022	-0.014	0.90	0.049	0.012	-0.037	0.79

Note: P-values indicate the statistical significance of a treatment indicator to predict the survey response. All variables from the school environment survey have been normalized using all schools in New York City to have mean zero and standard deviation one. Four schools (one control, three treatment), are missing environment survey outcomes, due to the fact that teachers in these schools did not complete the survey.

Table 2: Average Characteristics of Study Sample and Schools Citywide Serving Grades 3-8

	City	Sample	P-value on Test of Equality
Number of Principals/Schools	1092	223	
Average School Enrollment	660	712	0.03
Principal Characteristics			
Years of Experience as School's Principal	3.6	3.3	0.27
Years of Experience in School (Any Position)	5.2	4.7	0.17
Years of Experience as Assistant Principal (Any School)	2.7	2.5	0.41
Years of Experience as Teacher (Any School)	6.2	7.2	0.01
Principal Age	48.8	48.4	0.41
Principal is Black or Hispanic	47.7%	45.3%	0.47
Principal is Female	73.4%	79.4%	0.05
Student Characteristics			
On Free Lunch	84.9%	86.1%	0.35
Black or Hispanic	75.6%	72.5%	0.11
English Language Learners	11.5%	14.0%	0.00
In Special Education	13.8%	9.6%	0.00
Teaching Environment (Spring 2007)			
The Principal…			
Visits Classrooms to Observe the Quality of Teaching	0.00	0.03	0.65
Gives Me Regular and Helpful Feedback	0.00	-0.05	0.47
Places a High Priority on the Quality of Teaching	0.00	-0.02	0.72
Teachers in this School…			
Use Student Data to Improve Instructional Decisions	0.00	0.00	0.99
Receive Training in the Use of Student Data	0.00	-0.01	0.90

Note: Standard deviations of continuous variables shown in parentheses below mean values. P-values indicate the statistical significance of a treatment indicator to predict the survey response. All variables from the environment survey have been normalized using the citywide distribution to have mean zero and standard deviation one. Four sample schools (one control, three treatment) and 72 schools citywide are missing environment survey outcomes, due to the fact that teachers in these schools did not complete the survey.

Table 3: Summary Statistics on Teacher Level Variables

	Baseline Survey		Followup Survey	
	Control	Treatment	Control	Treatment
In Baseline Survey with Value Added Estimate	1184	1324	780	747
Principal's Rating (Scale from 1 to 6)				
Overall	4.32	4.31	4.51	4.57
	(1.11)	(1.12)	(1.04)	(1.06)
Math Instruction	4.21	4.23	4.46	4.50
	(1.09)	(1.13)	(0.99)	(1.03)
ELA Instruction	4.21	4.19	4.45	4.43
	(1.04)	(1.13)	(1.02)	(1.03)
Observations Made by Principal Last Year				
Formal	2.21	2.24	1.92	1.98
	(1.26)	(1.25)	(0.99)	(1.26)
Total	6.51	6.26	4.92	4.75
	(3.38)	(3.21)	(3.27)	(3.13)
Value-added Estimates				
Math, Multi-year, City Comparison	0.002	0.002	0.004	0.015
	(0.166)	(0.178)	(0.158)	(0.168)
Math, Multi-year, Peer Comparison	0.005	0.013	0.008	0.022
	(0.136)	(0.147)	(0.130)	(0.138)
Math, Single-year, Peer Comparison	-0.003	0.013	0.001	0.024
	(0.154)	(0.159)	(0.147)	(0.150)
ELA, Multi-year, City Comparison	-0.014	-0.002	-0.018	0.011
	(0.135)	(0.125)	(0.132)	(0.121)
ELA, Multi-year, Peer Comparison	-0.011	0.002	-0.016	0.010
	(0.092)	(0.091)	(0.086)	(0.089)
ELA, Single-year, Peer Comparison	-0.013	0.003	-0.019	0.012
	(0.105)	(0.109)	(0.101)	(0.109)
Teacher Experience in School Year 2006-2007				
None (First Year of Teaching was 2006-2007)	9.0%	10.8%	9.6%	10.4%
One Year	11.7%	10.8%	10.6%	8.6%
Two Years	10.9%	10.6%	11.0%	8.6%
Three Years	8.6%	10.9%	8.3%	12.4%
Four Years	7.4%	6.8%	6.9%	6.4%
Five to Nine Years	27.9%	26.8%	28.2%	28.1%
Ten or More Years	24.6%	23.2%	25.3%	25.4%
Turnover				
Employed in DOE, 2007-2008	91.0%	88.1%	*n/a*	*n/a*
Employed in Same School, 2007-2008	87.2%	83.4%	*n/a*	*n/a*
Employed in DOE, 2008-2009	84.0%	79.8%	93.1%	92.2%
Employed in Same School, 2008-2009	77.2%	72.7%	89.5%	89.2%
Teaching Math/English, 2008-2009	60.4%	56.2%	73.9%	72.3%

Note: Standard deviations in parentheses. Teachers for whom the principal reported more than 10 total observations made in the last year are given a value of 11.

Table 4: Principals' Pre-experimental Performance Evaluations and Value-Added

Panel A: "Overall" Performance	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Value-added, Multi-year, Peer	0.228**			0.208**	0.160**	0.266**	0.248**
	(0.024)			(0.052)	(0.037)	(0.057)	(0.024)
Value-added, Single-year, Peer		0.209**		0.023			
		(0.023)		(0.051)			
Value-added, Multi-year, Citywide			0.211**		0.093*	-0.045	
			(0.023)		(0.036)	(0.068)	
Teacher Experience Controls						√	√
School Fixed Effects							√
R-squared	0.05	0.04	0.05	0.05	0.06	0.08	0.32
Sample Size	2,507	2,507	2,507	2,507	2,507	2,507	2,507
Panel B: "Raising Student Achievement"	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Value-added, Multi-year, Peer	0.255**			0.207**	0.172**	0.301**	0.272**
	(0.026)			(0.054)	(0.041)	(0.063)	(0.025)
Value-added, Single-year, Peer		0.240**		0.056			
		(0.025)		(0.053)			
Value-added, Multi-year, Citywide			0.241**		0.114**	-0.053	
			(0.026)		(0.041)	(0.074)	
Teacher Experience Controls						√	√
School Fixed Effects							√
R-squared	0.07	0.06	0.06	0.07	0.07	0.10	0.38
Sample Size	2,507	2,507	2,507	2,507	2,507	2,507	2,507

Note: The dependent variable in Panel A is the principal's overall evaluation of the teacher; in Panel B it is the principals evaluation of the teacher's ability to raise student achievement. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 5: Pre-experimental Performance Evaluations and Value-Added, by Subject Area

Panel A: "Overall" Performance	Only Math	Only English	Math and English		
	(1)	(2)	(3)	(4)	(5)
Value-added in Math	0.385**		0.191**		0.103**
	(0.059)		(0.031)		(0.031)
Value-added in English		0.159*		0.216**	0.174**
		(0.075)		(0.030)	(0.030)
R-squared	0.51	0.38	0.32	0.33	0.34
Sample Size	544	456	1,507	1,507	1,507
Panel B: "Raising Student Achievement"	Only Math	Only English	Math and English		
	(6)	(7)	(8)	(9)	(10)
Value-added in Math	0.378**		0.226**		0.138**
	(0.065)		(0.032)		(0.032)
Value-added in English		0.219**		0.251**	0.180**
		(0.071)		(0.033)	(0.034)
R-squared	0.55	0.45	0.38	0.39	0.41
Sample Size	544	457	1,507	1,507	1,507

Note: Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. All specifications include school fixed effects and controls for teaching experience. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 6: Principals' Pre-experimental Ratings and Subject Specific Value-Added

	Overall Evaluation	Evaluation for Raising Student Achievement				
		Math/English	Math		English	
	(1)	(2)	(3)	(4)	(5)	(6)
Value-added, Math/English	0.004	0.054**				
	(0.010)	(0.012)				
Value-added, Math			0.028*			0.017
			(0.011)			(0.011)
Value-added, English				-0.007	0.049**	
				(0.010)	(0.012)	
Principal's Performance Evaluation						
Raising Math and/or English Achievement	0.880**					
	(0.014)					
Overall		0.878**				
		(0.015)				
Raising Math Achievement					0.924**	0.932**
					(0.015)	(0.014)
Raising English Achievement			0.873**	0.881**		
			(0.014)	(0.014)		
Restricted to Teachers of Math and English			√	√	√	√
R-squared	0.79	0.79	0.82	0.82	0.83	0.82
Sample Size	2,507	2,507	1,507	1,507	1,507	1,507

Note: Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 7: Performance Evaluations and the Precision of Value-Added Estimates

	Overall Evaluation				Student Achievement Evaluation			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-added	0.228**	0.078**	0.100**	0.100**	0.255**	0.106**	0.127**	0.129**
	(0.024)	(0.027)	(0.028)	(0.028)	(0.026)	(0.027)	(0.027)	(0.027)
Estimate Precision		0.146**	0.155**	0.154**		0.145**	0.151**	0.150**
		(0.027)	(0.024)	(0.024)		(0.028)	(0.026)	(0.027)
Value-added * Estimate Precision		0.108**	0.123**	0.111**		0.101**	0.118**	0.124**
		(0.028)	(0.030)	(0.038)		(0.035)	(0.031)	(0.038)
Teacher Experience and School FE			√	√			√	√
Years of Value-added Data FE				√				√
R-squared	0.05	0.09	0.35	0.35	0.07	0.10	0.40	0.40
Sample Size	2,507	2,507	2,507	2,507	2,507	2,507	2,507	2,507

Note: Value-added refers to estimates based on up to three years of data and comparisons with peers. Precision is measured as the inverse of the standard-error of the value-added estimate, normalized to have a minimum value of zero and a standard deviation of one. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 8: Value-Added and the Precision of Performance Evaluations

	(1)	(2)	(3)	(4)	(5)	(6)
Principal's Overall Evaluation	0.151**	0.103*	0.056			
	(0.032)	(0.046)	(0.084)			
Principal's Overall Evaluation	0.011+	0.016+	0.034*			
* Years of Experience as Principal	(0.007)	(0.008)	(0.016)			
Principal's Evaluation of Raising Achievement				0.158**	0.120**	0.054
				(0.032)	(0.046)	(0.082)
Principal's Evaluation of Raising Achievement				0.015*	0.018*	0.039**
* Years of Experience as Principal				(0.006)	(0.008)	(0.014)
Limited to Teachers with <10 Years Experience		√			√	
Limited to Teachers with <5 Years Experience			√			√
R-squared	0.08	0.06	0.08	0.10	0.07	0.09
Sample Size	1,215	815	363	1,215	815	363

Note: The dependent variable is the teacher's value-added estimate (combining math and English) based on three years of data and comparisons to peers, and only teachers with three years of data used in their value-added estimate are included in the sample. All specifications include a control for years of experience as principal, though in no regression is this coefficient statistically significant. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 9: Classroom Observations by Principals and Value-added

	Formal Observations				Total Observations			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-added	-0.081+	-0.094*	-0.052**	-0.026+	-0.075	-0.092	-0.089**	-0.035
	(0.041)	(0.039)	(0.015)	(0.015)	(0.099)	(0.099)	(0.028)	(0.028)
Overall Performance Evaluation				-0.105**				-0.217**
				(0.023)				(0.043)
Teacher Experience Fixed Effects		√	√	√		√	√	√
School Fixed Effects			√	√			√	√
R-squared	0.00	0.13	0.76	0.77	0.00	0.02	0.89	0.89
Sample Size	2,506	2,506	2,506	2,506	2,487	2,487	2,487	2,487

Note: Value-added refers to estimates based on up to three years of data and comparisons with peers. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 10: The Impact of Value-added Information on Performance Evaluations

	(1)			(2)			(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	0.123**	0.017	*0.106*	0.149**	0.038	*0.111*	0.087*	0.037	*0.05*	0.135**	0.054	*0.081*
	(0.030)	(0.026)	*[p=0.008]*	(0.033)	(0.025)	*[p=0.007]*	(0.040)	(0.033)	*[p=0.335]*	(0.047)	(0.036)	*[p=0.171]*
Overall Evaluation, Pre-experiment	0.793**	0.824**	*-0.031*	0.724**	0.760**	*-0.036*	0.758**	0.765**	*-0.007*	0.663**	0.714**	*-0.051*
	(0.033)	(0.035)	*[p=0.519]*	(0.042)	(0.040)	*[p=0.535]*	(0.071)	(0.067)	*[p=0.943]*	(0.077)	(0.079)	*[p=0.644]*
Estimate Precision												
* Value-added							0.072*	-0.013	*0.085*	0.057	-0.011	*0.068*
							(0.031)	(0.039)	*[p=0.088]*	(0.038)	(0.034)	*[p=0.183]*
* Overall Evaluation							-0.046	0.007	*-0.053*	-0.044	0.005	*-0.049*
							(0.029)	(0.035)	*[p=0.244]*	(0.030)	(0.039)	*[p=0.319]*
Years Principal Observes Teacher												
* Value-added							-0.081*	-0.016	*-0.065*	-0.082*	-0.02	*-0.062*
							(0.038)	(0.040)	*[p=0.239]*	(0.040)	(0.046)	*[p=0.309]*
* Overall Evaluation							0.115**	0.041	*0.074*	0.135**	0.031	*0.104*
							(0.034)	(0.038)	*[p=0.147]*	(0.038)	(0.045)	*[p=0.078]*
Experience Controls				√	√					√	√	
School Fixed Effects				√	√					√	√	
R-squared	0.56	0.57		0.69	0.68		0.57	0.58		0.70	0.68	
Sample Size	744	780		744	780		744	780		744	780	

Note: The dependent variable is the principal's overall evaluation of the teacher in the follow-up survey. Value-added refers to estimates based on up to three years of data and comparisons with peers. Precision is measured as the inverse of the confidence interval of the value-added estimate, normalized to have a minimum value of zero and a standard deviation of one. Years principal has supervised the teacher is equal to the minimum of the years of data used to construct the valued added estimate and the principals years of experience in the school. All specifications control for teacher experience fixed effects. Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 11: Impact of Value-added on Performance Evaluation, by Subject

	Math			English		
	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.147**	0.003	*0.144*			
	(0.032)	(0.028)	*[p=0.001]*			
Value-added, English				0.031	0.029	*0.002*
				(0.034)	(0.027)	*[p=0.963]*
Overall Evaluation, Pre-experiment	0.772**	0.819**	*-0.047*	0.818**	0.813**	*0.005*
	0.037	0.035	*[p=0.356]*	0.04	0.041	*[p=0.93]*
R-squared	0.57	0.57		0.55	0.55	
Sample Size	616	631		580	607	
	Math Only			English Only		
	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.192**	0.036	*0.156*			
	(0.054)	(0.051)	*[p=0.036]*			
Value-added, English				-0.088	-0.003	*-0.085*
				(0.075)	(0.08)	*[p=0.439]*
Overall Evaluation, Pre-experiment	0.765**	0.845**	*-0.08*	0.915**	0.842**	*0.073*
	(0.068)	(0.053)	*[p=0.354]*	(0.045)	(0.086)	*[p=0.453]*
R-squared	0.60	0.66		0.63	0.57	
Sample Size	156	161		108	129	
	Both Math & English					
	(5)			(6)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added, Math	0.121**	-0.028	*0.149*			
	(0.035)	(0.027)	*[p=0.001]*			
Value-added, English				0.025	0.034	*-0.009*
				(0.036)	(0.031)	*[p=0.85]*
Overall Evaluation, Pre-experiment	0.773**	0.804**	*-0.031*	0.798**	0.795**	*0.003*
	(0.043)	(0.046)	*[p=0.623]*	(0.047)	(0.046)	*[p=0.964]*
R-squared	0.55	0.53		0.54	0.53	
Sample Size	452	458		452	458	

Note: The dependent variable is the principal's evaluation of a teacher's overall effectiveness in the follow-up survey. Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 12: Impact of Value-Added Information on Classroom Observation

	Formal Observations		Total Observations	
	Levels	Changes	Levels	Changes
	(1)	(2)	(3)	(4)
Treatment School	0.064	0.165	-0.175	-0.210
	(0.161)	(0.126)	(0.513)	(0.525)
R-squared	0.00	0.01	0.00	0.00
Sample Size	1,523	1,523	1,520	1,520

Note: Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 13: Impact of Value-added Information on a Teachers' Propensity to Exit the School

Panel A: OLS	(1)			(2)			(3)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	-0.026*	-0.001	*-0.025*	-0.021+	0.009	*-0.03*	-0.020	0.010	*-0.03*
	(0.012)	(0.010)	*[p=0.11]*	(0.012)	(0.011)	*[p=0.065]*	(0.012)	(0.012)	*[p=0.077]*
Overall Evaluation, Pre-experiment				-0.018	-0.051**	*0.033*	-0.015	-0.034*	*0.019*
				(0.013)	(0.015)	*[p=0.097]*	(0.015)	(0.016)	*[p=0.386]*
Teacher Experience and School FE							√	√	
R-squared	0.01	0.00		0.01	0.02		0.22	0.19	
Sample Size	1,103	1,032		1,103	1,032		1,103	1,032	
Panel B: Logit	(4)			(5)			(6)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	-0.233*	-0.007	*-0.226*	-0.189+	0.095	*-0.284*	-0.159	0.134	*-0.293*
	(0.105)	(0.102)	*[p=0.123]*	(0.111)	(0.102)	*[p=0.06]*	(0.099)	(0.133)	*[p=0.077]*
Overall Evaluation, Pre-experiment				-0.160	-0.488**	*0.328*	-0.176	-0.379*	*0.203*
				(0.116)	(0.128)	*[p=0.058]*	(0.142)	(0.153)	*[p=0.331]*
Teacher Experience and School FE							√	√	
Sample Size	1,103	1,032		1,103	1,032		707	643	

Note: Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 14: Impact of Value-added on Propensity to Exit the School, by Subject

	Teachers of Only Math or Math and English					
	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added (in Math)	-0.355**	-0.03	*-0.325*	-0.304*	0.071	*-0.375*
	(0.116)	(0.126)	*[p=0.058]*	(0.125)	(0.127)	*[p=0.035]*
Overall Evaluation, Pre-experiment				-0.192	-0.441**	*0.249*
				(0.128)	(0.142)	*[p=0.193]*
Sample Size	936	844		936	844	
	Teachers of Only English					
	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added (in Math)	0.064	0.229	*-0.165*	0.059	0.251+	*-0.192*
	(0.288)	(0.142)	*[p=0.608]*	(0.291)	(0.143)	*[p=0.554]*
Overall Evaluation, Pre-experiment				0.053	-0.593**	*0.646*
				(0.222)	(0.23)	*[p=0.044]*
Sample Size	167	188		167	188	

Note: All specifications are logit regressions. Value-added refers to estimates based on up to three years of data and comparisons to peer teachers. Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 15: Student Achievement Gains, School Year 2008-2009

	Math					
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment School	0.024	0.024	0.022	0.040	0.057*	0.044+
	(0.017)	(0.017)	(0.019)	(0.025)	(0.024)	(0.023)
	[0.16]	[0.17]	[0.25]	[0.10]	[0.04]	[0.09]
Overall Evaluation (Pre-experiment)					0.074**	0.052**
					(0.011)	(0.011)
Value-added (Pre-experiment)						0.076**
						(0.010)
Teacher Experience Controls		√	√	√	√	√
Student-level Covariates			√	√	√	√
Restricted to Teachers in Pilot Sample				√	√	√
Sample Size	69,889	69,889	69,889	25,367	25,367	25,367
	English					
	(7)	(8)	(9)	(10)	(11)	(12)
Treatment School	-0.010	-0.012	0.011	-0.006	0.021	0.006
	(0.013)	(0.013)	(0.015)	(0.020)	(0.023)	(0.023)
	[0.45]	[0.37]	[0.47]	[0.76]	[0.35]	[0.80]
Overall Evaluation (Pre-experiment)					0.078**	0.067**
					(0.012)	(0.012)
Value-added (Pre-experiment)						0.048**
						(0.011)
Teacher Experience Controls		√	√	√	√	√
Student-level Covariates			√	√	√	√
Restricted to Teachers in Pilot Sample				√	√	√
Sample Size	67,835	67,835	67,835	23,603	23,603	23,603

Note: The dependent variables are gains in individual student test scores from 2008 to 2009, and regressions are estimated with school and teacher level random effects. P-values on tests of differences between treatment and control in brackets. *p<0.05, +p<0.1.

Table A1: Baseline Survey Responses for Treatment-Control Principals

	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C
Years of Experience as Evaluator	8.620	8.666	0.046	0.94
Only the Principal Contributed to the Survey	0.532	0.509	-0.023	0.73
Asst. Principal also Contributed to Survey	0.404	0.474	0.070	0.30
Lead Teacher also Contributed to Survey	0.083	0.117	0.034	0.41
Other Person also Contributed to Survey	0.128	0.16	0.032	0.50
Already Monitor Test Score Growth	0.807	0.803	-0.004	0.94
Top 2 Ways to Assess (Other than Observation) Include				
Student Work	0.892	0.857	-0.035	0.44
State Level Standardized Tests	0.775	0.75	-0.025	0.67
Feedback from Other Administrators	0.153	0.196	0.043	0.40
Feedback from Students	0.081	0.062	-0.019	0.59
Teacher Work Portfolio	0.045	0.045	-0.000	0.99
Feedback from Parents	0.018	0.036	0.018	0.42
Feedback from Other Teachers	0.009	0.036	0.027	0.18
Other School Related Tasks	0.009	0.018	0.009	0.57
Value Added Reports would be Extremely Useful for…				
Professional Development	0.818	0.83	0.012	0.81
Assessment of Staffing Needs	0.664	0.697	0.033	0.60
Assessment of Teachers	0.636	0.732	0.096	0.13
Assignment of Students to Teachers	0.564	0.679	0.115	0.08+
Tenure Decisions	0.545	0.607	0.062	0.35
Curricular Choices	0.436	0.526	0.090	0.18
Concerns Regarding Test Scores				
(1-5, 1 = Extremely Valid, 5 = Extremely Invalid)				
Tests Cannot Measure Other Important Outcomes	1.718	1.657	-0.061	0.63
Tests do not Measure Learning Well	3.064	3.179	0.115	0.39
Tests are Biased	3.155	3.161	0.006	0.97
Teachers are Not Primarily Responsible for Test Outcomes	3.591	3.839	0.248	0.12
Tests do not Measure Our Curriculum	3.591	3.697	0.106	0.48
Level of Agreement with Following Statements				
(1-5, 1 = Strongly Agree, 5 = Strongly Disagree)				
I am satisfied with teaching applicants at my school	2.550	2.58	0.030	0.81
I can select the best teachers from my applicants	2.211	2.125	-0.086	0.40
I know who the most effective teachers are in my school	1.284	1.259	-0.025	0.69
I can retain the most effective teachers in my school	1.769	1.786	0.017	0.88
I can dismiss the least effective teachers in my school	2.789	2.893	0.104	0.54
Anyone can be an effective teacher	3.266	3.393	0.127	0.41
I can improve my teachers' performance *(composite)*	1.884	2.000	0.116	0.17
Teachers in my school are cooperative/satisfied *(composite)*	1.927	1.944	0.017	0.81

Note: There are 112 treatment schools and 111 control schools. P-values indicate the statistical significance of a treatment indicator to predict the survey response.

Table A2: Follow-up Survey Responses for Treatment-Control Principals (Common Questions)

	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C
Only the Principal Contributed to the Survey	0.462	0.55	0.090	0.25
Asst. Principal also Contributed to Survey	0.462	0.39	-0.077	0.32
Lead Teacher also Contributed to Survey	0.121	0.12	-0.005	0.91
Other Person also Contributed to Survey	0.275	0.14	-0.134	0.03*
Top 4 Ways to Assess (Other than Observation) Include				
State Level Standardized Tests	0.957	0.93	-0.031	0.38
Student Work	0.817	0.84	0.022	0.70
Periodic Assessments	0.559	0.58	0.021	0.78
End of Course Exams	0.215	0.17	-0.042	0.49
Other Student Tests	0.075	0.11	0.036	0.42
Feedback from Other Administrators	0.591	0.59	0.001	0.99
Feedback from Students	0.290	0.26	-0.031	0.65
Feedback from Parents	0.183	0.15	-0.035	0.54
Feedback from Other Teachers	0.108	0.19	0.078	0.15
Teacher Work Portfolio	0.129	0.10	-0.030	0.54
Other School Related Tasks	0.075	0.09	0.011	0.79
To Evaluate Individual Teachers in Past Year, Principal Used				
Average State Test Scores	0.859	0.94	0.079	0.09+
Average State Test Scores by Subgroup	0.761	0.81	0.049	0.44
Average Growth in State Test Scores	0.815	0.91	0.097	0.07+
Value-Added Reports *(Treatment Only)*		0.55		
Percentage of Students Not Meeting Standards on State Tests	0.856	0.86	0.007	0.90
Percentage of Students by Proficiency Level	0.913	0.93	0.012	0.78
Change in Percentage of Students by Proficiency Level	0.846	0.88	0.029	0.59
If Using Student Tests to Assess An Individual Teacher, How Important is it to Consider the Following Isssue (1-5, 1=Not Important at All, 5 = Very Important)				
Mean of All 12 Items Below	3.612	3.81	0.196	0.07+
Teaching Experience	3.615	3.74	0.128	0.46
Prior Performance of Students on Standardized Tests	4.582	4.37	-0.211	0.08+
Percentage ELL/Special Education Students in Class	4.099	4.30	0.205	0.23
Class Size	3.578	3.81	0.232	0.21
The Number of Students who Entered the class mid-year.	3.533	4.01	0.479	0.01*
Which Teacher(s) the Students Had in the Previous Year	3.678	4.12	0.438	0.00*
If a Teacher Recently Started Teaching a New Grade/Subject	3.912	4.13	0.216	0.17
If a Teacher had a Personal Issue During the Year	3.367	3.66	0.292	0.10+
Things that Distracted the Teacher's Class on the Test Day	3.067	3.12	0.051	0.81
Outside Help a Teacher's Students Received	3.811	4.00	0.189	0.21
Help a Teacher Received from an Aide in the Classroom.	3.111	3.21	0.097	0.58
The Teacher's Performance in Teaching Non-tested Subjects	3.297	3.54	0.242	0.16

Note: This table is based on the survey responses of 82 treatment school principals and 93 control school principals who partially or fully completed the second portion of the follow-up survey. P-values indicate the statistical significance of a treatment indicator in a principal level regression.

Table A3: Follow-up Survey Responses for Treatment Principals

	Treatment Mean
Principal Received Professional Development	0.94
Principal Received Value-Added Reports	0.85
Principal Examined Value-Added Reports	0.84
Principal Shared the Reports with	
Assistant Principal	0.95
Lead Teacher	0.74
Teachers	0.51
School Support Organization	0.27
Superintendent	0.10
Network Leader	0.09
Union Representative	0.03
Parents	0.02
(1-5 Scale) The Value-added Reports...	
Contain Information Useful to Principals	4.29
Contain Information Useful to Teachers	4.05
Are Easy to Understand	3.36
Have Helped Me Better Understand Differences Between Teachers	3.59
Have Enhanced my Plans for Improving Instruction in my School	3.73
(1-5 Scale) How Useful Would Annual Value-Added Reports be for …	
Designing Professional Development for Teachers	3.76
Assigning Students to Teachers	3.89
Choices of Curricula or Instructional Programs	3.27
Assessing Staffing Needs	3.59
Teacher Evaluation	3.86
Principal is Confident that Value-Added Calculations Account for... (Yes = 1, No = 0)	
Teaching Experience	0.77
Prior Performance of Students on Standardized Tests	0.76
Percentage ELL/Special Education Students in a Teacher's Class	0.48
Class Size	0.40
The Number of Students who Entered the Class Mid-Year.	0.27
Which Teacher(s) the Students Had in the Previous Year	0.45
If a Teacher Recently Started Teaching a New Grade/Subject	0.53
If a Teacher had a Personal Issue During the Year	0.08
Things that Distracted the Teacher's Class on the Test Day	0.18
Outside Help a Teacher's Students Received (e.g., after-school)	0.10
Help a Teacher Received from an Aide in the Classroom.	0.13
The Teacher's Performance in Teaching Non-tested Subjects	0.07

Note: 84 treatment schools responded to the follow-up survey, but only 79 completed the second section (after evaluating their teachers) and only 66 principals who claimed to have received and examined the reports were asked the remainder of these questions.