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

Jonah E. Rockoff
Columbia Business School

Douglas O. Staiger
Dartmouth College

Thomas J. Kane
Harvard Graduate School of Education

Eric S. Taylor
Stanford University

February, 2011[*]

Abstract

Considerable theory regarding how employers learn about worker productivity remains untested. Examining the provision of objective estimates of teacher performance to school principals, we establish several facts supporting a simple Bayesian learning model with imperfect information. First, the correlation between performance estimates and prior beliefs rises with more precise objective estimates and more precise subjective priors. Second, new information exerts greater influence on posterior beliefs when it is more precise and when priors are less precise. Employer learning also affects job separation and productivity in schools, increasing turnover for teachers with low performance estimates and producing small test score improvements.

[*] Corresponding author: jonah.rockoff@columbia.edu. We thank seminar participants at Northwestern University, Stanford University (Business School and School of Education), Harvard University (Economics Department, Kennedy School, and School of Education), Columbia University (Business School and Teachers College), University of Virginia, University of Florida, UC Davis, University of Oregon, University of Tel Aviv, the Research Institute of Industrial Economics (IFN), and the London School of Economics 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.

A substantial theoretical literature focuses on how employers with imperfect information learn about employee productivity, with seminal work by Spence (1973) and Jovanovic (1979). In contrast, empirical research on employer learning has developed more slowly, particularly with regard to subjective evaluation of employees (see Farber and Gibbons, 1996; Oyer and Schaefer, 2011).[1] In this study, we use micro-data to examine how new information on worker performance is incorporated into employee evaluations. Specifically, we analyze the results of a randomized pilot program where principals in New York City schools received new performance measures on their teachers based on student test score outcomes.

We base our analysis on a simple Bayesian learning model in which principals use imperfect information to learn about teacher productivity. This provides us with several empirical predictions for the relationship between teacher performance data and principals' prior and posterior beliefs which we test. All of the model's predictions are borne out in our data. There is a strong relationship between student test-based performance measures—called "value-added" measures—and principals' baseline evaluations of teacher effectiveness, and this relationship is stronger when value-added estimates and principals' priors are relatively more precise. More importantly, principals who are provided with objective performance data incorporate this information into their posterior beliefs, and do so to a greater extent when the new information is more precise and when their priors are less precise.

We then investigate the impact of providing new information on managerial decision-making. First, although principals endogenously allocate time spent observing teachers, we find

[1] The most well-known empirical studies on this topic focus on the consistency of wage trajectories with employer learning (e.g., Murphy, 1986; Gibbons and Katz, 1991; Farber and Gibbons, 1996; Altonji and Pierret, 2001; Lange, 2007). Other related work has provided descriptive evidence on how managers form evaluations of employee productivity. Baker et al. (1988) and Murphy and Oyer (2001) document variation in the use of subjective and objective performance evaluations across occupations, while other studies highlight managers' reluctance to differentiate among employees (e.g., Medoff and Abraham, 1980; Murphy, 1992) and potential bias in subjective evaluations (see Prendergast and Topel, 1993).

no evidence that the receipt of "hard" performance data crowded out the collection of "soft" information through classroom observation. Second, we find that teachers with low performance estimates were more likely to exit their schools after this information was provided to principals. This change in employee retention implies a small improvement in average teacher productivity in schools where the information was provided, and we find changes in student achievement in line with these expectations.

This is, to our knowledge, the first rigorous and detailed test of how new information affects employers' subjective beliefs about worker productivity. In addition, our findings have importance implications for the current policy debate regarding performance estimates for teachers based on student outcomes. A series of papers by economists interested in educational production have demonstrated that productivity varies greatly across teachers (Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007), and two recent papers argue in favor of using estimates of teacher value-added—based on standardized tests—to evaluate teachers (Gordon et al., 2006; Kane and Staiger, 2008). Motivated in part by this research, the federal government has encouraged states and school districts to use student achievement growth in to measure teacher effectiveness as part of the incentives built into its $4.3 billion Race to the Top Fund. Nevertheless, some have questioned whether such measures are sufficiently reliable and valid indicators to be used in high-stakes personnel decisions (Economic Policy Institute, 2010; Corcoran, 2010). Our results suggest that value-added estimates provided new information to principals and helped them to make personnel decisions that improved student performance.

Research on evaluation and incentives in employment contracts provides several reasons why simply providing value-added estimates to school administrators, without any explicit link

to teacher pay or promotion, may improve the quality of education provided in public schools.[2] First, many teachers work for "idealistic reasons or because they enjoy working with children" (Dixit, 2002), and variation in performance among such "motivated agents" is likely due to persistent variation in skills and not effort provision.[3] Second, because teaching is a multi-dimensional task with multiple goals, employee evaluations should be based on holistic, subjective judgments rather than simple objective performance measures (Holmstrom and Milgrom, 1991) in order to avoid dysfunctional behaviors like cheating on tests (Jacob and Levitt, 2003).[4] Third, in addition to providing information to the principal, the verifiability of an objective performance measure can help the principal act on available information (Baker, Gibbons, and Murphy, 1994).

Our findings support the notion that providing measures of teacher productivity (based on student test scores) is useful to school principals and helps improve school quality. Principals incorporated the information into their posterior beliefs while taking into account the precision of the new information and the precision of their priors. Provision of this new information led to a change in patterns of retention, so that teachers with higher performance measures were more likely to stay and those with lower performance measures more likely to leave, and, in line with these changes in turnover, led to small improvements in student achievement the following year.

In Section 2 we describe the program we study, present comparisons of treatment and control groups, provide 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

[2] For broader reviews of this literature, see Prendergast (1999) and Gibbons (1998, 2005).
[3] See also Besley and Ghatak (2005). This description may not be entirely appropriate in developing countries, where teachers are often absent or are found present but not teaching when audited (Chaudhury et al. 2006).
[4] The empirical literature on rewarding teachers based on objective performance measures is limited, partially due to opposition to such contracts by unions (see Ballou, 2001). Randomized field studies in the U.S. have found little benefit to merit pay schemes (Turner and Goodman, 2009; Springer et al., 2010) but those outside of the U.S. have been more promising (Lavy, 2002, 2009; Muralidharan and Sundaraman, 2008).

analysis of the impact of the treatment. We examine the relationship between the value-added performance estimates and principals' prior beliefs regarding teachers in Section 4, and we estimate impacts of the provision of performance data in Sections 5 and 6. Section 7 concludes.

2. The Teacher Data Initiative

In 2007, the New York City Department of Education (hereafter NYC) began a program to estimate teachers' "value-added" based on student test scores and disseminate this information to principals in reports that they can understand. NYC officials believed that many principals had too little capacity to access and analyze data, too little scope with which to compare teachers within their schools, and too little training in evaluating teachers. They therefore felt that value-added estimates could provide principals with new and potentially useful information.[5] Finally, recent changes to management and accountability systems had given NYC principals more decision-making power and responsibility, and the value-added reports were seen as a means of empowering principals without imposing choices upon them.[6]

In order to understand the impact of this initiative on principals, teachers, and students, NYC piloted the provision of teacher performance estimates using a randomized control trial. In early summer 2007, NYC 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

[5] There is still considerable discussion surrounding the assumptions underlying value-added, particularly due to non-random matching of students and teachers (see Todd and Wolpin, 2003; Rothstein, 2009). Our paper does not focus on these concerns and 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). However, NYC officials were aware of this debate, and felt that principals would have local knowledge regarding student assignment that could help them interpret estimates of teacher value-added in the context of any peculiar matching of students and teachers.

[6] Principals in NYC had received greater power to allocate financial resources within their schools and, under the NYC accountability system, could earn up to $25,000 for good performance or be removed for poor performance.

of the grades 4 to 8.[7] 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 signed up to become part of the program and were sent a baseline survey on August 8, 2007. 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, and 223 principals completed the survey.[8] Schools were assigned a random number, sorted by number within grade configuration (i.e., elementary, middle, and K-8 schools), and 112 schools (73 elementary, 27 middle, and 12 K-8 schools) were selected into treatment group; 111 schools remain in the control group.

In order to participate, eligible principals had to volunteer and complete the baseline survey, and only about one in four did so. While this suggests checking for differences between volunteers and non-volunteers (which we do here), this level of participation is not unusually low. For example, in Project STAR, a well-known class size reduction experiment (see Krueger, 1999; Chetty et al., 2010) the fraction of eligible schools that participated was roughly one in five. Take-up was roughly one in four in a recent experiment offering free management consulting to mid-sized Indian manufacturing firms (Bloom et al., 2010). While we cannot rule

[7] NYC 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.

[8] The Battelle Memorial Institute, under contract with NYC, collected all survey data. Battelle also provided professional development to principals, estimated teacher value-added, and prepared the value-added reports. We do not discuss the value-added estimation methodology here, but details were provided to participating principals in a technical report which is 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 was not done for teachers of 8th grade (due to a lack of 9th grade test data) or for teachers observed for the first time in the school year 2006-2007 (the most recent year of data used to estimate value-added).

out selection of volunteers based on unobservable characteristics, we find little evidence that participating principals (schools) were substantially different from those serving students in grades 4-8 citywide (Table 1, leftmost columns).[9] On a wide range of observable characteristics, we find only a few statistically significant differences, and 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 (79 vs. 73 percent). Importantly, we find no significant differences in teachers' responses to a citywide survey from spring 2007, including the frequency with which the principal conducts classroom observations, the quality of feedback teachers receive from the principal, how much the principal prioritizes teaching, or the use of student data in instruction.

Two additional factors may be important in interpreting our findings. First, NYC launched a new accountability system in the fall of 2007 built partly around the math and English exams taken by students in grades 4-8, and principals thus had incentives to focus on teacher performance in these grades and subjects (Rockoff and Turner, 2010). On the other hand, the broader political climate may have reduced principals' incentives to use the value-added reports. The teachers' union did not support the program and filed a grievance in the fall of 2007, after principals had signed up for the program but before they received the reports. Principals receiving the reports were advised by NYC that they were *not* to be used for formal teacher evaluation during the pilot year, and in April 2008 teachers' unions successfully lobbied a change in state law so that teachers could "not be granted or denied tenure based on student

[9] It is also unclear how selection on unobservables may impact our findings. 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. Because we find little evidence of selection on observables, it is also quite possible that participation may have been based on exogenous factors, like the principal's availability to attend one of the information sessions conducted in summer 2007.

performance data."[10] Importantly, the identities of participating principals were not made public, but the pilot's existence was well known (New York Times, January 21, 2008).

A more immediate consideration is the internal validity of the research design. To assess the comparability of treatment and control groups, we compare the average characteristics of treatment and control principals/schools at baseline and test for statistically significant differences (Table 1, middle columns). 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 in the spring of 2007. Thus, we have great confidence that 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 in December of 2007.[11] 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. Principals who attended/viewed training sessions completed a short survey instrument to provide feedback to NYC; 95 percent 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. Importantly, NYC did not distribute reports to 16 treatment principals who did not attend or view a training session, but in our analysis these principals are included as part of the treatment group.

[10] In order to qualify for funding under the Race to the Top Competition, this legislation was repealed and New York will now incorporate student achievement growth measures into formal teacher evaluation starting in 2012.
[11] These training sessions were held over several evenings. The first two hours focused on the value-added estimation, a walk-through of a sample value-added report, and discussion of potential uses of the reports. Principals then received their teachers' reports, and the remaining hour was a question and answer session.

A sample value-added data report is included as Appendix Figure 1. Value-added measures were based only on the 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.[12]

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 the 16 treatment principals who did not receive reports; only 5 of them 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.

[12] The report also presents value-added estimates specific to subgroups (e.g., English Language Learners, Special Education students). 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 principals that responded to the follow-up survey and compare treatment and control principals on the same characteristics as the baseline sample (Table 1, rightmost columns). 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. For non-survey outcomes (e.g., teacher turnover, student test scores), we include all participating schools in our analysis, regardless of survey response.

In December of the following school year (2007-2008), NYC expanded dissemination of teacher data reports to all elementary and middle schools. This does not affect the vast majority of our analysis, since the 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, both treatment and control principals had received value-added reports when tests were taken in the spring of 2009. 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, this fact should be kept in mind when considering our findings.

2.1 Baseline and Follow-up Surveys

First and foremost, the baseline and follow-up surveys asked principals to provide a subjective evaluation of the performance of their teachers. Specifically, principals were given a list of the teachers in their schools who taught math and/or English to students in grades four through eight, and were asked to evaluate each teacher "overall," and then specifically "in terms of raising student achievement" in math, English, or both (for teachers of 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 experience. Evaluations were made on a six point scale: Exceptional (top 5 percent), Very Good (76-95th percentile), Good (51-75th), Fair (26-50th), Poor (6-25th), or Very Poor (bottom 5 percent).

Descriptive statistics for teachers in our sample are presented in Table 2. On a 1-6 scale, the mean overall evaluation was 4.3, with a standard deviation of 1.1 points.[13] Perhaps not surprisingly, correlations between the overall evaluation and the subject specific evaluations are quite high (0.87 for math and 0.88 for English), as is the correlation between subject specific evaluations for teachers of both subjects (0.91). While over three quarters of teachers received an overall evaluation indicating above median performance (Good, Very Good, or Exceptional), variation in these low-stakes responses is greater than variation in formal evaluations of NYC teachers, where 98 percent are deemed "Satisfactory" and just 2 percent "Unsatisfactory".[14]

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. 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.[15] 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

[13] From Exceptional to Very Poor, percentages in each group were 13.7, 33.6, 30.2, 17.2, 3.8, and 1.6, respectively.
[14] MacLeod (2003) demonstrates how compression in subjective evaluations is a feature of optimal contracting. See Weisberg et al. (2009) for more evidence on principals' reluctance to give teachers poor formal evaluations.
[15] Formal observations are part of the official NYC teacher evaluation system. Untenured teachers in elementary and middle schools in NYC 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. In our analysis we code "more than ten" observations as having a value of 11.

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.

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.[16] 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

[16] 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.

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 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 for 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 was student performance on state standardized tests, cited by 96 percent of control principals and 93 percent of treatment principals. Thus, these principals would almost certainly have used student tests to evaluate teachers 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 and treatment principals were also more likely to claim using 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.[17] This provides an initial indication that the treatment changed the set of information principals incorporated into their evaluations.

[17] 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.

Finally, treatment principals' confidence in whether the value-added methodology controlled for various factors largely accords with reality (see Table A3). For example, 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).[18]

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 NYC 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 2), suggesting participating schools was similar to other schools in NYC on average. Not surprisingly, 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 math than English, consistent with other studies from 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 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.

[18] 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 NYC or left NYC's teacher workforce. Roughly 89 percent of the teachers in the baseline survey still worked in NYC in the school year 2007-2008, and 85 percent were teaching in the same school. Roughly 82 percent still worked in NYC in the school year 2008-2009, and 75 percent were teaching in the same school.[19]

Finally, we have data on student test scores and demographics (i.e., gender, race/ethnicity, free lunch receipt, English language learner, and special education status) for the school year 2008-2009. 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

[19] We find marginally significant differences in exit rates between treatment and control schools between our baseline survey and the fall of 2007 (16 vs. 13 percent), prior to the release of the value-added reports. Given the number of variables tested, this is likely due to chance and should not affect the validity of our analysis. We also examine turnover during this "pre-experimental" period as a robustness check in, discussed in Section 6.2.

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.[20]

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

It is straightforward to show this simple model yields two intuitive empirical predictions:

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.

[20] 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.[21]

While we focus on how value-added information affected the principals' beliefs, we also use our data to test several reasonable hypotheses about how the value-added information affected the principals' actions. First, if classroom observation is costly, then treatment principals could respond to the provision of value-added data 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, either through changes in principals' posterior beliefs, or providing independent and verifiable confirmation of their priors, along the lines of Baker et al. (1994). Finally, changes in turnover or resource allocation due to new information may improve overall educational quality in treatment schools, in which case we would expect to see student achievement rise, relative to control schools.

4. Value-added Estimates and Principals' Priors

We start by testing Prediction 1 from our framework. The most basic part of this prediction—that value-added estimates and principals' prior beliefs should be positively

[21] We lack data on teachers' complete work histories but we know their years worked in value-added subjects/grades in the current school. To measure years during which the principal had likely observed the teacher, we take the minimum of this variable and the number of years the principal worked in the current school. We do know teachers' total years of experience, but this is likely to overestimate experience within a school for a large fraction of teachers.

correlated—has been shown before (e.g., Murnane, 1975; Armor et al., 1976; Jacob and Lefgren, 2008; Harris and Sass, 2009), and it will come as no surprise that we confirm this finding. However, in addition to a number of other new results, the test of how precision mediates the strength of this relationship is unique to our study.

4.1 Baseline Correlations of Value-Added with Prior Beliefs

We measure the relationship between value-added estimates and principals' prior beliefs using linear regression. First, we estimate 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.

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

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 3, 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).[22] 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 3, 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.[23]

Previous research has found that the relationship between value-added estimates and subjective beliefs regarding teacher effectiveness is somewhat stronger in math than English (e.g., Jacob and Lefgren, 2008; Rockoff and Speroni, 2010). We examine this by running regressions separately by teachers of math, English, or both math and English. Principals' prior beliefs were not clearly tied more strongly to one of the two subject areas, and we always estimate positive and significant coefficients on value-added (see Appendix Table A4).[24]

[22] Coefficient estimates on experience indicators are available upon request. Conditional on value-added, baseline evaluations were lowest for teachers who just completed their first year and highest for teachers with three to nine years of experience, while those with only a few years of experience or ten or more years were rated in the middle.
[23] We pool treatment and control groups for power, but our findings are not substantially or statistically different if we examine group separately. For example, in the specification shown in Column 14, the value-added coefficient is 0.277 for treatment schools only (standard error 0.029) and 0.263 for control schools only (standard error 0.044).
[24] For teachers of only math, the coefficient on value-added is 0.4, while for teachers of only English the effect size is 0.2. However, among teachers of both subjects, the coefficient estimates were larger for value-added in English than in math, either when estimated separately or together in the same regression.

In several additional tests, presented in Appendix Table A5, we find that principals' subjective evaluations are surprisingly nuanced in how they capture variation in teacher performance as measured by the value-added estimates. There is a very high correlation between principals' evaluations of teachers' "overall" performance and performance in "raising student achievement" (0.87). Nevertheless, when these evaluations are not in agreement, their differences are also in line with value-added estimates (Table A5, Columns 1 and 2).[25] Similarly, for teachers who teach both subjects, there is a very high correlation between the principal's evaluations of performance in raising math achievement and English achievement (0.91), but, again, the differences in subject-specific evaluations tend to be in line with differences in teacher value-added (Table A5, Columns 3-6).[26]

4.2 Baseline Relationships and the Precision of Information

Having established a consistent baseline relationship between value-added and principals' performance evaluations, we proceed to test the more novel 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.[27] Consistent with the Bayesian

[25] Specifically, value-added can predict the principals' evaluation of a teacher's ability to raise student achievement even while controlling for a principal's evaluation of overall performance, but value-added has no power to predict a principal's evaluation of overall performance while controlling for a principal's evaluation of a teacher's ability to raise student achievement.

[26] Conditional on the principal's evaluation of math performance, value-added in math does not predict the principal's evaluation of teacher performance in English, though value-added in English remains a significant predictor. The results are symmetric for the two subjects. In other words, where a principal's evaluations of a teacher's math performance and English performance differ, this difference tends to coincide with the teacher's objective performance estimates in math versus English.

[27] 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.

learning model, value-added estimates are more strongly correlated with principals' priors when those estimates have tighter confidence intervals (Table 4). Our estimates of the interaction of value-added and precision are positive and highly significant. The estimates are very similar regardless of whether we predict the principal's overall evaluation (Columns 1 to 3) or evaluation of the teacher's ability to raise student achievement (Columns 4 to 6), and also similar when 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 available to generate value-added estimates may tend to be both more effective, even conditional on total teaching experience, and have smaller confidence intervals. 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).[28]

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. Second, to incorporate the strength of principals' priors we focus on the length of the working relationship for each principal-teacher pairing. Specifically, we interact the evaluation with the years the principal has held that position in the school, and limit the sample to teachers with three years of value-added, i.e., teachers whom we are sure have worked in the school for at least three years. For this sample, we can be sure that 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.

[28] Another concern may be that the most precise estimates are for teachers of only math, since value-added was a stronger predictor of principals' evaluations for these teachers (see Appendix Table A5). However, we find positive and significant interactions of value-added and precision when separately examining teachers by subject area.

The empirical results are in line with this expectation and the Bayesian model. Both the main effect of principals' evaluations and the interaction with principal experience are positive and statistically significant (Table 5). 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 principals are more likely to have performance information on very experienced teachers that is not captured by the available value-added estimate: (a) teachers may have taught other non-tested subjects or grade levels, or (b) teachers may simply have had variation in their performance that is not captured by the last three years of data. 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).[29] 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.

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).

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

[29] We also estimated a specification interacting baseline evaluations with an indicator for whether the principal strongly agreed with the statement "I know who the most effective teachers are in my school." This interaction was positive, as we would predict, but small and statistically insignificant. This 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.

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 confirm our prediction that providing value-added estimates to principals had a significant impact on their 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 6, 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 estimate for the treatment group is smaller as expected (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.[30] 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

[30] 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 NYC experience is likely to misclassify many experienced teachers who have changed schools.

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

Thus, our findings are quite consistent with the Prediction 2 of the Bayesian 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 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 7). Specifically, we find positive significant effects of math value-added on posterior evaluations for teachers of math and for teachers of math and English for the treatment group, while the control group estimates are much smaller and not distinguishable from zero (Column Groups 1, 3, and 5). Meanwhile, we find no significant effects of English value-added estimates on posterior beliefs, and 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 measure

teachers' contributions to student achievement accurately (this was discussed with principals in the training sessions). It may also be that principals were more confident in their ability to gauge instructional quality in English, and thus 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, given that the evaluations provided by principals were unofficial and carried no stakes, it is important to test whether the provision of new information actually translated into changes in personnel decisions or the quality of education provided at the school. In this section, we examine 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 Classroom Observation

The time which principals spend observing teachers in the classroom is valuable and could be spent on other duties. Using baseline data, we first test whether principals allocate time spent on teacher observation according to prior knowledge and beliefs. We then examine whether providing principals with new "hard" data on teacher performance crowded out time spent on gathering "soft" data via classroom observation during the pilot year.

To examine whether principals' allocate their time observing teachers strategically, we regress observations made of teacher *i* in school *j* the year *prior* to the pilot (O_i) on the years the principal has observed the teacher (T_i), which we previously found to moderate the impact of new information (Table 6), the principal's baseline evaluation (R_i), and school fixed effects.

$$(4)\ O_{ij} = \lambda T_{ij} + \beta R_{ij} + \delta_j + \varepsilon_{ij}$$

We expect our estimate of λ to be negative, reflecting the declining value of new information as a principal learns over time. We also might expect β to be negative, since a goal of observation may be to identify ineffective teachers and provide them with constructive criticism.

Estimates for both formal and total observations provide clear support for the hypotheses that principals allocate more time to observing teachers who they know less well or who they believe are performing poorly.[31] An additional year of having observed the teacher in their current role reduces both formal and total observation frequency by about 0.25, while a standard deviation increase in the principal's baseline evaluation reduces formal observation frequency by 0.1 and total observation frequency by 0.2 (Table 8, Columns 1 and 3). One concern with these specifications is that the number of years a teacher has been observed by the principal will be positively correlated with teaching experience, and it may simply be experience rather than principals' knowledge that causes them to observe the teacher less often. Controlling for experience does cause the coefficients on years of observation to fall substantially, but they remain statistically significant (Table 8, Columns 2 and 4).

Despite the finding that principals do allocate their time strategically, we find no evidence that providing principals with objective performance data led them to spend less time in the classroom. Regressing observation frequency during the pilot year on an indicator for treatment we find no evidence that treatment affected principals' propensity to observe teachers in the classroom (Table 9). The point estimates are positive for formal observations and negative for total observations, and are all statistically insignificant regardless of whether we look at levels or changes from baseline. However, it should be kept in mind that, given the structure of

[31] We drop a few teachers from these regressions for whom the principal reported formal observations but left the question on total observations blank. We *include* 15 schools (6 treatment, 9 control) in which the principal was asked about at least five teachers and reported the same number of formal and total observations in each case. These principals likely have a uniform observation policy, and all of the coefficients increase slightly if we exclude them.

the program and the limitations of our data, the effects would have had to take place in the short window between when reports were received in December 2007 and the end of the pilot year.[32]

6.2 Employee Turnover

There are two main channels through which the new information provided in the value-added reports might impact employee turnover. First, it might have a direct impact by changing principals' posterior beliefs, lowering evaluations of teachers they might have sought to retain or raising evaluations of teachers they may have sought to let go. Second, along the lines of Baker et al. (1994), the presence of verifiable third-party information on performance may increase a principal's willingness to act on available information, even if the principal's posterior remains unchanged. Existing studies find that principals are reluctant to remove teachers for poor performance, even when the administrative costs of doing so are minimized (see Jacob, 2007, 2010), suggesting reputational or social costs associated with dismissing a teacher are important.

To examine how turnover was affected by the information on value-added, we use two regression specifications, shown by Equations 5a and 5b:

$$\text{(5a)}\quad E_{it+1} = \beta V_{it} + \varepsilon_{it+1}$$

$$\text{(5b)}\quad E_{it+1} = \gamma V_{it} + \lambda R_{it} + \zeta_{it+1}$$

E_{it+1} is an indicator for whether teacher *i* is no longer employed in the same school at time *t+1* (i.e., in the year after the pilot), 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. While it would be preferable to examine involuntary and voluntary exits separately, there is no way to separate

[32] We are also limited to the sample of principals who completed the follow-up survey and the teachers who remained working during the pilot year. Repeating the specifications shown in Table 8 for this subset of principals and teachers provides very similar coefficient estimates. Other regressions, not reported in Table 9 but available upon request, show no significant interaction between treatment status and other variables including value added, the precision of value-added, the years the principal has observed the teacher, or the principal's baseline evaluation.

these phenomena in our data. 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. Again, we run regressions separately for treatment and control groups and test for differences between the groups, clustering standard errors at the school level.

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 10). 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).[33]

As stated above, the impact of new information on retention outcomes may be due to "direct" effects on principals' evaluations or "indirect" effects on their willingness to act, conditional on the evaluation. To distinguish these two explanations, we take advantage of the fact that only the information on *math* value-added had an impact on principals' posterior beliefs,

[33] 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.

despite the fact that math and English value-added were both strongly related to principals' priors at baseline. If the impact of information on turnover is due to direct effects on principals' posteriors, then these impacts should be driven by math value-added, not English.

This is precisely what we find. We regress exit outcomes on subject-specific value-added measures, first examining all teachers of math and English; those that teach both subjects appear in both regressions. Exit is negatively related to math value-added for math teachers in the treatment group but not the control group (Table 11, Column Group 1) and the difference in coefficients is statistically significant (p-value 0.04), while English value-added is not significantly related to exit for English teachers (and has coefficients quite close to zero) in either group (Table 11, Column Group 2). To be sure that these differences are not driven by sample, we estimate the same specifications for teachers of both math and English (Column Groups 3 and 4). These results are robust to including value-added of both subjects in the same regression (Column Group 5) and including teacher experience and school fixed effects (Column Group 6). This supports the interpretation that the primary impact of new information was to change principals' posterior beliefs, and this change in beliefs in turn altered patterns of employment.

6.3 Student Achievement

Providing managers with information on employee performance could be detrimental to firm productivity if that information is invalid or misleading. Thus, in addition to showing how principals incorporated information into their evaluations and personnel decisions, it is important to test whether the information led to improvements in school productivity. If the value-added estimates are valid predictors of how teachers will perform in the future, the effect of information on turnover suggests we should see a slight increase in student achievement in math, due to

selective retention. Productivity in treatment schools might also rise if principals use this information to reallocate resources, such as providing training to low performing teachers.

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 a treatment school. Taking advantage of the randomization of treatment, we allow for random effects at the school and teacher level to account for the nested structure of the data and increase efficiency. We find a small but marginally significant improvement in math achievement gains of 0.024 student level standard deviations (p-value 0.16) among treatment schools (Table 12, Column 1). This estimate is insensitive to adding additional controls for teacher experience—though we lack experience data on teachers hired after the pilot—or student characteristics and grade level covariates (Column 2 and 3).[34] As expected, we find no evidence of significant improvements in English; the treatment coefficients range from -0.01 to 0.01 depending on our control variables (Columns 7 to 9) and their p-values are all above 0.4.

Due to turnover or reassignment to non-tested grades or subjects, fewer than half of the students included in these regressions were taught by a teacher who was working in these schools during the pilot study. When we limit our sample to students taught by teachers in the pilot, the point estimate on treatment rises to 0.04 in math and has a p-value of 0.10 (Column 4), while the estimate in English remains quite close to zero (Column 10). Thus, math achievement gains produced by teachers in the pilot were significantly higher in the treatment schools.

When we include principals' overall evaluations of teachers at baseline as a covariate (Columns 5 and 11) we find positive significant coefficients (0.074 in math, 0.078 in English).

[34] 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 students in middle schools).

Moreover, the treatment effect estimates rise (0.057 in math, p-value 0.04, 0.021 in English, p-value 0.35), implying that, among teachers from the pilot, those in the treated group were originally rated worse, on average, than those in the control group. This is consistent with our finding that the new information caused treatment principals to keep some teachers of whom they held a low opinion and 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 positive and significant coefficients in both subjects—the treatment coefficients shrinks back towards zero. That is, conditional on the principal's baseline opinion, treatment school teachers from the pilot had higher value-added than control school teachers. Importantly, the positive significant coefficients on principals' baseline evaluations and teachers' value-added suggests that both measures are, as we posited in our conceptual framework, imperfect but useful estimates of teacher effectiveness.[35]

7. Conclusion

We study a pilot experiment providing "value-added" estimates of teacher performance to New York City school principals to learn about how managers evaluate employee job performance and how objective performance data influences this process. Using detailed micro-data, we present a number of empirical facts consistent with a simple Bayesian model where principals learn over time with imperfect information. First, the positive relationship between value-added and principals' prior beliefs about teacher effectiveness 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' priors. Second, principals change their evaluations of teachers in response to the new information contained in the value-

[35] 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.

added estimates, and the impact of this new information rises with its precision and falls with the precision of the principals' prior beliefs. Thus, value-added estimates are capturing a dimension of teacher performance valued by school principals and are providing new information to principals on this performance dimension.

Importantly, the provision of new information also had real impacts on personnel decisions and productivity. Teachers with lower value-added estimates were more likely to exit their schools, and students show small, marginally significant improvements in test scores that are consistent with these changes in selective retention. Importantly, comparisons of impacts across subject areas (math vs. English) suggest that these changes arose via the influence of new information on principals' posterior beliefs. Despite the fact that objective performance data in both English and math bear strong relationships with principals' priors, the changes in principals' posterior beliefs, patterns of teacher turnover, and student test score improvements all are driven by the provision of performance data in math.

In addition to furthering our knowledge of employer learning and subjective evaluation, our results add to the growing literature on the benefits of government provision of information in public settings where acquisition may be costly, such as restaurant hygiene (Jin and Leslie, 2003), hospital quality (Dranove et al., 2003; Hibbard et al., 2005; Chassin et al., 2010), and school quality (Hastings and Weinstein, 2008; Andrabi et al., 2009). Overall, our results support the notion that performance data on teachers can be useful to principals in managing their schools, and provide modest support for current federal and state policies moving school districts towards the use of student achievement in evaluating teachers.

However, the privacy of such information may affect its usefulness as a policy tool. In New York City, information on teacher value-added is private to principals in the school where a

teacher is currently employed. While our study shows that this information tends to increase turnover of low value-added teachers, many of these teachers found another job within the school district, and Boyd et al. (2007) find that low value-added teachers who transfer schools continue to perform poorly. Thus, the privacy of performance information helps to create a "dance of the lemons" (Ravitch, 2007), where a lack of information at the hiring stage allows teachers with low productivity to continue teaching.

On the other hand, the limitations of objective performance data may be a greater concern when the public release of information puts it in the hands of individuals with no other information at their disposal. In August 2010, the Los Angeles Times generated controversy by publishing value-added estimates for teachers and, as of this writing, a judge in New York City has ruled that the city must release its value-added estimates as part of a Freedom of Information Act request by several local newspapers. The general public may not have the training or complimentary sources of teacher performance information that helped the principals in our study to correctly interpret these measures. The usefulness of this information to interested parties such as parents and the incentives generated by public information are important questions for further research.

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.

Altonji, J.G. and Pierret, C.R. (2001) "Employer Learning and Statistical Discrimination," Quarterly Journal of Economics, 116(1): 313-350.

Andrabi, T., Das, J., and Khwaja, A.I. (2009), "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets," Unpublished Working Paper.

Armor, David, Patricia Conry-Oseguera, Millicent Cox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools*. Santa Monica, CA: Rand Corporation.

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.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd , H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., and Shepard, L.A. (2010) "Problems with the Use of Student Test Scores to Evaluate Teachers," Economic Policy Institute Paper #278.

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.

Bloom, N., Eifert, B., Mahajan, A. McKenzie, D., and Roberts, J. (2010) "Does Management Matter? Evidence from India," Unpublished Manuscript.

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

Chassin, M.R., Loeb, J.M., Schmaltz, S.P., Wachter, R.M. (2010), "Accountability Measures – Using Measurement to Promote Quality Improvement," New England Journal of Medicine, 363(7): 683-688.

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.

Corcoran, S. (2010) "Can Teachers be Evaluated by their Students' Test Scores? Should They Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice," Working Paper, Annenberg Institute for School Reform.

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

Dranove, D., Kessler, D., McClellan, M. and Satterthwaite, M. (2003) "Is More Information Better? The Effects of 'Report Cards' on Health Care Providers," Journal of Political Economy, 111(3): 555-588.

Farber, H.S. and Gibbons, R. (1996) "Learning and Wage Dynamics," Quarterly Journal of Economics, 111(4): pp. 1007-1047.

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

Gibbons R. and Katz, L. (1991) "Layoffs and Lemons," Journal of Labor Economics 9(4): 351-380.

Goldhaber, D. and Hansen, M. (2009) "Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions," University of Washington Center on Reinventing Public Education Working Paper 2009_2.

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. (2006) "Value-added Models and the Measurement of Teacher Quality," University of Florida Working Paper.

Harris, D.N. and Sass, T.R. (2009) "What Makes for a Good Teacher and Who Can Tell?" Calder Center Working Paper #30.

Hastings, J.S. and Weinstein J.M. (2008) "Information, School Choice, and Academic Achievement: Evidence from Two Experiments," Quarterly Journal of Economics, 123(4): 1373-1414.

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.

Hibbard, J.H., Stockard, J., and Tusler, M. (2005) "Hospital Performance Reports: Impact on Quality, Market Share, and Reputation," Health Affairs, 24(4): 1150-1160.

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.

Jin, G.Z. and Leslie, P. (2003) “The Effect Of Information On Product Quality: Evidence From Restaurant Hygiene Grade Cards,” Quarterly Journal of Economics, 118(2): 409-451.

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.

Koedel, C. and Betts, J.R. (2009) “Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique,” Education Finance and Policy 6(1): 18-42.

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.

Macleod, W.B. (2003) “Optimal Contracting with Subjective Evaluation,” American Economic Review, 93(1): 216-240.

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. (1986) “Incentives, Learning, and Compensation: A Theoretical and Empirical Investigation of Managerial Labor Contracts,” RAND Journal of Economics 17(1): pp. 59-76.

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.

New York Times (2008), “New York Measuring Teachers by Test Scores,” by Jennifer Medina, January 21, 2008, page A1.

Oyer, P. and Schaeffer, S. (2011) “Personnel Economics: Hiring and Incentives” in Orley Ashenfelter and David Card (Eds.), Handbook of Labor Economics, Great Britain, North Holland, pp. 1769-1823.

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. (2010) “Short Run Impacts of Accountability on School Quality,” American Economic Journal: Economic Policy, 2(4): 119-147.

Rothstein, J. (2009) “Student Sorting and Bias in Value-added Estimation: Selection on Observables and Unobservables,” Education Finance and Policy, 4(4): 537-571.

Spence, M. (1973) “Job Market Signaling,” Quarterly Journal of Economics, 87(3): 355-374.

Springer, M.G., Ballou, D., Hamilton, L. Le, V., Lockwood, J.R., McCaffrey, D.F., Pepper, M. and Stecher, B.M. (2010) “Teacher Pay for Performance: Experimental Evidence From the Project on Incentives in Teaching,” Working Paper, National Center on Performance Incentives.

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: Comparability Between Sample and Popululation and Between Treatment and Control at Baseline and Follow-up

	Population vs. Sample			*Treatment vs. Control (Baseline)*			*Treatment vs. Control (Follow-up)*		
	Population Mean	Sample Mean	P-value on Difference	Control Mean	Treatment Mean	P-value on Difference	Control Mean	Treatment Mean	P-value on Difference
Number of Principals/Schools	1092	223		112	111		84	94	
Total Enrollment	660	712	0.03	705	717.8	0.79	715	724.9	0.85
Principal Characteristics (Spring 2007)									
Years of Experience as Principal (in School)	3.6	3.3	0.27	3.3	3.2	0.81	3.2	3.2	0.98
Years of Experience as Assistant Principal	2.7	2.5	0.41	2.4	2.7	0.40	2.4	2.8	0.43
Years of Experience as Teacher	6.2	7.2	0.01	6.8	7.6	0.36	6.6	7.8	0.18
Years of Experience in School (Any Position)	5.2	4.7	0.17	4.4	5.1	0.31	4.4	4.8	0.53
Principal Age	48.8	48.4	0.41	48.0	48.8	0.50	48.2	49.9	0.16
Principal is Black or Hispanic	47.7%	45.3%	0.47	48.6%	41.9%	0.32	47.9%	44.1%	0.61
Principal is Female	73.4%	79.4%	0.05	81.1%	77.7%	0.53	80.9%	77.4%	0.57
Student Characteristics (Grades 4-8, 2007)									
On Free Lunch	84.9%	86.1%	0.35	87%	85.4%	0.64	85.7%	85.0%	0.81
English Language Learners	11.5%	14.0%	0.00	15%	13.3%	0.30	14.7%	13.0%	0.30
In Special Education	13.8%	9.6%	0.00	10%	9.8%	0.79	9.2%	10.2%	0.40
Black or Hispanic	75.6%	72.5%	0.11	72%	72.8%	0.91	71.9%	72.9%	0.83
Teacher Survey Results (Spring 2007)									
The Principal…									
Visits Classrooms to Observe the Quality of Teaching	0.00	0.03	0.65	0.071	-0.03	0.44	0.091	-0.026	0.44
Gives Me Regular and Helpful Feedback	0.00	-0.05	0.47	-0.047	-0.069	0.86	-0.026	-0.101	0.59
Places a High Priority on the Quality of Teaching	0.00	-0.02	0.72	-0.027	0.004	0.80	-0.031	0.034	0.64
Teachers in this School…									
Use Student Data to Improve Instructional Decisions	0.00	0.00	0.99	0.096	0.077	0.87	0.093	0.058	0.80
Receive Training in the Use of Student Data	0.00	-0.01	0.90	0.036	0.022	0.90	0.049	0.012	0.79

Note: P-values indicate the statistical significance of the difference between the two groups of principals/schools. Teacher survey variables have been normalized using schools city-wide to have mean zero and standard deviation one. One control and three treatment schools are missing survey data.

Table 2: Summary Statistics on Teacher Level Variables

	Control	Treatment	P-value on Difference
In Baseline Survey with Value Added Estimate	1184	1324	
Principal's Rating (Scale from 1 to 6)			
Overall	4.32	4.31	0.90
	(1.11)	(1.12)	
Math Instruction	4.21	4.23	0.87
	(1.09)	(1.13)	
ELA Instruction	4.21	4.19	0.89
	(1.04)	(1.13)	
Observations Made by Principal Last Year			
Formal	2.21	2.24	0.84
	(1.26)	(1.25)	
Total	6.51	6.26	0.59
	(3.38)	(3.21)	
Value-added Estimates			
Math, Multi-year, Citywide	0.002	0.002	0.96
	(0.166)	(0.178)	
Math, Single-year, Citywide	0.006	0.014	0.57
	(0.177)	(0.188)	
ELA, Multi-year, Citywide	-0.014	-0.002	0.24
	(0.135)	(0.125)	
ELA, Single-year, Citywide	-0.007	0.011	0.13
	(0.144)	(0.141)	
Teacher Experience at Baseline			
None (First Year was 2006-2007)	9.0%	10.8%	0.28
One Year	11.7%	10.8%	0.62
Two Years	10.9%	10.6%	0.85
Three Years	8.6%	10.9%	0.08
Four Years	7.4%	6.8%	0.62
Five to Nine Years	27.9%	26.8%	0.61
Ten or More Years	24.6%	23.2%	0.57
Turnover			
Exited DOE Prior to Fall 2007	9.0%	11.9%	0.07
Exited School Prior to Fall 2007	12.8%	16.4%	0.06
Exited DOE Between Fall 2007 and 2008	6.9%	8.2%	0.32
Exited School Between Fall 2007 and 2008	12.4%	13.7%	0.45

Note: P-values indicate the statistical significance of the difference between the two groups of teachers. 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 3: 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 4: 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 5: 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 6: 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 7: 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 8: Teacher Observation by Principals at Baseline

	Formal Observations		Total Observations	
	(1)	(2)	(4)	(5)
Years Principal Observes Teacher	-0.272**	-0.090**	-0.221**	-0.076+
	(0.027)	(0.027)	(0.038)	(0.039)
Overall Performance Evaluation	-0.116**	-0.107**	-0.228**	-0.222**
	(0.022)	(0.022)	(0.040)	(0.041)
Teacher Experience (No Experience Omitted)				
1 Years Experience		-0.045		-0.268+
		(0.082)		(0.137)
2 Years Experience		-0.150+		-0.314*
		(0.088)		(0.128)
3 Years Experience		-0.353**		-0.413**
		(0.113)		(0.154)
4 Years Experience		-0.649**		-0.642**
		(0.108)		(0.148)
5-9 Years Experience		-0.806**		-0.670**
		(0.100)		(0.139)
10+ Years Experience		-0.884**		-0.815**
		(0.102)		(0.148)
R-squared	0.71	0.77	0.89	0.89
Sample Size	2,487	2,487	2,487	2,487

Table 9: 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 10: 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 11: Impact of Value-added on Propensity to Exit the School, by Subject

	Teachers of Math			Teachers of English		
	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.031*	0.006	*-0.037*			
	(0.013)	(0.012)	*[p=0.037]*			
Value-added in English				0.000	0.003	*-0.003*
				(0.014)	(0.013)	*[p=0.875]*
Overall Evaluation, Pre-experiment	-0.020	-0.041**	*0.021*	-0.013	-0.052**	*0.039*
	(0.014)	(0.015)	*[p=0.306]*	(0.015)	(0.016)	*[p=0.076]*
Sample Size	936	844		872	819	
	Teachers of Math and English					
	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.036*	0.003	*-0.039*			
	(0.016)	(0.015)	*[p=0.076]*			
Value-added in English				-0.000	-0.011	*0.011*
				(0.028)	(0.015)	*[p=0.729]*
Overall Evaluation, Pre-experiment	-0.008	-0.037*	*0.029*	-0.017	-0.035*	*0.018*
	(0.015)	(0.017)	*[p=0.201]*	(0.016)	(0.017)	*[p=0.441]*
Sample Size	705	631		705	631	
	Teachers of Math and English					
	(5)			(6)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added in Math	-0.043*	0.010	*-0.053*	-0.042*	-0.006	*-0.036*
	(0.017)	(0.014)	*[p=0.016]*	(0.018)	(0.02)	*[p=0.181]*
Value-added in English	0.016	-0.015	*0.031*	0.015	-0.001	*0.016*
	(0.028)	(0.015)	*[p=0.329]*	(0.028)	(0.019)	*[p=0.636]*
Overall Evaluation, Pre-experiment	-0.011	-0.036*	*0.025*	-0.007	-0.031	*0.024*
	(0.015)	(0.017)	*[p=0.27]*	(0.016)	(0.02)	*[p=0.349]*
Teacher Experience and School FE				√	√	
Sample Size	705	631		705	631	

Note: All specifications are linear regressions that include teacher experience and school fixed effects. 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. Significance of coefficients are denoted as follows: **p < 0.01, *p<0.05, +p<0.1; p-values on tests of differences between treatment and control are shown in brackets.

Table 12: Impacts on 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)
	[p=0.16]	[p=0.17]	[p=0.25]	[p=0.10]	[p=0.04]	[p=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)
	[p=0.45]	[p=0.37]	[p=0.47]	[p=0.76]	[p=0.35]	[p=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.

Appendix Figure 1: Sample Value Added Report

Page 1

NYC Department of Education

Value-added Data for Teachers Initiative

Teacher:	**Swain, Winthrop**	**Years in Current Grade/Subject:**	**3**
Grade:	**5.0th Grade**	**Experience Category:**	**10+Yrs**
School:	**PS 006 Lillie D. Blake**	**Classroom Quintile:**	**Fourth**
Year:	**2006-2007**		

Teacher Performance

The Difference-from-Predicted gain in the average student proficiency level for this teacher is the difference between the average actual gain of all the teacher's students and the average predicted gain for students with similar characteristics.

Teacher Compared to Citywide Teacher Performance Horizon - All schools, all teachers; same grade

	Sample Size	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)	Citywide Horizon Teacher value-added relative to range of results for all in same grade in the City
ELA - This year (lower / upper bound)	40	.22	.07	.15* (.02,.27)	76.4% (56%,96.9%) -0.31 .15* 0.29
ELA - History: up to 3 years (lower / upper bound)	144	.11	.04	.07* (.00,.14)	67.4% (54.9%,79.9%) -0.30 .07* 0.25
Math - This year (lower / upper bound)	43	.40	.17	.23* (.09,.37)	78.8% (62.1%,95.5%) -0.45 .23* 0.41
Math - History: up to 3 years (lower / upper bound)	152	-.03	-.09	.06 (-.01,.13)	62.4% (52.6%,72.2%) -0.40 .06 0.34

Teacher Compared to Peer Teacher Performance Horizon – Similar experience, similar classrooms; same grade

	Sample Size	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)	Peer Teacher Horizon Teacher value-added relative to range of results for all teachers in the same grade with similar experience and similar classrooms
ELA - This year (lower / upper bound)	40	.22	.14	.07 (-.04,.19)	69.8% (43.3%,96.2%) -0.23 .07 0.21
ELA - History: up to 3 years (lower / upper bound)	144	.11	.09	.03 (-.04,.09)	60.1% (45.3%,75%) -0.25 .03 0.21
Math - This year (lower / upper bound)	43	.40	.22	.18* (.05,.31)	75.4% (59.7%,91.1%) -0.47 .18* 0.39
Math - History: up to 3 years (lower / upper bound)	152	-.03	-.04	.01 (-.06,.08)	54.4% (41.3%,67.6%) -0.28 .01 0.25

Note: The lower and upper bound means that there is a very high probability (95%) that the teacher's actual contribution to student gains in proficiency falls within this interval. The (*) means that there is a very high probability that the contribution is positive (or negative). All comparisons are among teachers in the same grade.

Appendix Figure 1: Sample Value Added Report

Page 2

NYC Department of Education

Value-added Data for Teachers Initiative

Teacher: **Swain, Winthrop**

Teacher Performance by Student Characteristics

Teacher's value-added for sub-groups of students compared to teacher's value-added overall for history: up to 3 years

Types of Student	Sample Size / (% of Sample)	Actual Gain	Predicted Gain	Difference from Predicted (Teacher's Value Added)
English Language Arts				
All Students	144 (100%)	0.11	0.04	0.07*
Citywide:				
Bottom Third	94 (62.8%)	0.27	0.16	0.10*
Middle Third	39 (29.3%)	-0.13	-0.14	0.01
Top Third	11 (7.9%)	-0.32	-0.37	0.04
School				
Bottom Third	51 (32.5%)	0.39	0.24	0.16*
ELL	-	-	-	-
Special Education	15 (10.1%)	0.19	0.02	0.17
Mathematics				
All Students	152 (100%)	-0.03	-0.09	0.06
Citywide:				
Bottom Third	106 (64.2%)	0.11	0.01	0.10*
Middle Third	37 (28.4%)	-0.33	-0.30	-0.03
Top Third	9 (7.4%)	-0.46	-0.45	-0.02
School				
Bottom Third	48 (25.2%)	0.24	0.14	0.11
ELL	10 (6.8%)	-0.14	0.01	-0.15
Special Education	15 (9.1%)	-0.01	-0.11	0.11

The (*) means that there is a very high probability that the contribution is positive (or negative).

Teacher Percentile

The percent of teachers in the comparison group whose value added falls below this teacher

	English Language Arts		Mathematics	
Comparison Teachers	This Year	History: up to 3 years	This Year	History: up to 3 years
All teachers, all schools	88	77	90	69
Teachers with similar experience, similar classrooms	86	71	86	56

Note: All comparisons are among teachers in the same grade

Basic Student Progress

The percent of students in the teacher's classroom making at least the predicted gain

	English Language Arts		Mathematics	
	This Year	History: up to 3 years	This Year	History: up to 3 years
This Teacher	70.0%	60.4%	72.1%	53.3%
All teachers, all schools	48.4%	47.3%	49.9%	47.3%
Teachers with similar experience, similar classrooms	56.7%	51.8%	48.5%	54.3%

Note: All comparisons are among teachers in the same grade

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.

Appendix Table A4: Value-added and Propensity to Exit Pre-experiment (Placebo Test)

Panel A: OLS	(1)			(2)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	-0.003	-0.009	*0.006*	0.015	0.005	*0.01*
	(0.014)	(0.011)	*[p=0.736]*	(0.013)	(0.01)	*[p=0.542]*
Overall Evaluation, Pre-experiment				-0.074**	-0.076**	*0.002*
				(0.014)	(0.014)	*[p=0.92]*
R-squared	0.00	0.00		0.04	0.05	
Sample Size	1,323	1,184		1,323	1,184	
Panel B: Logit	(3)			(4)		
	Treatment	Control	*Difference*	Treatment	Control	*Difference*
Value-added	-0.025	-0.086	*0.061*	0.106	0.049	*0.057*
	(0.099)	(0.101)	*[p=0.666]*	(0.091)	(0.092)	*[p=0.66]*
Overall Evaluation, Pre-experiment				-0.517**	-0.646**	*0.129*
				(0.099)	(0.103)	*[p=0.367]*
Sample Size	1,323	1,184		1,323	1,184	

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.