Preliminary Draft, Please do not cite or circulate without authors' permission

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
Harvard Graduate School of Education

May, 2009*

Abstract

A large body of research suggests that improving teacher quality is an important channel for raising student achievement. However, in order to selectively retain highly effective teachers, or provide training to teachers who need it, school administrators must be able to accurately evaluate teacher performance. One potential tool to improve teacher evaluation is the use of information based on student achievement outcomes. To assess the merits of providing such information to school administrators, the New York City Department of Education disseminated reports on the performance of individual teachers to a randomly selected subset of volunteer principals in the school year 2007-2008. Several important facts emerge from our analysis of this pilot program. First, the estimates of teacher performance reported to principals were correlated with principals' preexisting beliefs about teacher effectiveness, as reflected in both their opinions and their actions (e.g., classroom observations). Second, principals changed their subjective ratings of teachers in response to the information in the reports, establishing that value-added reports provided new information that principals felt was useful. Finally, our results suggest that the information in the reports raised the probability that teachers with low performance estimates left their school at the end of the pilot year. Collectively, these results suggest that information on how individual teachers impact student achievement may help principals raise teacher quality in their schools.

* Corresponding author: jonah.rockoff@columbia.edu. 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.

Introduction

There is a large body of evidence that teachers vary greatly in their ability to raise student achievement (e.g., Rockoff (2004), Rivkin et al. (2005), Aaronson et al. (2007)). If true, then policies that accurately measure teacher effectiveness and selectively retain effective teachers (or help less effective teachers improve) can lead to higher student achievement. However, we are only beginning to understand how the individuals entrusted with teacher personnel decisions (namely, principals) construct and use measures of teacher effectiveness.

Today, the most common method for evaluating teacher effectiveness is a school principal's subjective evaluation, based largely on classroom observations. While principals' evaluations may distinguish teachers with very high or very low effectiveness (Jacob and Lefgren (2008), Harris and Sass (2007)), the accuracy of such evaluations may be limited by various factors such as limited data, difficulty in gauging how classroom composition may affect the learning environment, limited understanding of what constitutes effective teaching generally or for a specific subject, or personal biases. The extent to which school principals act upon their own subjective evaluations is also unclear. Principals may be reluctant to dismiss a teacher due to institutional or social costs (see Jacob (2007)), misalignment of incentives, or uncertainty about the accuracy of their own judgments.

A body of recent research has focused on using student test based estimates of teacher "value-added" to identify teachers who are most effective at raising student achievement. Although there is still considerable discussion surrounding the validity of the assumptions underlying these estimates (Todd and Wolpin (2003), Rothstein (2009)), two recent papers (Gordon et al. (2006), Kane and Staiger (2008)) make a strong case for using value-added estimates as an input into teacher related policy decisions. One option for using these estimates

is to provide them to principals, and allow principals to incorporate this information into their decision-making process. Assuming the estimates provide new and useful information, principals might disseminate the practices of successful teachers or work to improve or replace teachers who perform poorly.

Whether such a policy would have any effect at all is unclear. For example, principals' subjective evaluations of teacher effectiveness may already account for all of the variation in value-added estimates that is correlated with future teacher performance. Principals might also place little faith in the accuracy of value-added estimates and therefore give them little or no weight when making teacher personnel decisions. Empirical research on this issue is quite limited. An increasing number of school districts estimate teacher value-added (e.g., through programs like EVASS) and a few (e.g., Seattle, Dallas, Denver, and the state of Tennessee) have started to use value-added estimates for teacher policy decisions such as bonus payments. However, to the best of our knowledge, there is no empirical research on the impact of providing these data to principals.

Given these outstanding questions, the New York City Department of Education (hereafter the DOE) conducted a pilot program during the school year 2007-2008 to assess the impact of providing principals with measures of teacher effectiveness. A group of principals, randomly selected from among a larger group of volunteers, received "value-added reports" on individual teachers at their school. Each report contained several measures of teacher effectiveness based on regression analyses of students' standardized test score performance. Treatment principals were given training in the methodology that created these measures and how to interpret the reports.

To evaluate the pilot program, we first confirm that the randomization created groups of treatment and control principals (schools) that were quite similar. The successful randomization allows us to draw causal inferences regarding the impact of providing the value-added reports to principals. We also present evidence that, consistent with earlier research, there is a statistically significant positive relationship between the value-added measures reported to principals and the principals' pre-existing beliefs regarding teacher effectiveness.

We then look for differences between treatment and control groups on a set of outcomes measured after the year in which the reports were distributed. We present evidence that treated principals' beliefs regarding teacher effectiveness were significantly affected by the value-added reports they received, particularly for math instruction. We also find some evidence of an increase in the probability that teachers receiving low value-added reports exited their school. These findings represent a substantial contribution to understanding how principals incorporate information on teacher quality into personnel decisions, and suggest that the provision of value-added estimates to principals may be a useful tool for the improvement of teacher quality.

The paper proceeds as follows. We describe the pilot program in Section 2, and in Section 3 we provide comparisons of treatment and control groups at baseline, and discuss sample attrition. In Section 4, we present results on the relationship between teachers' estimated value-added and principals' pre-existing beliefs regarding teacher performance. In Section 5, we present results on the impact of the treatment, and Section 6 concludes.

2. The Value-Added Data Initiative

In the fall of 2006, the DOE began development of a new initiative that would provide principals with data to help them identify high and low performing teachers. Its basic elements

were the development of the internal capacity to estimate teacher value-added, design and dissemination of reports on teacher value-added to principals, and training of principals to understand and interpret the reports. In order to understand the likely impact of this initiative on principals, teachers, and students, the DOE decided to pilot the initiative with a subset of schools, using randomized treatment and controls groups.

In the early summer of 2007, the DOE invited principals to participate in the pilot program. Invitations were limited to principals whose schools contained any of the grades 4-8, who themselves had been in the school for all of the school year 2006-2007, who were not known to be leaving for the school year 2007-2008, and whose schools were not among a subset of middle schools with known data problems in the linkage of teachers to students.[1] These principals received information about the pilot via e-mail, a link to a web site with additional information, and an invitation to attend presentations related to research on value-added and the DOE initiative. From this group of about 1,000 principals, 335 initially expressed interest in becoming part of the program. These principals were sent a baseline survey on August 8, 2007 and were told that completion of the survey by September 21st was required to be eligible to participate in the pilot. They were also informed that only a randomly selected subset of eligible principals would be provided teacher reports in order to study the impacts of the program. 223 principals completed the survey, and half of these principals (112) were selected into the treatment group. The randomization was accomplished by generating a random number for each school, sorting by number within Elementary, Middle, and K-8 schools, and selecting the first 73, 27, and 12 cases within each group, respectively. Neither the identities of the schools who were eligible to receive the reports nor those that were selected to receive them were ever made

[1] Grade level limitations were based on the fact that students in New York are tested annually in math and English Language Arts in grades 3 through 8, and the methodology used by the DOE to estimate value-added relies on past test scores as a control variable.

public. However, principals themselves may have shared this information and were never advised to keep it private.

The baseline survey was divided into two parts.[2] First, principals were given a list of teachers who, based on DOE records, taught math and/or English Language Arts (ELA) to students in any of grades four to eight, i.e., teachers for whom value-added estimates were possible. Principals were asked to confirm that each teacher had indeed taught in these areas (98.2 percent of teachers were confirmed). Principals then evaluated each teacher using a six point scale; first evaluating the teacher "overall," and then specifically "in terms of raising student achievement" in mathematics or ELA (or both for teachers who taught both subjects).[3] In all evaluations, principals were asked to compare the teacher to all "teachers you have known who taught the same grade/subject," not just the teachers within their school or teachers with similar levels of experience. In addition to their rating(s) of the teacher, principals were asked to provide the number of formal classroom observations and total classroom observations they made of the teacher during the past school year.

The second part of the baseline survey asked principals about a number of topics related to how they measured teacher effectiveness in their schools and their use of student test score data in that process. For example, they were asked about devices they use to assess teachers other than classroom observation, their beliefs regarding potential benefits of (and concerns with) measuring teacher performance using student test scores, and their level of satisfaction with their ability to attract and retain high quality teachers in their schools. Summary statistics

[2] A complete copy of the baseline survey is included in the Appendix. The baseline and follow-up surveys were conducted by the Battelle Memorial Institute. Battelle also performed the estimation of teacher value-added, aided in the provision of professional development to principals, and prepared the value-added reports.

[3] Specifically, principals were asked to rate the teachers as "Exceptional (top 5%)" "Very Good (76-95th percentile)" "Good (51-75th)" "Fair "(26-50th)" "Poor (6-25th percentile)" or "Very Poor (bottom 5%)".

on principals' responses to the second part of the baseline survey are provided in Appendix Table A1.

Treatment principals were invited to participate in professional development sessions, held in early December 2007, to increase their understanding of the DOE's value-added methodology and the data reports. The sessions lasted three hours, with two hours for an explanation of the statistical concepts behind value-added modeling, a walk-through of a sample value-added report, and discussion of uses of the information in schools. Principals then received their teachers' reports, and the remaining hour was devoted to answering principals' questions. Of the treatment principals, 71 attended a session in person, while 24 participated in an online session (similar to a conference call, but with a presentation viewed via computer), and 1 viewed a videotape of a session.[4] The DOE did not distribute value-added reports to the 16 treatment principals who did not receive any professional development.[5] Principals receiving reports were invited to a follow-up professional development session in April, and approximately 30 attended.

Appendix Figure 1 shows a sample value-added data report, which contains a number of different measures of value-added.[6] Four different measures of value-added are reported for each teacher in each subject. The first two value-added scores (see the top table on page 1) compare the teacher to all teachers in the city teaching the same grade-level. The second two (see the bottom table on page 1) compare the teacher only to other teachers within the city who

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

[5] In our analysis below, we include all treatment principals whenever possible and estimate "intent to treat" effects of the pilot program. Thus, readers interested in the "treatment on treated" effect should scale up our estimates by roughly 15 percent.

[6] We do not discuss the value-added estimation methodology here. The details can be found in a technical report put out as part of the value-added data initiative (Battelle, 2009). Essentially, the methodology uses linear regression to predict student test scores based on prior information (e.g., past test scores), and generates value-added estimates by applying an empirical Bayes estimator to the residuals from these regressions.

have similar levels of teaching experience and who work in classrooms with similar student composition. For each comparison group the teacher's value-added is measured two ways: based on up to three years of prior data, and based only on the prior year. A 95 percent confidence interval is reported for all estimates based on the estimated variance of the value-added measure.[7] In order to be more familiar to principals, the value-added measures were reported in "proficiency rating units," a scale based on the state examinations and used by the DOE in its school accountability system. Since each value-added measure compared teachers only to others teaching the same grade-level, multiple reports were generated for teachers of multiple grades; our data indicate that multiple reports were distributed for 25 percent of middle school teachers and 10 percent of elementary school teachers.

In late May 2008, a follow-up survey was sent to principals to be completed by mid-July.[8] Two treatment principals and one control principal had sent word to the DOE that they no longer wanted to be involved with the study and were not sent the follow-up survey. However, 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.

The follow-up survey first asked principals to evaluate individual teachers in the same manner as in August 2007, allowing one to measure if the principal's opinions changed over the intervening period. If a principal did rate a teacher differently than they did at baseline, they were asked to provide a reason for the change. In the second part of the follow-up survey, both treatment and control groups were asked a set of questions about the importance of various

[7] The report also shows where the teacher sits in the distribution of value-added estimates and transforms the value-added estimates into percentiles. Additionally, the report presents value-added measures specific to student subgroups (e.g., English Language Learners, Special Education students, students who scored in the bottom third of the school's distribution in the prior year). In our analysis below, we restrict our attention to the value-added estimates estimating using all types of students.

[8] A complete copy is included in the appendix.

issues when using students' standardized test scores to assess individual teachers; treatment principals were then asked about their confidence that the value-added calculations addressed these issues. Treatment principals were also asked a number of questions about the reports they received, including their opinions on the usefulness of the reports for various purposes and whether they shared the reports with administrators and/or teachers in their school.

Summary statistics on principals' responses to the second part of the follow-up survey are provided in Appendix Tables A2 and A3, the latter focusing on questions asked only of the treatment principals. There are a few facts in these tables worth noting here. First, when asked to report the top four factors (other than observation) influencing their evaluation of teachers, principals included statewide standardized tests (the same tests used to create the value-added estimates) more than any other source of information; these tests were cited by 96 percent of control principals and 93 percent of treatment principals. Thus, principals who volunteered to receive the value-added reports already seem to be using standardized test outcomes as a means of evaluating teacher effectiveness. Second, principals in both treatment and control schools reported using many different test based measures, including average test scores, average test score growth, and the percentage of students meeting proficiency standards.

Principals in both groups mainly agreed on the importance of considering various factors when using student test scores to assess individual teachers, with three exceptions. Treatment principals placed more importance on considering the number of students entering a class mid-year and the teacher the student had in the prior year. Both of these variables were controlled for in the DOE's value-added methodology and were discussed in professional development sessions which may have sensitized treatment principals to these issues. Treatment principals also put less importance on controlling for students' prior test scores, and the difference with

control schools is marginally significant. It is less clear to us what is driving this difference, since treatment principals placed greater importance on controlling for every other factor about which they were asked, and the DOE value-added methodology does control for students' prior test scores. Finally, treatment principals' confidence in how the DOE value-added methodology accounted for various factors largely accords with the details of the methodology. For example, principals were highly confident that the methodology accounted for factors such as teaching experience, prior test scores, and class size (all of which are control variables in the value-added estimation) and expressed little confidence that the methodology accounted for factors such as whether a teacher had a classroom aide or whether a teacher's students received outside help (neither of which were control variables).[9]

Before proceeding with our analysis, it is important to mention a number of external factors that may have affected principals' use of the value-added reports. In June of 2007, the DOE met with officials of the teachers' union (the United Federation of Teachers) to discuss the pilot. The union did not support the idea, and it filed a formal grievance on the matter in October of 2007. The DOE, partly in response to the union's opposition, advised treatment principals when they received the reports that they were *not* to be used for teacher evaluation (e.g., tenure decisions) during the pilot year.[10] Also, though the identities of participating principals were held confidential, the existence of the pilot was made public and widely reported in the popular press, including making the front page of the New York Times on January 21, 2008. In April 2008, with considerable support from teachers' unions, the New York State legislature amended

[9] 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 such as whether a teacher had a personal issue or whether students were distracted on the day of the test by construction noise.
[10] They were also advised not to share the pilot data with parents.

the law governing teacher tenure, adding the statement that teachers "shall not be granted or denied tenure based on student performance data."

3. Comparison of Treatment and Control Groups

In order to ensure that the randomization was successful, we first compare the average characteristics of treatment and control principals (schools) at baseline, and test for any statistically significant differences between the two groups (Table 1, left side). We find the treatment and control groups are very similar with regard to principals' characteristics (work experience and demographics), students' characteristics (poverty, demographics, and participation in special programs), and responses by their teachers to a series of questions on a citywide survey from the spring of 2007. These results indicate that the randomization was successful in producing a high degree of similarity between the two groups at baseline.[11]

Between the baseline and follow-up surveys, attrition of principals was somewhat higher in the treatment group (see Table 2, Panel A). Of the 110 treatment principals invited to take the follow-up survey, 84 began survey and 79 completed it, while, of the 110 control principals invited, 94 began the survey and 93 completed it.[12] The difference between the groups seems to be driven in part by the principals who did not receive professional development or value-added reports; only 5 of these 16 principals completed the follow-up survey.

Likewise, teacher attrition was also higher for the treatment group. Among teachers with a value-added estimate and a principal rating in the baseline survey, roughly 56 percent in the treatment group were evaluated by the same principal in the follow-up survey, relative to 66

[11] Note that, in addition to these characteristics, we can compare treatment and control principals based on their responses to the baseline survey. Average responses for treatment and control principals were only statistically different at the ten percent level on just one item: 68 percent of treatment principals said that reports on student test score growth would be very useful for assignment of students to teachers, versus 56 percent for control principals.
[12] Nearly all principals who began the follow-up survey completed the first section in which they rated teachers.

percent in the control group. This difference was partly due to greater attrition of treatment principals, but also due to the fact that more teachers in treatment schools left their baseline school before the start of the school year 2007-2008 (18 percent vs. 13 percent). It is highly unlikely that this difference in turnover could have been due to the treatment, since principals were not placed into treatment and control groups until after the start of the school year and did not receive professional development and value-added reports until several months later.

Nevertheless, greater attrition in the treatment schools raises a concern that the characteristics of principals and teachers for whom follow-up survey data is available could be different in the treatment and control groups. We investigate whether the treatment and control groups were similar after attrition by limiting our sample to principals responding to the follow-up survey and analyzing the same characteristics as we did for the baseline sample (Table 1, right side). As at baseline, there were no significant differences between the two groups, and on some measures the two groups converge. Although we cannot test for differences on dimensions other than those we can observe in our data, these results support the notion that the treatment and control principals who responded to the follow-up survey were comparable. Of course, any outcome measured outside of the follow-up survey (e.g., teacher turnover) can be analyzed without limiting our sample to survey respondents.

4. Value-added and Principals' Responses in the Baseline Survey

Before examining any impact of the value-added reports, it is useful to first establish how these estimates relate to principals' pre-existing beliefs regarding the effectiveness of their teachers. Table 3 (left side) presents summary statistics on teachers at baseline. On a 0-5 scale, the mean principal rating was roughly 3.2 in both math and ELA, with a standard deviation of

about 1.1 for both subjects. On average, teachers received 2.2 formal observations and 6.4 total observations in the prior year. Consistent with idea that classroom observation is the mainstay of teacher evaluation, only 3 percent of teachers in the baseline sample were not formally observed by their principal in the prior year, and only 0.4 percent of teachers were not observed either formally or informally.

Average teacher value-added estimates are quite close to zero at baseline for both groups, suggesting that these schools collectively did not have significantly more or less effective teachers than those which did not participate. The variation in value-added estimates is, not surprisingly, higher for estimates using only one year of data and higher for estimates that do not adjust for years of teaching experience. Variation is also smaller for ELA than math, consistent with estimates from other studies in New York and elsewhere (e.g., Kane et al. (2008), Kane and Staiger (2008), Rivkin et al. (2005)). Roughly half of the teachers had at least five years of experience, while about one third had less than three years of experience (our proxy for being untenured).

To examine the relationship between value-added estimates reported to principals and baseline ratings, we use a simple linear regression specification, shown by Equation 1.

$$(1)\quad R_{ijt} = \alpha + \beta VA_{ijt} + \delta X_{it} + \varepsilon_{ijt}$$

The rating given to teacher i for subject j at time t (R_{ijt}) is specified as a function of the value-added estimate in subject j at time t (VA_{ijt}) and other teacher characteristics (X_{it}). We begin our analysis by omitting teacher and principal characteristics from the covariates and then test to see how the addition of these controls influences our findings.

The results of these regressions are quite similar for English Language Arts (Table 4a) and math (Table 4b). In the first two columns of each table we show that principals' pre-

experimental ratings were significantly higher for teachers who had high value-added estimates, both for the multi-year estimates and the single year estimates that compare teachers to their peers (i.e., those with similar experience and similar classrooms of students). In the third column, we test the relative strength of these two value-added estimates to predict principals' ratings, and find that the multi-year estimates dominate estimates based only on only the past year of student performance, particularly for math. Thus, principals' opinions are based more on teachers' long term average performance than their most recent results. In Columns 4 and 5, we show that the multi-year estimates for value-added using the peer comparison are also stronger predictors of ratings than those based on citywide comparison of teachers, though the latter are still statistically significant when both are included in the regression. For simplicity, we present results that use the multi-year peer comparison estimates in the remainder of our analysis, though these results are similar if we instead use the citywide estimates.

In Columns 6 and 7, we include more control variables, first adding indicators for various levels of teaching experience and then adding principal fixed effects. The initial results are quite robust to these alternate specifications. It is worth noting that, conditional on value-added, principals pre-experimental ratings tend to be lowest for teachers who just completed their first year, and highest for teachers with three to nine years of experience, while teachers with only a few years of experience or ten or more years of experience tend to be rated in the middle.

In Column 8 we add an interaction between the value-added estimate and whether the teacher has more than three years of experience, an imperfect proxy for whether a teacher has tenure.[13] If either teacher value-added estimates or principal ratings become more precise over time (as both estimation processes are based on more data), one might hypothesize that the predictive power of the value-added measures will be higher for more experienced teachers. We

[13] Teachers receive tenure if they continue to teach in New York after serving for three years.

find that this is indeed the case for both ELA and math. The value-added estimate is a significant predictor of the principal's evaluation at the start of teachers' careers (the first coefficient in Column 8), but it is even stronger for veteran teachers (the sum of the first and second coefficients in Column 8).

For teachers who teach both math and ELA, we asked the principal to evaluate their effectiveness in both subjects separately. These separate ratings allow us to examine the principals' evaluation of differential effectiveness by subject area by asking whether teachers with higher value-added in a particular subject are more likely to be rated higher in that subject by their principal. To do so, we estimate the following regression specifications:

$$\text{(2a)}\quad R_{ijt} = \alpha + \pi R_{ikt} + \beta VA_{ijt} + \varepsilon_{ijt}$$
$$\text{(2b)}\quad R_{ijt} = \alpha + \pi R_{ikt} + \beta VA_{ikt} + \varepsilon_{ijt}$$

The notation here is the same as in Equation 1, except the *k* subscript denotes ELA when *j* denotes math, and vice versa. If principals' subject-specific ratings are related to subject specific value-added, then value-added measures for math should have more power to predict ratings in math than ratings in ELA, and vice versa. To keep our sample the same in all regressions, we restrict our attention to teachers with value-added estimates and ratings in both subjects. For simplicity, we do not include additional controls, but our results are not sensitive to the inclusion of teacher experience controls or principal fixed effects.

For both math and ELA, we find that, conditional on the principal's rating in the other subject, principals' ratings are significantly related to value-added in the same subject, but not to value-added in the other subject (Table 5). In other words, for any two teachers with the same math rating, the one with a higher ELA rating tended to have a higher ELA value-added score, but not a higher math value-added score. Likewise, for any two teachers with the same ELA rating, the one with a higher math rating tended to have a higher math value-added score, but not

a higher ELA value-added score. However, it is also important to point out that there is a high correlation across subjects for both principals' ratings (0.91) and value-added estimates (0.44). Thus, among teachers of both math and ELA, those rated as highly effective in one subject also tend to be rated highly in the other.

Next, we examine the relationship between the frequency of principals' classroom observations and teachers' value-added estimates using a similar regression specification (i.e., Equation 1). We believe ours is the first paper to examine this relationship and to ask whether value-added is correlated not only with principals' opinions but also with their actions. Since we have a single measure for observations, we use the average of math and ELA value-added for teachers with estimates in both subject areas. Here, we find evidence that principals conducted more formal classroom observations with teachers who had lower value-added estimates (Table 6, Column 1). The estimated relationship is robust to the inclusion of teacher experience controls (Column 2) and principal fixed effects (Column 3), though examining only the variation in observations within principals reduces the magnitude of the coefficient by half. It is also notable that principals steadily reduce the frequency of their formal observations as teachers gain more experience.

For total observations, we find similar results, suggesting that the results for formal observations are were not driven by a simple substitution between informal and formal methods. The estimated effect with principal fixed effects (Column 6) is noticeably higher than for formal observations only, suggesting that, at least within principals, there is additional informal observation of low value-added teachers.

Before turning to the impacts of the pilot program, it is worth noting that all of the results reported in this section are quite similar when we analyze treatment and control principals

separately. Thus, before the experiment, the value-added estimates were significantly and similarly predictive of opinions and actions of both sets of principals. In the next section, by contrast, we present results suggesting that the provision of the value-added reports created some important differences in how treatment and control principals evaluated their teachers by the end of the pilot year.

5. Impacts of Treatment on Principals Post-Experimental Ratings and other Outcomes

A first order question for the impact of providing principals with teacher value-added reports is: Did the value-added information affect principals' opinions regarding teacher effectiveness? We examine this issue by estimating regressions of principals' post-experimental ratings on teacher value-added while controlling for principals' pre-experimental ratings (and other covariates), as shown by Equation 3.

$$(3)\quad R_{it+1} = \alpha + \lambda R_{it} + \beta VA_{it} + \delta X_{it} + \varepsilon_{ijt}$$

The rating given to teacher *i* at time *t+1* (R_{it+1}) is specified as a function of the teacher's previous rating (R_{it}), the value-added estimate at time *t* (VA_{it}) and other teacher characteristics (X_{it}). We run these regressions separately for the treatment and control groups, and then compare the coefficients on VA_{it} between the two groups. Since only the treatment group received value-added reports, we might expect a significant positive relationship only for this group. However, if principals' opinions regarding teacher effectiveness tend to converge with value-added over time, then we might expect a significant positive relationship for both groups, but a stronger relationship for the treatment group. The results are shown in Table 7, with ELA in the top panel and math in the panel below. All regressions include teacher experience controls, though our

results are not sensitive to their inclusion, and we present regressions with and without principal fixed effects.

Our results suggest that the impact providing value-added reports to principals' ratings was weak and statistically insignificant for ELA, but much stronger and highly statistically significant for math. In regressions without principal fixed effects (Column Group 1), we find no effect of value-added in math on post-experimental ratings for the control group, but a large and highly significant effect for the treatment group. When we include principal fixed effects (Column Group 2), the effect for the control group is marginally significant (suggesting some general convergence over time), though the effect for the treatment group is still significantly higher and the difference between the groups remains both significant and similar in magnitude.[14]

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 ELA exams (i.e., given in January) increased principals' concerns about the ability of the value-added methodology to accurately measure a teacher's individual contribution in ELA. It may also be that principals were more confident in their ability to gauge the quality of ELA instruction, and therefore put less weight on the information in the value-added reports. Unfortunately, we cannot confirm or reject these or other potential explanations.

The ultimate goal of providing principals with information on value-added is to raise student achievement via improvements in teacher effectiveness. One of the channels through which this can occur is through selective retention of more effective teachers. A change in

[14] In an alternate specification, we replaced the continuous value-added measure with an indicator equal to one if the teacher was in the bottom quartile of the value-added distribution. Under this alternative approach we find very similar results—a marginally significant difference between treatment and control (in the expected direction) for math, but an insignificant difference for ELA. The difference between treatment and control is, however, relatively less statistically significant in math.

principals' opinions of teachers may or may not, however, result in changes in personnel decisions. We therefore examine how the propensity of teachers to exit their schools after the pilot year was related to their value-added estimates, and whether this relationship differed between treatment and control schools. If principals use information on teacher effectiveness to help them selectively retain teachers, one might expect to see a significant relationship in the treatment group—with lower value-added teachers more likely to exit the school—and no relationship in the control group.

Of the 2,173 teachers who remained in their baseline school during the school year 2007-2008, 273 (about 13 percent) did not continue to teach in the same school the following year.[15] Figure 1 shows the fraction of teachers that exited their school following the school year 2007-2008 broken down by treatment and control and the teacher's value-added quartile. For both ELA and math, teachers in the bottom quartile of value-added were noticeably more likely to exit the school if they were in the treatment group. In math, the exit probability declines steadily across the quartiles for the treatment group, but not for the control group; in ELA, exit rates for teachers in the higher quartiles of value-added were similar in treatment and control groups. This graphical evidence suggests that providing value-added information may have shifted teacher retention in the treatment group toward retaining higher value-added teachers.

We investigate these trends formally using regression analysis. Our specification posits a teacher's exit outcome as a function of the value-added estimate, the principal's pre-experimental rating of the teacher, and teacher experience, as shown by equation 4:

$$(4)\quad E_{it+1} = \alpha + \lambda R_{ijt} + \beta VA_{ijt} + \delta X_{it} + \varepsilon_{ij}$$

[15] The probability of exit was slightly higher for treatment schools (13.5 percent) than control schools (12 percent), but regressing an indicator for leaving on a treatment indicator, we find this difference is not statistically significant.

An indicator variable (E_{it+1}) equal to one if teacher i exited the school at time $t+1$ is specified as a function of the value-added estimate in subject j at time t (VA_{ijt}), the principal's rating in subject j at time t (R_{ijt}), and other teacher characteristics (X_{it}). In an alternate specification we replace the continuous value-added score with an indicator variable equal to one if the teacher is in the bottom quartile of value-added scores. Note that while exiting is a singular event for any given teacher, we predict the event using ELA and math value-added and ratings separately to investigate the relative importance of effectiveness in each subject. Results for ELA and math are presented in Table 8, and we present specifications with and without principal fixed effects. In Column Sets 1 and 2 we include value-added as a linear term, while Column Sets 3 and 4 compare teachers with value-added estimates in the bottom quartile to all other teachers. As with our analysis of principals' post-experimental ratings, we run regressions separately for treatment and control groups and then test for differences between the groups.

Our results provide suggestive evidence that the value-added reports did indeed cause a stronger negative relationship between teacher effectiveness and exit propensity. In specifications without principal fixed effects, the coefficient on value-added in both ELA and math is negative for the treatment schools and positive for the control schools. For math, the treatment group coefficient is marginally statistically significant (p-value 0.11) and we can reject the equality of the coefficients at the 10 percent level. For ELA, the difference in the coefficients between treatment and controls is of similar magnitude as math, but not precisely estimated (p-value 0.29). Adding principal fixed effects to the regression does not change the qualitative nature of the results. We do not find a linear effect of value-added in ELA, but for math the coefficient is negative and statistically significant (p-value 0.03) and we can reject equality of the coefficients for the two groups at the 14 percent level.

When we focus on teachers in the bottom quartile of the value-added estimates, the regression results are consistent with the graphical evidence and again suggest that lower value-added teachers had a higher propensity to leave due to the treatment. Teachers with ELA value-added estimates in the bottom quartile were more likely to exit their schools than teachers with value-added in the top three quartiles. In specifications with and without principals fixed effects, this difference was significantly different from zero and significantly different from the control group coefficient (which was of the opposite sign).

The results in Table 8 also suggest that, conditional on the information in the value-added reports, teachers' exit probabilities were more strongly related to principals' pre-existing opinions regarding teacher effectiveness in the control group. Thus, principals do seem to use their subjective evaluations of teachers to make personnel decisions, but treatment principals' decisions seemed to place some weight on the value-added reports and less on their prior beliefs.

One potential concern with these results on teachers' exit propensity is that treatment principals may simply have been more adept or inclined to screen out teachers with low value-added independent of their receiving a report showing actual value-added scores. We can test this alternate hypothesis by examining the probability that a teacher exited the school before the start of the pilot year (i.e., between the school years 2006-2007 and 2007-2008). Doing so, we find no evidence to support this alternate explanation (Table 9). Indeed, these "placebo tests" show a nearly identical relationship between value-added and the propensity to exit prior to the pilot in treatment and control schools. Moreover, the coefficients on principals' pre-existing beliefs regarding teacher effectiveness are also nearly identical between the two groups and both highly significant. Again, this supports the notion that treatment and control principals made

personnel decisions based on their subjective evaluations of teacher effectiveness, and did so to a similar degree prior to the start of the experiment.[16]

There are two important issues to consider when thinking about the implications of these results on teacher exiting for student achievement. First, it is unclear whether current value-added estimates are valid predictors (conditional on baseline ratings) of how teachers will perform in the future, even though the value-added estimates are correlated with principals' baseline ratings. Prior research (especially Kane and Staiger, 2008) suggests that value-added estimates do predict future performance, but the only way to directly test the impacts of teacher exit on student achievement is to examine student outcomes in these schools during the school year 2008-2009, which are not yet available. The second issue to consider is whether these teacher exits represent movements into other schools or exits from the teaching profession. Even if we take as given that the value-added estimates are valid predictors of future teacher performance, movement of a low value-added teacher from one school to another is unlikely to result in any net benefit to students.

Of the 273 teachers who left their baseline school after the school year 2007-2008, 25 percent (68 teachers) moved to teach in another school within the DOE. In order to be check whether these "school-to-school movers" are driving the results in Table 8, we examine the relationship between value-added and the propensity to either exit the DOE or move to another school within the DOE using a multinomial logistic regression. Leaving the New York public schools is, of course, an imperfect measure of exiting the teaching profession, but it is the best proxy we have in our data. For simplicity, we only present results for the linear specification in math and the quartile specification in ELA, since these correspond with the most salient findings

[16] Of course, one could interpret these coefficients to mean that teacher exits have a negative effect on principals' opinions of teacher effectiveness, but we find this alternative "reverse causality" explanation hard to believe.

from Figure 1 and Table 8. In specifications without fixed effects, we find point estimates consistent with the notion that the overall school exit results were driven by both movements to other schools and movements out of the district (Table 10, Column Groups 1 and 3). When we add principal fixed effects (Column Groups 2 and 4), the coefficient on math value-added for movement to another school becomes large and very imprecisely estimated for the control group, but the other estimates remain largely unchanged. Thus, if teachers with low-value estimates are ineffective at raising student achievement, our results suggest the treatment certainly produced benefits to students in the treatment group schools, but that benefits to students more broadly may have been dampened by movement of low value-added teachers into other schools.

Another way to investigate how principals use the value-added reports is to examine the formal evaluations given to teachers by principals. Regulations require that principals perform classroom observations of all teachers and evaluate them as either satisfactory or unsatisfactory. However, receiving an unsatisfactory rating is a relatively rare occurrence. Of the 2,114 teachers who remained in their baseline schools during the pilot year, 39 (1.84 percent) were given an unsatisfactory rating in the school year 2007-2008.[17] Nevertheless, we believe an analysis of these ratings is useful since, unlike exit behavior, the ratings are chosen by principals, not teachers. We use the same specifications as in our exit analysis, and, for simplicity, present only results that include principal fixed effects and teacher experience controls (Table 11). Here we find suggestive evidence that teachers with lower value-added estimates in math were more likely to be rated unsatisfactory in treatment schools but not in control schools. For ELA, however, differences between treatment and control schools in the relationship between value-added estimates and these formal ratings are insignificant.

[17] The portion rated unsatisfactorily was slightly higher in the treatment group (2 percent) than the control group (1.5 percent) but this difference was not statistically significant.

Finally, we have examined several potential sources of heterogeneity in the impacts of receiving the treatment on post-experimental outcomes. Particularly, we have estimated specifications that allow impacts to differ between untenured and tenured teachers, between elementary and middle schools, between schools that received a B or higher accountability grade and those receiving a C or lower, and between treatment schools in which the principal shared value-added reports with teachers and those in which the principal did not.[18] In general, the impacts did not differ significantly by subgroup and, therefore, are not reported. The only notable differences that were statistically significant were that (1) teacher exit was more strongly related to the pre-experimental rating for treatment principals who shared their value-added reports with teachers, and (2) for both treatment and control schools, teachers with 0-2 years of experience were more likely to exit the school before the school year 2007-2008 if their principal rated them poorly prior to the experiment.

6. Conclusion

A number of important facts emerge from our analysis. First, even before reports were distributed, value-added was correlated with principals' beliefs about teacher effectiveness, as reflected in both their opinions (teacher ratings) and their actions (teacher observations). Second, principals changed their ratings of teachers in response to the information in the value-added report. Thus, the value-added reports provided new information to principals that influenced their evaluation of teacher performance. Finally, our results suggest that providing the information in the value-added reports raised the probability that low value-added teachers would leave the school.

[18] Details on New York's school accountability system can be found in Rockoff and Turner (2008). Essentially, a school receiving a low accountability grade faces consequences (including the removal of the principal) if they do not improve student performance on standardized achievement exams and other measures of school quality.

The implications of our findings regarding teachers' exit probabilities are somewhat unclear, since teachers with lower value-added estimates in the treatment schools seem to have been more likely both to move out of teaching in New York City schools and to move to another public school within the district. In the context of this pilot program, the value-added reports constituted private information for the principal who received them, and principals considering hiring a teacher from a treatment school would not know if a teacher received high or low value-added scores. The privacy of information on teacher performance may create an important inefficiency in the teacher labor market, and contribute to the continued employment of poor performing teachers in public schools, a process dubbed the "dance of the lemons" by education professionals (see Ravitch (2007)). This suggests that if the value-added reports contain information useful to principals about their current teachers, they might also provide useful information to principals considering experienced applicants from other schools. The introduction of such information into the teacher labor market may especially benefit poor and low performing students; research by Boyd et al. (2007) suggests that low value-added teachers working in schools serving more advantaged students tend to transfer to schools serving disadvantaged student populations and continue to perform poorly in the years after they transfer.

Overall, our findings suggest that value-added reports provide principals with useful information about teacher effectiveness. However, in future work, we will further evaluate the benefits of providing this information to principals by asking whether test scores of students in the treatment schools improved relative to control schools. Student test scores could improve either because the information was used to help teachers improve their classroom performance or through the exit of lower performing teachers from the school. Given the explicit prohibition of principals from using the value-added information for teacher evaluation and tenure decisions,

one might expect the impacts through the latter channel to be muted in our treatment group. Nevertheless, this experiment may provide a lower bound on the potential impact that providing such reports have on student test scores.

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.

The Battelle Memorial Institute, Health and Life Sciences Division (2009) "NYC Teacher Data Initiative Technical Report: Development of Model to Measure Teacher Value-Added."

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

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

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

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

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

Kane, T. J., Rockoff, J. and Staiger, D. O. (2006) "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City" NBER Working Paper #12155.

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

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

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

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

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

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

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.

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

	Baseline (112 Treatment, 111 Control)				*Followup (84 Treatment, 94 Control)*			
	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C	Control Mean	Treatment Mean	Treatment - Control	P-value H_0: T=C
Principal Characteristics								
Years of Experience as Principal (in School)	3.3	3.2	-0.1	0.79	3.2	3.2	0.0	0.94
Years of Experience as Assistant Principal	2.4	2.7	0.3	0.4	2.4	2.8	0.3	0.43
Years of Experience as Teacher	6.8	7.3	0.6	0.45	6.6	7.6	1.1	0.23
Years of Experience in School (Any Position)	4.4	5.0	0.6	0.32	4.3	4.8	0.4	0.53
Principal Age	48.0	48.8	0.8	0.5	48.2	49.9	1.7	0.16
Principal is Black or Hispanic	48.6%	41.9%	-6.7%	0.32	47.9%	44.1%	-3.8%	0.61
Principal is Female	81.1%	77.7%	-3.4%	0.53	80.9%	77.4%	-3.5%	0.57
Student Characteristics								
On Free Lunch	86.7%	85.4%	-1.3%	0.59	85.9%	84.9%	-1.0%	0.73
English Language Learners	15.7%	14.1%	-1.6%	0.28	15.6%	13.8%	-1.8%	0.31
In Special Education	9.5%	9.5%	0.0%	0.99	9.2%	9.7%	0.5%	0.63
Black	42.8%	40.8%	-2.0%	0.58	41.5%	41.7%	0.2%	0.97
Hispanic	29.5%	31.8%	2.3%	0.55	30.4%	31.1%	0.7%	0.88
School Environment (Teacher Survey, Spring 2007)								
The Principal…								
Places Children's Learning Needs Ahead of Other Interests	-0.04	-0.04	-0.001	0.99	-0.039	-0.012	0.027	0.85
Is an Effective Manager	-0.08	-0.058	0.018	0.89	-0.059	-0.018	0.041	0.79
Supports Me	-0.14	-0.113	0.026	0.84	-0.119	-0.099	0.02	0.89
Visits Classrooms to Observe the Quality of Teaching	0.07	-0.03	-0.101	0.44	0.091	-0.026	-0.117	0.44
Gives Me Regular and Helpful Feedback	-0.05	-0.069	-0.022	0.86	-0.026	-0.101	-0.075	0.59
Places a High Priority on the Quality of Teaching	-0.03	0.004	0.031	0.8	-0.031	0.034	0.065	0.64
Teachers in this School…								
Use Student Data to Improve Instructional Decisions	0.10	0.077	-0.019	0.87	0.093	0.058	-0.035	0.8
Receive Training in the Use of Student Data	0.04	0.022	-0.014	0.9	0.049	0.012	-0.037	0.79

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

Table 2: Attrition Between Baseline and Follow-up Surveys

Panel A: Principals	All		Treatment		Control	
Completed Baseline Survey	223	100.0%	112	100.0%	111	100.0%
Did Not Withdraw from Pilot Before Follow-up	220	98.7%	110	98.2%	110	99.1%
Responded to Follow-up Survey	178	79.8%	84	75.0%	94	84.7%
Completed Follow-up Survey	172	77.1%	79	70.5%	93	83.8%
Panel B: Teachers	All		Treatment		Control	
In Baseline Survey with Value Added Estimate	2,509	100.0%	1,324	100.0%	1,185	100.0%
Taught in Same School in School Year 2007-08	2,114	84.3%	1,087	82.1%	1,027	86.7%
Principal Responded to Follow-up Survey	1,775	70.7%	866	65.4%	909	76.7%
In Value-Added Subject/Grade in 2007-08	1,557	62.1%	757	57.2%	800	67.5%
Rated in Follow-up Survey by the Same Principal	1,527	60.9%	747	56.4%	780	65.8%

Table 3: Summary Statistics on Teacher Level Variables

	Baseline Survey		Followup Survey	
	Control	Treatment	Control	Treatment
In Baseline Survey with Value Added Estimate	1185	1324	780	747
Teacher Experience in School Year 2006-2007				
None (First Year of Teaching was 2006-2007)	9.0%	10.8%	9.6%	10.4%
One Year	11.7%	10.8%	10.6%	8.6%
Two Years	10.9%	10.6%	11.0%	8.6%
Three Years	8.6%	10.9%	8.3%	12.4%
Four Years	7.4%	6.8%	6.9%	6.4%
Five to Nine Years	27.9%	26.8%	28.2%	28.1%
Ten or More Years	24.6%	23.2%	25.3%	25.4%
Principal's Rating (Scale from 0 to 5)				
Math Instruction	3.21	3.23	3.46	3.50
	(1.09)	(1.13)	(0.99)	(1.03)
ELA Instruction	3.20	3.19	3.45	3.43
	(1.04)	(1.13)	(1.02)	(1.03)
Observations Made by Principal Last Year				
Formal	2.21	2.24	1.92	1.98
	(1.26)	(1.25)	(0.99)	(1.26)
Total	6.51	6.26	4.92	4.75
	(3.38)	(3.21)	(3.27)	(3.13)
Value-added Estimates				
Math, Multi-year, City Comparison	0.004	0.002	0.004	0.016
	(0.166)	(0.181)	(0.159)	(0.171)
Math, Multi-year, Peer Comparison	0.006	0.013	0.009	0.022
	(0.137)	(0.150)	(0.132)	(0.142)
Math, Single-year, Peer Comparison	-0.001	0.013	0.002	0.025
	(0.155)	(0.163)	(0.148)	(0.154)
ELA, Multi-year, City Comparison	-0.013	-0.001	-0.018	0.012
	(0.135)	(0.126)	(0.132)	(0.122)
ELA, Multi-year, Peer Comparison	-0.011	0.002	-0.016	0.010
	(0.092)	(0.093)	(0.086)	(0.091)
ELA, Single-year, Peer Comparison	-0.012	0.004	-0.019	0.013
	(0.106)	(0.110)	(0.101)	(0.111)

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

Table 4a: Principals' Pre-experimental Ratings of Teacher Performance and Value-Added, ELA

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-added Score, Multi-year, Peer Comparison	2.432**		1.391*		1.521**	2.503**	2.919**	2.424**
	(0.322)		(0.577)		(0.488)	(0.309)	(0.338)	(0.411)
Value-added Score, Single-year, Peer Comparison		2.048**	1.046*					
		(0.269)	(0.494)					
Value-added Score, Multi-year, Citywide Comparison				1.679**	0.938**			
				(0.222)	(0.341)			
Value-added Score, Multi-Year, Peer Comparison								0.919
* Teacher Experience > 2 Years								(0.582)
Teacher Experience (10+ Years is Omitted Group)								
0 Years Experience						-0.494**	-0.406**	-0.407**
						(0.099)	(0.092)	(0.092)
1 Years Experience						0.097	0.061	0.049
						(0.113)	(0.103)	(0.102)
2 Years Experience						0.038	-0.075	-0.084
						(0.099)	(0.088)	(0.088)
3 Years Experience						0.167+	0.160+	0.158+
						(0.095)	(0.096)	(0.096)
4 Years Experience						0.259**	0.272**	0.266**
						(0.094)	(0.090)	(0.089)
5-9 Years Experience						0.175*	0.176*	0.173*
						(0.072)	(0.069)	(0.069)
Principal Fixed Effects							Y	Y
R-squared	0.043	0.042	0.046	0.041	0.050	0.074	0.389	0.390
Sample Size	1,963	1,963	1,963	1,963	1,963	1,942	1,942	1,942

Note: The dependent variable is the principal's rating of the teacher's effectiveness in English Language Arts instruction. Each column reports estimates from separate regressions; Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1. Sample: Teachers with value-added available for reporting in the 2007-08 school year and a rating in ELA from their principal in the intial survey.

Table 4b: Principals' Pre-experimental Ratings of Teacher Performance and Value-Added, Math

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Value-added Score, Multi-year, Peer Comparison	2.082**		2.056**		1.268**	2.140**	2.089**	1.752**
	(0.216)		(0.458)		(0.362)	(0.226)	(0.212)	(0.252)
Value-added Score, Single-year, Peer Comparison		1.678**	0.027					
		(0.181)	(0.390)					
Value-added Score, Multi-year, Citywide Comparison				1.685**	0.845**			
				(0.166)	(0.282)			
Value-added Score, Multi-Year, Peer Comparison								0.617+
* Teacher Experience > 2 Years								(0.350)
Teacher Experience (10+ Years is Omitted Group)								
0 Years Experience						-0.345**	-0.430**	-0.423**
						(0.102)	(0.099)	(0.099)
1 Years Experience						0.021	-0.095	-0.096
						(0.105)	(0.098)	(0.099)
2 Years Experience						0.052	-0.048	-0.05
						(0.091)	(0.077)	(0.077)
3 Years Experience						0.221*	0.228*	0.227*
						(0.094)	(0.092)	(0.092)
4 Years Experience						0.361**	0.272**	0.269**
						(0.098)	(0.087)	(0.087)
5-9 Years Experience						0.284**	0.267**	0.261**
						(0.069)	(0.065)	(0.065)
Principal Fixed Effects							Y	Y
R-squared	0.074	0.059	0.074	0.071	0.081	0.106	0.396	0.397
Sample Size	2,048	2,048	2,048	2,048	2,048	2,026	2,026	2,026

Note: The dependent variable is the principal's rating of the teacher's effectiveness in mathematics instruction. Each column reports estimates from separate regressions; Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1. Sample: Teachers with value-added available for reporting in the 2007-08 school year and a rating in math from their principal in the intial survey.

Table 5: Principals' Pre-experimental Ratings and Subject Specific Value-Added

	Math Rating		ELA Rating	
	(1)	(2)	(3)	(4)
Value-added Score, Math	0.237		0.115	
	(0.087)**		(0.083)	
Value-added Score, ELA		-0.044		0.526
		(0.127)		(0.142)**
Principal's Pre-experimental Rating in Other Subject	0.887	0.894	0.914	0.907
	(0.014)**	(0.014)**	(0.013)**	(0.014)**
R-squared	0.82	0.82	0.82	0.82
Sample Size	1527	1527	1527	1527

Note: "Rating in Other Subject" denotes ELA when the dependent variable is rating, and vice versa. Standard errors (in parentheses) are clustered by school. **p < 0.01, *p<0.05, +p<0.1.

Table 6: Principals' Pre-experimental Classroom Observations and Value-added

	Formal Observations			Total Observations		
	(1)	(2)	(3)	(4)	(5)	(6)
Value-added, Math/ELA Average	-0.723*	-0.905*	-0.451**	-0.642	-0.995	-0.856**
	(0.364)	(0.361)	(0.136)	(0.860)	(0.922)	(0.270)
Teacher Experience (10+ Years is Omitted Group)						
0 Years Experience		1.220**	1.040**		1.417**	0.966**
		(0.119)	(0.094)		(0.368)	(0.142)
1 Years Experience		1.100**	0.920**		1.038**	0.610**
		(0.124)	(0.079)		(0.326)	(0.106)
2 Years Experience		0.934**	0.811**		0.722*	0.571**
		(0.109)	(0.079)		(0.291)	(0.107)
3 Years Experience		0.572**	0.520**		0.474	0.351**
		(0.124)	(0.082)		(0.330)	(0.120)
4 Years Experience		0.341**	0.219**		-0.077	0.115
		(0.103)	(0.064)		(0.360)	(0.109)
5-9 Years Experience		0.179*	0.068+		0.045	0.107
		(0.085)	(0.038)		(0.219)	(0.071)
Principal Fixed Effects			Y			Y
R-squared	0.004	0.134	0.763	0.000	0.024	0.887
Sample Size	2,508	2,485	2,485	2,489	2,466	2,466

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

Table 7: The Impact of Value-added Information on Principals' Post-Survey Ratings of Teacher Effectiveness

	(1)			(2)		
	Treatment	Control	Difference	Treatment	Control	Difference
English Language Arts (ELA)						
Principal's Pre-experiment Rating	0.639**	0.663**	-0.024	0.550**	0.656**	-0.105
	(0.042)	(0.044)		(0.053)	(0.046)	
Value-added Score, Multi-year, Peer Comparison	0.285	0.036	0.249	0.651	0.582	0.069
	(0.429)	(0.411)		(0.451)	(0.419)	
Experience Controls	Y	Y		Y	Y	
Principal Fixed Effects				Y	Y	
R-squared	0.439	0.419		0.605	0.628	
Sample Size	583	607		583	607	
Math						
Principal's Pre-experiment Rating	0.581**	0.662**	-0.08	0.525**	0.651**	-0.125*
	(0.042)	(0.039)		(0.050)	(0.045)	
Value-added Score, Multi-year, Peer Comparison	1.320**	0.244	1.076**	1.398**	0.427+	0.971**
	(0.254)	(0.247)		(0.335)	(0.220)	
Experience Controls	Y	Y		Y	Y	
Principal Fixed Effects				Y	Y	
R-squared	0.473	0.468		0.648	0.637	
Sample Size	615	631		615	631	

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

Table 8: Teachers' Propensity to Exit Schools and Value-added

	(1)			(2)			(3)			(4)		
ELA	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Value-added Score	-0.117	0.094	-0.211	-0.013	0.081	-0.094						
	(0.139)	(0.146)	[0.291]	(0.143)	(0.179)	[0.656]						
Bottom Quartile, Compared to Top Three Quartiles							0.068*	-0.044+	0.112**	0.077*	-0.038	0.115**
							(0.030)	(0.024)	[0.003]	(0.034)	(0.027)	[0.005]
Principal's Pre-experimental Rating	-0.003	-0.039**	0.036+	-0.002	-0.02	0.018	-0.001	-0.039**	0.038*	0.005	-0.020	0.026
	(0.012)	(0.014)		(0.014)	(0.018)		(0.012)	(0.014)		(0.013)	(0.017)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Principal Fixed Effects				Y	Y	Y				Y	Y	Y
R-squared	0.06	0.048		0.267	0.232		0.065	0.05		0.273	0.234	
Sample Size	854	814		854	814		854	814		854	814	
Math												
Value-added Score	-0.155	0.062	-0.218	-0.211*	-0.007	-0.204						
	(0.096)	(0.093)	[0.101]	(0.098)	(0.110)	[0.136]						
Bottom Quartile, Compared to Top Three Quartiles							0.011	-0.031	0.042	-0.004	-0.016	0.013
							(0.035)	(0.026)	[0.336]	(0.035)	(0.031)	[0.767]
Principal's Pre-experimental Rating	-0.017	-0.031*	0.013	-0.004	-0.008	0.005	-0.024*	-0.031*	0.007	-0.013	-0.010	-0.003
	(0.011)	(0.014)		(0.015)	(0.016)		(0.011)	(0.013)		(0.014)	(0.015)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Principal Fixed Effects				Y	Y	Y				Y	Y	Y
R-squared	0.062	0.035		0.241	0.218		0.058	0.035		0.236	0.218	
Sample Size	916	838		916	838		916	838		916	838	

Note: 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 9: Teachers' Propensity to Exit Schools Prior to the Experiment and Value-added (Placebo Test)

	(1)			(2)		
ELA	Treatment	Control	Difference	Treatment	Control	Difference
Value-added Score	0.088	0.136	-0.048			
	(0.166)	(0.146)	[0.815]			
Bottom Quartile, Compared to Top Three Quartiles				-0.036	-0.025	-0.011
				(0.034)	(0.032)	[0.801]
Principal's Pre-experimental Rating	-0.054**	-0.052**	-0.002	-0.055**	-0.050**	-0.005
	(0.016)	(0.018)		(0.016)	(0.018)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y
Principal Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.23	0.223		0.231	0.222	
Sample Size	1,014	928		1,014	928	
Math						
Value-added Score	0.043	0.040	0.003			
	(0.112)	(0.105)	[0.984]			
Bottom Quartile, Compared to Top Three Quartiles				0.006	-0.023	0.030
				(0.036)	(0.042)	[0.564]
Principal's Pre-experimental Rating	-0.046**	-0.044**	-0.001	-0.043**	-0.045**	0.001
	(0.016)	(0.015)		(0.015)	(0.015)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y
Principal Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.187	0.202		0.186	0.203	
Sample Size	1,079	947		1,079	947	

Note: 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 10: Value-added and Exiting the District vs. Moving Schools

	Math						***ELA***					
	(1)			(2)			(3)			(4)		
	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference	Treatment	Control	Difference
Move to Another School in the District												
Value-added Score	-3.132*	-0.276	-2.855	-2.628	-5.298	2.670						
	(1.554)	(1.799)	[0.229]	(1.740)	(4.561)	[0.584]						
Bottom Quartile, Compared to Top Three Quartiles							0.534	-0.454	0.988	0.471	-0.903	1.374
							(0.487)	(0.524)	[0.166]	(0.808)	(0.802)	[0.226]
Principal's Pre-experimental Rating	-0.026	-0.282	0.256	0.267	-0.008	0.275	-0.011	-0.522+	0.511	0.05	-0.351	0.401
	(0.196)	(0.303)	[0.477]	(0.414)	(0.418)	[0.639]	(0.235)	(0.269)	[0.151]	(0.383)	(0.320)	[0.421]
Exit Teaching in the District												
Value-added Score	-0.587	0.898	-1.485	-1.523	0.911	-2.434						
	(0.787)	(1.004)	[0.243]	(1.073)	(1.656)	[0.216]						
Bottom Quartile, Compared to Top Three Quartiles							0.569*	-0.466	1.035*	0.846*	-0.499	1.345*
							(0.257)	(0.327)	[0.013]	(0.385)	(0.425)	[0.019]
Principal's Pre-experimental Rating	-0.235+	-0.376*	0.14	-0.161	-0.230	0.069	-0.004	-0.344**	0.340+	0.07	-0.225	0.296
	(0.134)	(0.148)	[0.480]	(0.205)	(0.202)	[0.810]	(0.132)	(0.127)	[0.064]	(0.178)	(0.183)	[0.246]
Teacher Experience Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Principal Fixed Effects				Y	Y	Y				Y	Y	Y
R-squared	0.084	0.062		0.399	0.422		0.082	0.064		0.423	0.406	
Sample Size	916	838		916	838		854	814		854	814	

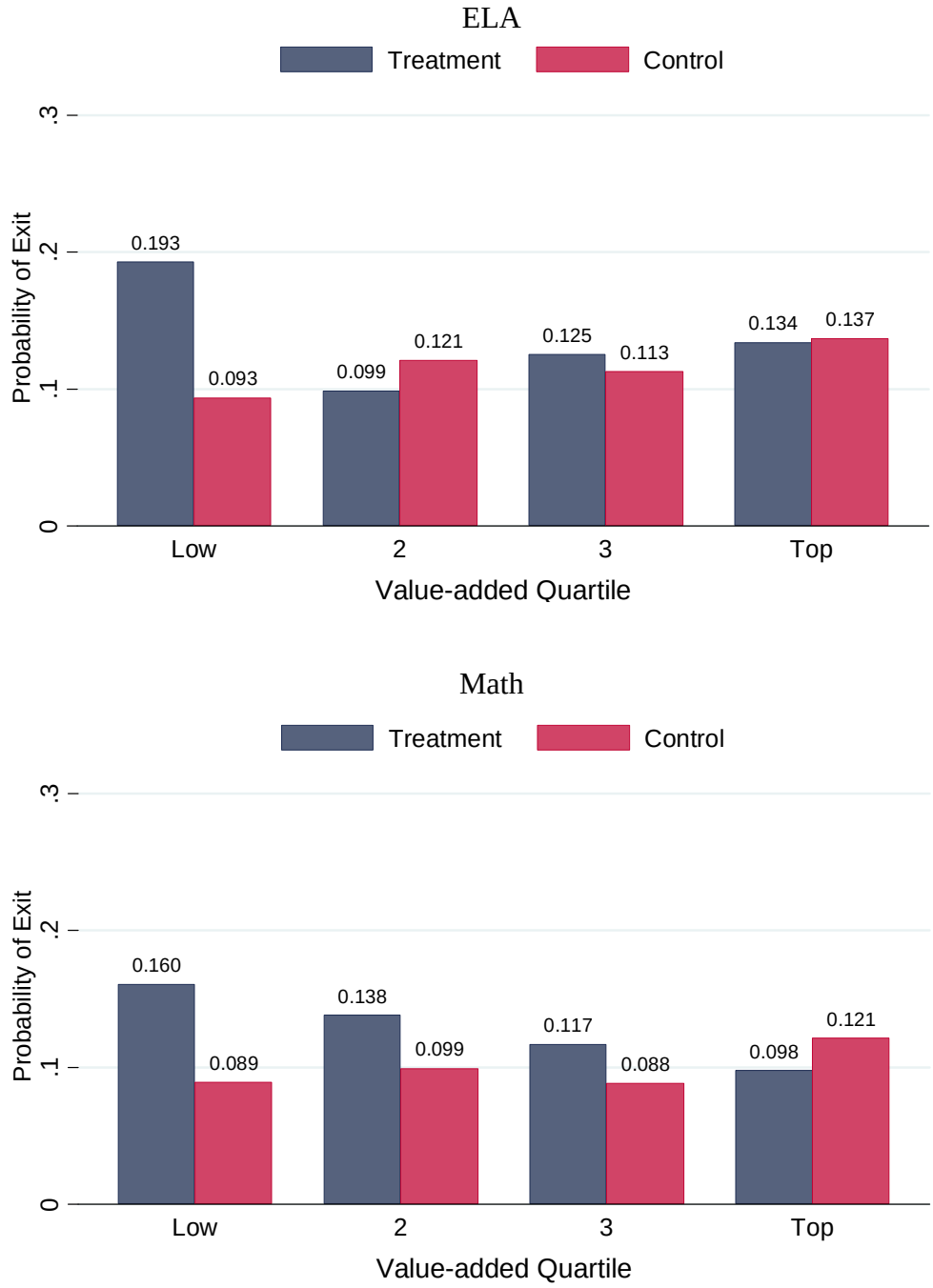
Note: Standard errors (in parentheses) are clustered by school; p-values on the test of differences in brackets. **p < 0.01, *p<0.05, +p<0.1.

Table 11: Teachers' Propensity to Receive an Unsatisfactory Evaluation and Value-Added

	(1)			(2)		
ELA	Treatment	Control	Difference	Treatment	Control	Difference
Value-added Score	0.007	-0.017	0.024			
	(0.092)	(0.043)	[0.799]			
Bottom Quartile, Compared to Top Three Quartiles				0.015	0.024+	-0.009
				(0.017)	(0.014)	[0.655]
Principal's Pre-experimental Rating	-0.054**	-0.052**	-0.002	-0.055**	-0.050**	-0.005
	(0.016)	(0.018)		(0.016)	(0.018)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y
Principal Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.214	0.207		0.215	0.212	
Sample Size	854	814		854	814	
Math						
Value-added Score	-0.065	0.024	-0.089+			
	(0.049)	(0.027)	[0.088]			
Bottom Quartile, Compared to Top Three Quartiles				0.035+	-0.013	0.048*
				(0.020)	(0.016)	[0.048]
Principal's Pre-experimental Rating	-0.046**	-0.044**	-0.001	-0.043**	-0.045**	0.001
	(0.016)	(0.015)		(0.015)	(0.015)	
Teacher Experience Controls	Y	Y	Y	Y	Y	Y
Principal Fixed Effects	Y	Y	Y	Y	Y	Y
R-squared	0.241	0.252		0.245	0.252	
Sample Size	916	838		916	838	

Note: 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.

Figure 1: Probability of Exit by Value-Added Quartile, Treatment vs. Control Group



Note: The sample is limited to teachers working in the same school in the school years 2006-2007 and 2007-2008 who were rated by their principal in the baseline survey and have a value-added estimate. Exit is defined as working in a different school or not working in any school in the school year 2008-09.

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 Teachers (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	0.46	0.55	0.09	0.25
Asst. Principal also Contributed to Survey	0.46	0.39	-0.08	0.32
Lead Teacher also Contributed to Survey	0.12	0.12	-0.01	0.91
Other Person also Contributed to Survey	0.28	0.14	-0.13	0.03*
Top 4 Ways to Assess Teachers (Other than Observation) Include				
State Level Standardized Tests	0.96	0.93	-0.03	0.38
Student Work	0.82	0.84	0.02	0.7
Periodic Assessments	0.56	0.58	0.02	0.78
End of Course Exams	0.22	0.17	-0.04	0.49
Other Student Tests	0.08	0.11	0.04	0.42
Feedback from Other Administrators	0.59	0.59	0.00	0.99
Feedback from Students	0.29	0.26	-0.03	0.65
Feedback from Parents	0.18	0.15	-0.04	0.54
Feedback from Other Teachers	0.11	0.19	0.08	0.15
Teacher Work Portfolio	0.13	0.10	-0.03	0.54
Other School Related Tasks	0.08	0.09	0.01	0.79
To Evaluate Individual Teachers in Past Year, Principal Used				
Average State Test Scores	0.86	0.94	0.08	0.09+
Average State Test Scores by Subgroup	0.76	0.81	0.05	0.44
Average Growth in State Test Scores	0.82	0.91	0.10	0.07+
Value-Added Reports *(Treatment Only)*		0.55		
Percentage of Students Not Meeting Standards on State Tests	0.86	0.86	0.01	0.9
Percentage of Students by Proficiency Level	0.91	0.93	0.01	0.78
Change in Percentage of Students by Proficiency Level	0.85	0.88	0.03	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)				
Teaching Experience	3.62	3.74	0.13	0.46
Prior Performance of Students on Standardized Tests	4.58	4.37	-0.21	0.08+
Percentage ELL/Special Education Students in a Teacher's Class	4.10	4.30	0.21	0.23
Class Size	3.58	3.81	0.23	0.21
The Number of Students who Entered the class mid-year.	3.53	4.01	0.48	0.01*
Which Teacher(s) the Students Had in the Previous Year	3.68	4.12	0.44	0.001*
If a Teacher Recently Started Teaching a New Grade/Subject	3.91	4.13	0.22	0.17
If a Teacher had a Personal Issue During the Year	3.37	3.66	0.29	0.1+
Things that Distracted the Teacher's Class on the Test Day	3.07	3.12	0.05	0.81
Outside Help a Teacher's Students Received (e.g., after-school)	3.81	4.00	0.19	0.21
Help a Teacher Received from an Aide in the Classroom.	3.11	3.21	0.10	0.58
The Teacher's Performance in Teaching Non-tested Subjects	3.30	3.54	0.24	0.16

Note: There are 82 treatment schools and 93 control schools. P-values indicate the statistical significance of a treatment indicator to predict the survey response.

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: 82 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.