Subjective and Objective Evaluations of Teacher Effectiveness

Jonah E. Rockoff
Columbia Business School and NBER

Cecilia Speroni*
Teachers College, Columbia University

December 2009

Abstract

A substantial literature documents large variation in teacher effectiveness at raising student achievement, providing motivation to identify highly effective and ineffective teachers early in their careers. In this paper, we document how two programs designed to evaluate new teachers' skills have predictive power for the achievement gains made by teachers' future students. We find that these subjective evaluations have substantial power compared with objective measures of teacher effectiveness taken from a teacher's first year in the classroom.

* Email: jonah.rockoff@columbia.edu; cs2456@columbia.edu. Jonah Rockoff thanks the Smith Richardson Foundation for financial assistance and Lesley Turner for her work on the subjective evaluations data.

"I have an open mind about teacher evaluation, but we need to find a way to measure classroom success and teacher effectiveness. Pretending that student outcomes are not part of the equation is like pretending that professional basketball has nothing to do with the score."

- Arne Duncan, U.S. Secretary of Education, Remarks to the Education Writers Association April 30th, 2009

A large body of research demonstrates the importance of teacher effectiveness in raising student achievement. This literature has extensive roots (e.g., Hanushek (1971), Brophy and Good (1986)), and has grown due to the availability of large administrative datasets that link student outcomes to classroom teachers (e.g., Sanders and Rivers (1996), Rockoff (2004), Rivkin et al. (2005), Harris and Sass (2006), Aaronson et al. (2007), and Clotfelter et al. (2007)). Two stylized facts from this work are that (1) teacher effectiveness (sometimes referred to as "value-added") varies widely and (2) outside of teaching experience, the characteristics used to certify and pay teachers bear little relation to student outcomes. These findings provide motivation to understand better how effective and ineffective teachers can be identified early in their careers.

In this paper, we measure the extent to which a set of subjective and objective evaluations of teacher effectiveness can predict teachers' future impacts on student achievement. The subjective evaluations come from an alternative certification program that evaluates its applicants prior to the start of their teaching careers and a mentoring program in which experienced educators work with new teachers and submit subjective evaluations of their effectiveness throughout the school year. Our objective measures of effectiveness are estimates of teachers' impacts on student achievement in the first year of their careers.

We find that both subjective and objective evaluations bear significant relationships with the achievement of the teachers' future students. In addition, when both subjective and objective evaluations are entered as predictors in a regression of future students' test scores, their

coefficients are only slightly attenuated. Thus, each type of evaluation contains information on teacher effectiveness that is distinct from the other.

Notably, we also find evidence of significant variation in the leniency with which standards of evaluation were applied by some evaluators of new teachers. Specifically, variation in subjective evaluations *within* evaluators is a much stronger predictor of teacher effectiveness than variation between evaluators. This highlights the importance of reliability in the procedures used to generate subjective evaluations.

I. Prior Literature

Several recent studies have examined how objective data on student learning from early in a teacher's career can be used to predict how teachers will impact student outcomes in the future. For example, Gordon et al. (2006) take measures of the effectiveness of teachers in Los Angeles using data from the first two years of their careers and, grouping teachers by quartiles, examine students' outcomes in these teachers' classrooms during the following year. They find large differences across quartiles—students with teachers in the top quartile gained 10 percentile points more than those assigned to teachers in the bottom quartile, about half the national Black-White achievement gap—and conclude that using data on student performance to identify and selectively retain teachers could yield large benefits for student achievement.

Tempering such findings is the reality that sampling variation and classroom level idiosyncratic shocks introduce noise into measures of teacher effectiveness solely based on student test scores, so that some teachers who initially appear effective may perform poorly in the future, and vice versa. Of equal concern is that estimates of teacher effectiveness may be biased if some teachers are persistently assigned students that are more or less difficult to teach

in ways that administrative data sets do not measure. For these reasons, it is important to understand how other measures of effectiveness can be used to achieve greater stability and accuracy in measures of effective teaching. Moreover, it is unlikely that an evaluation system purely based on student outcomes data would ever be implemented (see Weingarten (2007)).

There is a considerable literature on the power of subjective teaching evaluations to predict gains in student achievement. The largest focus has been on evaluations of teachers by the school principal, motivated by principals' authority in making personnel decisions.[1] A second strand of work examines the relation between teacher effectiveness and formal evaluations based on classroom observation protocols or "rubrics" (e.g. Holtzapple (2003), Schacter and Thum (2004), Gallagher (2004), Kimball et al. (2004), and Milanowski (2004)). With few exceptions, principal evaluations and classroom observations have been found to have significant predictive power to predict student achievement.[2]

The findings from these studies are quite encouraging, but there are two notable shortcomings that limit what we can learn from them about identifying effective new teachers using subjective evaluations. First and foremost, they investigate the power of evaluations to predict the exam performance of current, not future, students. A teacher may be highly rated because she has a group of students who are well behaved, cohesive, and highly motivated in ways that cannot be controlled for using regression analysis and available data. A stronger test of the power of these evaluations would be to predict gains produced by the teacher with a new group of students in a subsequent year (as done by Gordon et al. (2006) using objective

[1] This topic has been studied over a long period of time by educators (e.g., Hill (1921), Brookover (1945), Gotham (1945), Anderson (1954), Medley and Coker (1987), Manatt and Daniels (1990), Wilkerson et al. (2000)), but economists have also made significant contributions (e.g., Murnane (1975), Armor et al. (1976), Harris and Sass (2007), Jacob and Lefgren (2008), Rockoff et al. (2009)).

[2] For example, Jacob and Lefgren (2008) find that a one standard deviation increase in a principals' evaluation of a teacher is associated with higher test score performance of 0.10 and 0.05 standard deviations in math and English, respectively.

performance data). Second, it is unclear the extent to which principal evaluations are based on the results of prior student exams.

Ours is the first study to focus on subjective evaluations made prior to or just at the start of a teacher's career. It is also one of the few studies that tests how multiple sources of subjective evaluation predict teacher effectiveness.[3] Because our data are administrative, rather than survey based, we also use a relatively large sample, i.e., thousands of teachers, rather than hundreds. In addition, our study is distinct from prior work (outside of Tyler et al. (2009)) in that both sets of subjective evaluations we examine were made by professionals as part of their job, and one was a high-stakes evaluation. This is important to the extent that individuals change the way they do assessments in different contexts.

II. Data and Descriptive Statistics

Our analysis uses data on students and teachers in the public schools of New York City. First are administrative data on demographics, behavior, and achievement test scores in math and English for students in grades 3 to 8 in the school years 2003-04 through 2007-08. These data also link students to their math and English teacher(s). We also use data on teachers' characteristics: demographics, possession of a master's degree, type of certification/program, and teaching experience (as proxied by their position in the salary schedule).

Data on subjective evaluations come from two programs for new teachers in New York City. The first program is the New York City Teaching Fellows (TF), an alternative path to

[3] Most studies of subjective evaluations by different groups—principals, peer teachers, students, parents, and the teachers themselves—only examine correlations among these measures (e.g., Epstein (1985), Peterson (1987)). We know of two studies that examine the relation between multiple subjective evaluations and teacher effectiveness (Anderson (1954) and Wilkerson (2000)), but both are based on very small samples.

teaching certification through which about a third of new teachers in New York City are hired.[4] After submitting an application, approximately 60 percent of applicants are invited for a day-long interview process, which includes a mock teaching lesson, a written essay on a topic not given in advanced, a discussion with other candidates, and a personal interview.

Starting with applications for school year 2004-2005, applicants brought in for interviews have been rated on a 5-point scale.[5] In order to be accepted into the program, candidates must receive one of the top three evaluations; only about five percent of applicants receiving either of the two lowest evaluations are accepted into the program, based on a review by a committee that makes final recruitment decisions. Because very few candidates received the second-lowest evaluation (reserved for borderline cases), we combine Fellows receiving the two lowest evaluations into one group for our analysis. We use evaluations on TF applicants who began teaching in the school years 2004-2005 through 2006-2007.

The second source of subjective evaluations data is a program which provided mentoring to new teachers in New York City during the school years 2004-2005 through 2006-2007.[6] Under this centrally administered program, a group of trained, full-time mentors worked with new teachers over the course of their first year to improve their teaching skills. Typically, a

[4] Fellows are required to attend an intensive pre-service training program designed to prepare them to teach and to pursue a (subsidized) master's degree in education while teaching in a public school. Boyd et al. (2006) and Kane et al. (2008) provide more detailed descriptions and analyses of this program.

[5] The first evaluations on a 5 point scale were entered starting in November of 2003. Applicants that had already been interviewed in September and October were assigned a mark regarding acceptance or rejection and, sometimes, a designation of "top 20" or "borderline." We use these marks to recode these candidates under the 5 point scale in the following manner: "top 20" applicants are given the best evaluation, accepted candidates with no additional designation are given the second best evaluation, "borderline" accepted candidates are given the third best evaluation, "borderline" rejected applicants are given the second lowest evaluation, and rejected applicants with no additional designation are given the lowest evaluation. Personal correspondence with Teaching Fellows program administrators confirmed that these classifications are appropriate.

[6] See Rockoff (2008) for a detailed description and analysis of this program. Mentoring is required for all new teachers in New York State. The New York City mentoring program targeted all new teachers in school years 2004-2005 and 2005-2006, but in 2006-2007 it did not serve teachers at roughly 300 "empowerment" schools that were given greater autonomy (including control of how to conduct mentoring) in return for greater accountability. The mentoring program did not continue in the school year 2007-2008, when all principals were given greater autonomy.

mentor would meet with each teacher once every one or two weeks, starting sometime between late September and mid-October and extending through June.

As part of this program, mentors submitted ongoing evaluations of teachers' progress in mastering a detailed set of teaching standards. Mentors provided monthly summative evaluations and bimonthly formative evaluations of teachers on a five point scale.[7] Summative and formative evaluations are highly correlated (coefficient of correlation 0.85) and we therefore average them into a single measure of teacher effectiveness.

Mentors were unable to observe teachers prior to the start of the school year, so all of their evaluations may be partially affected by the students to whom teachers were assigned in their first year. Nevertheless, it is still interesting to ask whether mentors' impressions after only a few meetings with the teacher are predictive of performance in the first year. We therefore calculate mentors' evaluations of teachers using evaluations submitted up until November 15, in addition to all evaluations submitted in reference to teacher performance from March through June. The latter is only used to examine teacher effectiveness the following year.

[7] Formative evaluations were much more detailed than summative evaluations. Teachers were rated on six competencies: engaging and supporting all students in learning, creating and maintaining an effective environment for student learning, understanding and organizing subject matter for student learning, planning instruction and designing learning experiences for all students, assessing student learning, and developing as a professional educator. Moreover, each of these competencies had between 5 and 8 items. However, not all mentors rated teachers in all competencies, and, when they did, evaluations were highly correlated (and often identical) across competencies. Results of a simple factor analysis (available upon request) reveal that variation in evaluations for all competencies was mainly driven by a single underlying trait. Thus, we construct a single formative evaluation using the average of all non-missing subcategory evaluations. All evaluations were entered electronically into a centralized database. However, some mentors did not complete all evaluations for their teachers, and some evaluations were submitted quite late. We drop the two percent of evaluations that were submitted more than 60 days after the month to which they related. As one might expect, the distribution of evaluations changed considerably over the course of the school year. In the early months of the year, most teachers received the lowest evaluation, so the distribution is skewed with long right hand tail. By the end of the year, the distribution is more normally distributed; some teachers were still at the lowest stage and others had reached the top, but most were somewhere in the middle. Because evaluation data was not completed every month for every teacher, we account for variation in the timing of teachers' evaluations by normalizing evaluations by the month and year they were submitted.

The individuals working as evaluators (TF interviewers and mentors) had all been trained on a set of evaluation standards, but it is possible that some individuals were "tougher" in applying these standards than others. Fortunately, over the course of this period each TF interviewer saw dozens of applicants, and each mentor worked with roughly 15 teachers per year (some working for multiple years). In addition, interviewers were assigned randomly to TF applicants, and Rockoff (2008) shows that, conditional on a teacher's subject area, the pairing of mentors with new teachers appears quasi-random. We therefore examine specifications that separate variation in absolute evaluation levels from relative variation within evaluators. To do so, we measure the average of the evaluations given out by each mentor (TF interviewer) and include these averages in our regression specifications as additional covariates.

Because we are interested in how both subjective and objective evaluations relate to teacher effectiveness, we restrict the analysis to teachers who taught tested subjects (math and/or English) and grades (four to eight).[8] Table 1 provides descriptive statistics for teachers in these grades and subjects who received subjective evaluations; for comparison purposes, we also include statistics based on other teachers working in the same years, subjects, and grades throughout New York City. Not surprisingly, teachers with evaluations are considerably younger and less likely to have a master's degree than other teachers, and possess little teaching experience. They are also teaching students who are more likely to be Black or Hispanic and have lower prior test scores, reflecting the tendency for higher turnover (and thus more hiring) in schools serving these students. While Teaching Fellows and mentoring program are distinct,

[8] We also implement a few additional sample restrictions, following Kane et al. (2008). Specifically, we drop classrooms in any school-year cell for which less than 75 percent of the students were successfully matched to a teacher, classrooms with less than 7 or more than 45 students tested, classrooms with more than 25 percent special education students, and classrooms taught by teachers listed as working in multiple schools or who left mid-year.

there is considerable overlap between them: 27 percent of mentored teachers were Teaching Fellows and 90 percent of the Teaching Fellows received mentoring evaluations.

We present a second set of summary statistics in Table 2, grouping new teachers by their subjective evaluations. Mentored teachers are divided by tercile of their beginning-of-year evaluation and TF teachers by their evaluation score, combining the lowest two evaluations into a single group. The table also displays the p-values from a test for whether the mean values for each characteristic are statistically different across these groups.[9] These tests indicate that teachers receiving higher mentor evaluations are slightly less likely to teach Hispanic or Black students, and students receiving Free/Reduced Price Lunch; they also have slightly larger classes. More importantly, their students have substantially higher *prior* test scores than those taught by teachers that received lower evaluations.[10] Thus, if mentor evaluations are valid measures of teaching skills, then more highly skilled new teachers are, on average, being hired by schools with higher achieving students. In contrast, we find little systematic variation in the student characteristics for Fellows who received different evaluations during recruitment.

Since most Teaching Fellows also received mentor evaluations, we present the average mentor evaluations from the beginning and end of the first year by TF evaluation (bottom of Table 2). Interestingly, the relationship between the two evaluations at the start of the year is fairly weak. Fellows receiving initially unacceptable evaluations (i.e., the two lowest scores of the 5 point scale) received the lowest mentor evaluations on average, but Fellows with the third highest TF evaluations (i.e. on the border of acceptance into the program) received the highest average mentor evaluation. In contrast, the relationship between TF evaluations and mentor

[9] These tests are based on the results of student (teacher) level linear regressions of student (teacher) characteristics on group level indicator variables, allowing for clustering at the teacher (school) level.
[10] These differences are eliminated if we look at residuals of prior test scores from a regression of school fixed effects. Moving from raw to residual scores, the gaps between top and bottom tercile mentored teachers move from 0.15 to 0.03 in math and from 0.10 to 0.01 in English.

evaluations at the *end* of the first year are monotonic. It is also worth noting that Teaching Fellows received low evaluations by mentors on average, though there is little evidence Teaching Fellows are less effective than other new teachers (Kane et al. (2008)).

III. Methodology and Regression Estimates

Our main analysis is based on regressions of the following form:

$$(1)\quad A_{ikt} = \boldsymbol{\gamma Eval}_k + \boldsymbol{\beta X}_{it} + \boldsymbol{\lambda T}_{ikt} + \sum_{g,t} \pi_{gt} D^g_{it} + + \sum_z \pi_z D^z_{it} + \varepsilon_{ikt}$$

where A_{ikt} is the standardized achievement test score for student i taught by teacher k in year t, $Eval_k$ is a vector of (subjective and/or objective) evaluations of teacher effectiveness, X_{it} are student level control variables (including prior achievement), T_{ikt} are controls for teacher and classroom level characteristics, D^g_{it} is an indicator for whether student i is in grade g in year t, π_{gt} is a grade-year fixed effect, D^z_{it} is an indicator for whether student i attends a school located in zip code z in year t, π_z is a zip code fixed effect, and ε_{ikt} is an idiosyncratic error term.

We first examine the power of subjective evaluations made prior to hire or at the start of the school year to predict the achievement gains of students assigned to new teachers. Next, we estimate the predictive power of subjective and objective evaluations of teacher effectiveness based on teachers' performance in their first year (or during recruitment) to predict the achievement gains of the students whom they teach on the second year.[11]

The objective evaluations of teacher effectiveness are estimates of a teacher's impact on student test scores in their first year. The empirical Bayes' method we use follows closely that of

[11] A potential concern is that teachers who perform poorly in their first year are more likely to leave the teaching profession or be assigned to non-tested grades or subjects in their second year. We examine both types of attrition using regression analysis and find no evidence that teachers receiving lower TF evaluations or mentor evaluations were more likely to exit teaching or not be linked with students in the following year. These results (available upon request) support the idea that results from teachers' second years are not materially affected by endogenous attrition.

Kane et al. (2008), and is a fairly standard procedure in the estimation of a teacher's value-added. Our method differs only in that our estimates are taken from a series of regressions, rather than just one. Each regression uses two years of data—which are needed to implement the empirical Bayes estimator—and produces objective evaluations of performance for a single cohort of first-year teachers. For example, to estimate value-added for teachers who began their careers in school year 2005-2006, we use data from 2004-2005 and 2005-2006. In this way, we avoid using data from teachers' second years to evaluation their first-year performance.[12]

While we focus on new teachers with subjective evaluations, the regressions also include students taught by other teachers in the same school-year cells. These teachers are included by setting the evaluation(s) variable(s) to zero and including an indicator variable for missing evaluation(s). Using this larger pool of teachers allows us to gain grater precision in the coefficient estimates for zip-code fixed effects and controls for school and classroom characteristics. For ease of interpretation, mentor evaluations, student test scores, and teachers' first year value-added estimates have all been standardized to have mean zero and standard deviation of one. Standard errors are clustered at the teacher level.

Estimates of the power of subjective evaluations to predict student achievement in a teacher's first year are shown in Table 3. The coefficients on TF evaluations and mentor evaluations from the start of the school year for math achievement are both positive (0.015 and 0.016) and statistically significant (Columns 1 and 3).[13] Notably, if we add a control for the average evaluation given out by mentors, we find it has a negative significant coefficient,

[12] We lack value-added estimates on some teachers that received subjective evaluations and were linked to students in their second year of teaching, but were not linked their first year. To include these teachers in our analysis, we set their value-added estimates to zero and include a variable indicating that these teachers were missing an estimate.
[13] In separate regressions (available upon request) we replace the linear TF evaluation term with indicator variables for each evaluation score. The coefficients indicate a monotonic positive relationship between evaluations and student achievement, but the results are driven mostly by the top and bottom groups. The difference in student achievement on average between the middle two groups of teachers is quite small.

indicating important variation in how mentors applied evaluation standards (Column 4).[14]

Coefficients on both types of evaluations for reading achievement (Columns 8-11) are positive but quite small and statistically insignificant. It is important to note, however, that estimates of variance in teacher effectiveness are considerably smaller for English than math, both in New York City and elsewhere (Kane et al. (2008), Kane and Staiger (2008)). Thus, we lack sufficient power in our sample to identify effects in English of the same *proportional* magnitude as the effects we find for math.

We also estimate specifications that include both TF evaluations and mentor evaluations from the start of the school year, identifying their coefficients solely from variation across teachers with both types of evaluations. Inclusion of both sets of evaluations at the same time has little effect on either coefficient, consistent with the weak correlation between the two evaluations (Table 2). Interestingly, the coefficient on mentor evaluations from the start of the school year is considerably larger in English for this subsample of teachers (i.e., Teaching Fellows who receive mentoring services) than for all mentored teacher (0.03 vs. 0.005, Columns 13 and 11) and statistically significant. This change in the coefficient is driven by a stronger relationship between student achievement and mentor evaluations for Teaching Fellows; adding a control for whether a teacher is a Teaching Fellow does not materially change the coefficient on mentor evaluations in the regression that includes all mentored teachers.

We then proceed to examine student achievement in a teacher's second year of teaching. First, we show that the value-added estimates are highly significant predictors of the achievement of teacher's students in the second year (Table 4, Columns 1 and 7), with more

[14] Take, for example, a mentor working with 15 teachers. If one of these teachers received a one standard deviation increase in their evaluation, we would expect their students to score 0.02 standard deviations higher in math. However, if *all* of the teachers working with this mentor received a similar increase in their evaluations, we would expect test scores to rise by only 0.007 standard deviations.

variation in achievement predicted in math (0.09) than English (0.02).[15] These results are consistent with prior research (e.g., Gordon et al. 2006, Kane and Staiger 2008).

In both math and English, the relationships between TF evaluations from recruitment and student achievement in the second year are statistically insignificant (Table 4, Columns 2, 3, 8, 9). However, evaluations by mentors—as well as variation in evaluations within mentors—bear a substantial positive relationship with student achievement in teachers' second years. In math, mentors' evaluations both at the beginning and end of the school year have significant positive coefficients (0.032 and 0.054, respectively). Furthermore, the coefficients on these predictors remain significant (0.024 and 0.032, respectively) when we include both of them and the objective evaluation in the same regression. In English, the end of year mentor evaluation is a statistically significant predictor of student achievement in a teacher's second year with a coefficient (0.024) that is slightly larger than (and robust to the inclusion of) our objective evaluation of first-year performance.[16]

IV. Conclusion

Using detailed data on students and teachers in New York City, we find evidence that teachers who receive higher subjective evaluations either prior to hire or in their first year of teaching produce greater gains in achievement with their future students. Consistent with prior research, we also find evidence that teachers who produce greater test score gains in their first

[15] The coefficient for math is consistent with a stable value-added model, i.e., the standard deviation of value added in math for first year teachers is very close to the coefficient in the regression. For English, the coefficient is only half the size of the standard deviation in value added we estimate among first year teachers. We investigated this issue further and found that the decreased power of first year value added to predict second year value added drops in the school year 2005-2006, when the English test in New York State was moved from March to January and the format of the test changed in grades five, six, and seven.

[16] Notably, in all specifications, the coefficient on the average evaluation given out by mentors at the end of the school year is negative and statistically significant, indicating important variation in how mentors applied the teaching standards on which they were trained to evaluate teachers. Indeed, the magnitude of these coefficients suggests that variation in average evaluations across mentors bears little relationship with student achievement.

year also produce greater gains, on average, in their second year. More importantly, we find that—conditional on objective data on first year performance—subjective evaluations still present significant and meaningful information about a teacher's future success in raising student achievement.

Knowledge regarding the power of subjective evaluations and objective performance data has important implications for designing teacher evaluation systems, merit pay, and other polices whose goal is improving teacher quality and student achievement. All school districts evaluate teachers, but evaluation policies are not typically based in high quality empirical research and in many cases produce little differentiation among teachers (see Weisberg et al. 2009). Given the current era of increased accountability for schools and the research demonstrating the importance of teacher quality, it is likely that districts will begin to implement policies that put greater stress on teacher effectiveness.

As this process unfolds, policymakers will need to have a better understanding of the power and limitations of the measures they use in establishing incentives and accountability for teachers. Our results, and those of prior work, suggest that evaluation systems which incorporate both subjective measures made by trained professionals and objective job performance data have significant potential to help address the problem of low teacher quality. However, we also find that the application of standards can vary significantly across individuals responsible for making evaluations, and the implementation of any evaluation system should address this issue.

References

Aaronson, Daniel, Lisa Barrow, and William. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics*, 25(1): 95-135.

Anderson, Harold M. 1954. "A Study of Certain Criteria of Teaching Effectiveness." *Journal of Experimental Education*, 23(1): 41-71.

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 Corp.

Brookover, Wilbur B. 1945. "The Relation of Social Factors to Teaching Ability." *Journal of Experimental Education*, 13(4): 191-205.

Brophy, Jere, and Thomas L. Good. 1986. Teacher Behavior and Student Achievement. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*. 3rd ed., 238-375, New York: Simon and Schuster.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff. 2008. "The Narrowing Gap in New York City Teacher Qualifications and Its Implications for Student Achievement in High-Poverty Schools." *Journal of Policy Analysis and Management*, 27(4):793-818.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. "How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement." *Education Finance and Policy*, 1(2): 176-216.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. "How and Why Do Teacher Credentials Matter for Student Achievement?" NBER Working Paper 12828

Epstein, Joyce L. 1985. "A Question of Merit: Principals' and Parents' Evaluations of Teachers." *Educational Researcher*, 14(7): 3-10.

Gallagher, H. Alix. 2004. "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education*, 79(4): 79-107.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. *The Hamilton Project: Identifying Effective Teachers Using Performance on the Job*. Washington, DC: The Brookings Institution.

Gotham, R.E. 1945. "Personality and Teaching Efficiency." *Journal of Experimental Education*, 14(2): 157-165.

Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review, Papers and Proceedings* 61(2): 280-288.

Harris, Douglas N., and Tim R. Sass. 2006. "Value-Added Models and the Measurement of Teacher Quality." Unpublished Manuscript, Florida State University.

Harris, Douglas N., and Tim R. Sass. 2007. "What Makes for a Good Teacher and Who Can Tell?" Unpublished Manuscript, Florida State University.

Hill, C.W. 1921. "The Efficiency Ratings of Teachers." *The Elementary School Journal*, 21(6): 438-443.

Holtzapple, Elizabeth. 2003. "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System." *Journal of Personnel Evaluation in Education*, 17(3): 207-219.

Jacob, Brian A., and Lars J. Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Evaluation in Education." *Journal of Labor Economics*, 26(1): 101-136.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review*, 27(6): 615-631.

Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." NBER Working Paper 14607.

Kimball, Steven M., Brad White, Anthony T. Milanowski, and Geoffrey Borman. 2004. "Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of Education*, 79(4): 54-78.

Manatt, Richard P. and Bruce Daniels. 1990. "Relationships Between Principals' Ratings of Teacher Performance and Student Achievement." *Journal of Personnel Evaluation in Education* 4(2): 189-201.

Medley, Donald M., and Homer Coker. 1987. "The Accuracy of Principals' Judgments of Teacher Performance." *Journal of Educational Research*, 80(4): 242-247.

Milanowski, Anthony. 2004. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence From Cincinnati." *Peabody Journal of Education*, 79(4): 33-53.

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

Peterson, Kenneth D. 1987. "Teacher Evaluation with Multiple and Variable Lines of Evidence." *American Educational Research Journal*, 24(2): 311-317.

Rivkin, Steven G., Eric A. Hanushek, and John Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–458.

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

Rockoff, Jonah E. 2008. "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City." NBER Working Paper 13868.

Rockoff, Jonah E., Douglas O. Staiger, Eric Taylor, and Thomas J. Kane. 2009. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." Unpublished Manuscript, Columbia University.

Sanders, William L., and June C. Rivers. 1996. *Cumulative and Residual Effects of Teachers on Future Student Academic Achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Schacter, John, and Yeow M. Thum. 2004. "Paying for High- and Low- Quality Teaching." *Economics of Education Review*, 23(4): 411-440.

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. 2009. "Using Student Performance Data to Identify Effective Classroom Practices." Draft Working Paper. Providence, R.I.: Brown University, and Cambridge, Mass.: Harvard University.

Weingarten, Randi. 2007. "Using Student Test Scores to Evaluate Teachers: Common Sense or Nonsense?" *New York Times* Advertisement, March 2007.

Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. *The Widget Effect*. Brooklyn, NY: The New Teacher Project.

Wilkerson, David J., Richard P. Manatt , Mary A. Rogers , and Ron Maughan. 2000. "Validation of Student, Principal, and Self-Ratings in 360° Feedback® for Teacher Evaluation." *Journal of Personnel Evaluation in Education*, 14(2): 179-192.

Table 1: Descriptive Statistics by Teacher Program

	Mentored Teachers	Teaching Fellows	Other NYC Teachers
Number of Teachers in Analysis Sample	3,198	1,023	17,777
Teacher characteristics			
Teaching Fellow	27%	100%	*n/a*
Received Mentoring	100%	90%	*n/a*
Age	29.5	30.3	39.9
Years of Teaching Experience	0.53	0.39	4.67
Has Master Degree	36%	21%	76%
Student characteristics			
White	10%	7%	15%
Hispanic	45%	49%	38%
Black	34%	36%	32%
English Language Learner	10%	11%	8%
Receives Free/Reduced Price Lunch	71%	74%	65%
Prior Math Test Score (standardized)	0.03	-0.03	0.19
Prior English Test Score (standardized)	0.01	-0.04	0.17

Notes: Student characteristics for evaluated teachers (mentored or teaching fellow) are based on classrooms linked to them in their first year of teaching. For a small number of teachers, first year classroom data is not available and second year data is used. Teachers' characteristics are from their first year teaching. Statistics for "Other NYC Teachers" are based on all other teachers working during the school years 2004-2005 through 2007-2008.

Table 2: Descriptive Statistics by Evaluation

	Mentor Evaluation, *Sept-Nov, N(0,1)*				**Teaching Fellow Evaluation, *During Recruitment***				
	Bottom Tercile	Middle Tercile	Top Tercile	P-value	4/5 (Bottom)	3	2	1 (Top)	P-value
Student Characteristics									
White	8%	7%	13%	0.00	8%	8%	6%	6%	0.54
Hispanic	48%	48%	44%	0.02	49%	44%	51%	52%	0.01
Black	35%	36%	31%	0.05	35%	40%	34%	35%	0.05
English Language Learner	11%	10%	10%	0.97	11%	11%	10%	11%	0.84
Free/Reduced Price Lunch	73%	73%	70%	0.03	74%	75%	73%	75%	0.50
Special Education	0.2%	0.3%	0.3%	0.60	0.4%	0.3%	0.4%	0.4%	0.29
Class size	25.9	26.1	26.6	0.05	26.6	26.3	26.6	26.5	0.76
Prior Math Test Score, *N(0,1)*	-0.05	-0.01	0.10	0.00	-0.07	-0.04	-0.02	-0.01	0.83
Prior English Test Score, *N(0,1)*	-0.03	-0.02	0.07	0.02	-0.04	-0.03	-0.05	-0.04	0.89
Teacher Characteristics									
Years of Teaching Experience	0.44	0.36	0.54	0.02	0.18	0.45	0.41	0.36	0.77
Has Master Degree	36%	35%	38%	0.22	10%	19%	23%	24%	0.31
Mentor Evaluation, *Sept-Nov, N(0,1)*					-0.32	0.01	-0.09	-0.15	
Mentor Evaluation, *Mar-Jun, N(0,1)*					-0.27	-0.06	-0.05	0.01	

Notes : Student characteristics are based on students (grade 4 to 8) in classrooms with an evaluated teacher (mentored or teaching fellow) during their first year of teaching. For a small number of evaluated teachers, first year classroom data is not available and second year data is used. Teachers' characteristics are from their first year teaching. The p-value corresponds to a test that group level indicator variables are significant predictors of the student (teacher) charateristic in a student (teacher) level linear regression that allows for clustering at the teacher (school) level.

Table 3: Subjective Evaluations and Student Achievement in a Teacher's First Year

	Teaching Fellows		Mentored Teachers		TF & Mentored		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math							
TF Evaluation, *4 point scale*	0.015	0.011			0.016		0.016
	(0.008)*	(0.008)			(0.009)*		(0.009)*
TF Interviewer Average Evaluation		0.016			0.008		0.008
		(0.011)			(0.011)		(0.011)
Mentor Evaluation, *Sept-Nov, N(0,1)*			0.016	0.021		0.021	0.018
			(0.008)*	(0.009)**		(0.018)	(0.017)
Mentor Average Evaluation				-0.014		-0.012	-0.011
				(0.008)*		(0.014)	(0.014)
Observations	399,982	399,982	399,982	399,982	399,982	399,982	399,982
Teachers	8,287	8,287	8,287	8,287	8,287	8,287	8,287
Teachers with Evaluations	529	529	1,868	1,868	477	477	477
R^2	0.67	0.67	0.67	0.67	0.67	0.67	0.67
	Teaching Fellows		Mentored Teachers		TF & Mentored		
English	(8)	(9)	(10)	(11)	(12)	(13)	(14)
TF Evaluation, *4 point scale*	0.007	0.009			0.008		0.010
	(0.009)	(0.009)			(0.010)		(0.010)
TF Interviewer Average Evaluation		-0.002			-0.007		-0.011
		(0.011)			(0.011)		(0.011)
Mentor Evaluation, *Sept-Nov, N(0,1)*			0.005	0.005		0.030	0.032
			(0.006)	(0.007)		(0.014)**	(0.014)**
Mentor Average Evaluation				-0.000		0.001	0.003
				(0.007)		(0.015)	(0.015)
Observations	351,494	351,494	351,494	351,494	351,494	351,494	351,494
Teachers	8,311	8,311	8,311	8,311	8,311	8,311	8,311
Teachers with Evaluations	431	431	1,885	1,885	398	398	398
R^2	0.62	0.62	0.62	0.62	0.62	0.62	0.62

Notes: Standard errors (in parentheses) are clustered at the teacher level. All regressions control for students' sex, race, cubic polynomials in previous test scores, prior suspensions and absences, and indicators for English Language Learner, Special Education, grade retention, and free or reduced price lunch status. These controls are also interacted with grade level. The regressions also control for teacher experience (indicators for each year up to six years of experience and an indicator for seven or more years of experience), classroom and school-year demographic averages of student characteristics, and class size. In addition, all regressions include year, grade, year-grade, and zip-code fixed effects. * significant at 10%, ** significant at 5%.

Table 4: Subjective and Objective Evaluations and Student Achievement in a Teacher's Second Year

Math	All Second Year Teachers	Teaching Fellows		Mentored Teachers		
	(1)	(2)	(3)	(4)	(5)	(6)
Objective Evaluation *(Value Added Year 1)*	0.088		0.095			0.085
	(0.006)**		(0.010)**			(0.006)**
TF Evaluation, *4 point scale*		0.009	0.005			
		(0.012)	(0.010)			
TF Interviewer Average Evaluation		-0.005	-0.004			
		(0.012)	(0.011)			
Mentor Evaluation, *Sept-Nov, N(0,1)*				0.032		0.024
				(0.009)**		(0.008)**
Mentor Average Evaluation *(Sept-Nov)*				-0.018		-0.014
				(0.011)*		(0.011)
Mentor Evaluation, *Mar-Jun, N(0,1)*					0.054	0.032
					(0.009)**	(0.008)**
Mentor Average Evaluation *(Mar-Jun)*					-0.052	-0.031
					(0.012)**	(0.011)**
Observations	389,530	389,530	389,530	389,530	389,530	389,530
Teachers	7,678	7,678	7,678	7,678	7,678	7,678
Teachers with Evaluations	1,821	501	501	1,755	1,755	1,755
R^2	0.67	0.67	0.67	0.67	0.67	0.67
English	All Second Year Teachers	Teaching Fellows		Mentored Teachers		
	(7)	(8)	(9)	(10)	(11)	(12)
Objective Evaluation *(Value Added Year 1)*	0.020		0.015			0.019
	(0.005)**		(0.009)*			(0.005)**
TF Evaluation, *4 point scale*		0.001	-0.001			
		(0.009)	(0.009)			
TF Interviewer Average Evaluation		0.001	0.001			
		(0.010)	(0.011)			
Mentor Evaluation, *Sept-Nov, N(0,1)*				0.007		0.001
				(0.007)		(0.007)
Mentor Average Evaluation *(Sept-Nov)*				-0.010		-0.004
				(0.008)		(0.009)
Mentor Evaluation, *Mar-Jun, N(0,1)*					0.024	0.021
					(0.006)**	(0.006)**
Mentor Average Evaluation *(Mar-Jun)*					-0.032	-0.029
					(0.008)**	(0.009)**
Observations	340,428	340,428	340,428	340,428	340,428	340,428
Teachers	7,805	7,805	7,805	7,805	7,805	7,805
Teachers with Evaluations	1,796	405	405	1,743	1,743	1,743
R^2	0.61	0.61	0.61	0.61	0.61	0.61

Notes: Standard errors (in parentheses) are clustered at the teacher level. All regressions control for students' sex, race, cubic polynomials in previous test scores, prior suspensions and absences, and indicators for English Language Learner, Special Education, grade retention, and free or reduced price lunch status. These controls are also interacted with grade level. The regressions also control for teacher experience (indicators for each year up to six years of experience and an indicator for seven or more years of experience), classroom and school-year demographic averages of student characteristics, and class size. In addition, all regressions include year, grade, year-grade, and zip-code fixed effects. * significant at 10%, ** significant at 5%.