Labour Economics 18 (2011) 687–696

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

Subjective and objective evaluations of teacher effectiveness: Evidence from New York City ☆

Jonah E. Rockoff [a,*], Cecilia Speroni [b]

[a] Columbia Business School and NBER, United States
[b] Teachers College, Columbia University, United States

ARTICLE INFO

Article history:
Received 2 April 2010
Received in revised form 14 February 2011
Accepted 19 February 2011
Available online 21 March 2011

Keywords:
Teachers
Employee evaluation

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. Using data from New York City public schools, we estimate whether subjective evaluations of teacher effectiveness have predictive power for the achievement gains made by teachers' future students. We find that these subjective evaluations have substantial power, comparable with and complementary to objective measures of teacher effectiveness taken from a teacher's first year in the classroom.

> "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 Cantrell 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 two sources: 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 evaluations of new teachers' effectiveness throughout the school year. The objective evaluations of effectiveness we use 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 teachers' future students. Moreover, 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.

As we explain below, we design our analysis to avoid a mechanical relation between our objective measure of teacher effectiveness – value-added from a teacher's first year – and future performance. However, the result that an objective measure predicts a similar future objective measure is not surprising. Indeed, if the value-added measure were a biased measure of the teacher's contribution, it may still predict a future, similarly biased measure of that contribution. In this sense, the positive correlation between subjective and objective measures and the finding that they both predict future performance increases our confidence in each measure.

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, for evaluations by mentors, variation in evaluations *within* evaluators is a much stronger predictor of teacher

☆ We thank Brian Jacob and participants at the American Economics Association Meetings for their very helpful comments. Jonah Rockoff thanks the Smith Richardson Foundation for financial assistance and Lesley Turner for her work on the subjective evaluations data.

* Corresponding author.
E-mail addresses: jonah.rockoff@columbia.edu (J.E. Rockoff), cs2456@columbia.edu (C. Speroni).

0927-5371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.labeco.2011.02.004

effectiveness than variation *between* evaluators. This highlights the importance of reliability in the procedures used to generate subjective evaluations.

The paper proceeds as follows. We provide a brief summary of previous literature in Section 1 and describe our data in Section 2. Our methodology and empirical estimates are presented in Section 3, and Section 4 concludes.

1. 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. Goldhaber and Hansen (2009) draw similar conclusions in their analysis of data from North Carolina.

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 with students that are more or less difficult to teach in ways that administrative dataset 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 any system of teacher evaluation purely based on student test score data would ever be implemented, given considerable opposition from teachers' unions (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 power to predict student achievement. For example, Jacob and Lefgren (2008) find that a one standard deviation increase in a principal's evaluation of a teacher is associated with higher test score performance of 0.10 and 0.05 standard deviations in math and English, respectively.[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 performance data).[3] Second, it is unclear the extent to which principal evaluations represent a subjective assessment of teacher effectiveness or whether they are influenced by objective data on the performance of a teacher's previous students.

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.[4] 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.[5]

2. 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).

Using the linked student–teacher data, we can objectively evaluate teachers' impacts on student test scores in their first year using an empirical Bayes' method. This estimation of a teacher's value-added is a fairly standard procedure and follows closely the method described in Kane et al. (2008). However, rather than obtain a single estimate of teacher value-added using all years of data, we run a series of regressions, each of which uses two years of data, and the residuals from each regression are used to produce estimates for a single cohort of first-year teachers (e.g., data from 2004–05 and 2005–06 are used to estimate value-added for teachers who began their careers in school year 2005–06). This avoids using data from teachers' second years to evaluate their first-year performance.[6]

[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, 2009; Jacob and Lefgren, 2008; Rockoff et al., 2010).

[2] Another related set of studies focus on teachers who are certified by the National Board of Professional Teaching Standards (NBPTS) via review of a portfolio which includes student work, a self-assessment, and sample video of classroom instruction (e.g., Cavalluzzo, 2004; Goldhaber and Anthony, 2007; Cantrell et al. (2007); Harris and Sass, 2007). The evidence, while mixed, generally suggests that NBPTS selects more effective teachers among its applicants and that teachers certified by NBPTS are more effective than teachers who lack this certification.

[3] We view this as a stronger test because we are interested in the question of whether subjective evaluations can identify persistent differences in the impact of teachers on student achievement. This has direct implications for policies such as the awarding of tenure. However, variation in actual performance within teachers over time could also cause the contemporaneous relationship between evaluations and performance to be stronger than the non-contemporaneous relationship. If one were interested in awarding an annual performance bonus then the contemporaneous relationship would clearly be of great interest.

[4] 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 et al. (2000)), but both are based on very small samples.

[5] Because we analyze real teacher (candidate) evaluations done by professionals as part of their jobs, we also take the cardinal variation of these subjective measures largely as given. Our results should therefore be interpreted in this context—we report the relationship of student achievement to variation in evaluations made under particular schemes by particular individuals, and the magnitudes of our estimates may not generalize to other contexts.

[6] 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.

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 teaching certification taken by about a third of new teachers in New York City.[7] After submitting an application, approximately 60% 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 advance, 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.[8] 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.[9] 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 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.[10] Summative and formative evaluations are highly correlated (coefficient of correlation 0.84) and we therefore average them into a single measure of teacher effectiveness. 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 were not completed every month for every teacher, we account the timing of teachers' evaluations by normalizing evaluations by the month and year they were submitted.

Mentors could not observe teachers prior to the start of the school year, and their evaluations may be 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. We use evaluations submitted from March through June to examine effectiveness in teachers' second years.

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). 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

Table 1
Descriptive statistics by teacher program.

	Mentored teachers	Teaching Fellows	Other NYC teachers
Number of teachers	3181	1003	14,820
Teacher characteristics			
Teaching Fellow	27%	100%	*n/a*
Received mentoring	100%	90%	*n/a*
Age	29.5	30.3	39.6
Years of teaching experience	0.53	0.39	4.90
Has master degree	36%	21%	79%
Student characteristics			
Black or Hispanic	79%	85%	69%
English language learner	10%	10%	9%
Receives free/Reduced price lunch	71%	75%	67%
Prior math test score (standardized)	0.03	−0.03	0.20
Prior English test score (standardized)	0.01	−0.05	0.18

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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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									
Black or Hispanic	83%	84%	75%	0.00	84%	86%	85%	87%	0.59
English language learner	11%	10%	10%	0.97	11%	11%	10%	10%	0.86
Free/Reduced price lunch	73%	73%	70%	0.03	74%	76%	73%	76%	0.29
Special education	0.2%	0.3%	0.3%	0.60	0.4%	0.2%	0.4%	0.3%	0.20
Class size	25.9	26.1	26.6	0.05	26.6	26.3	26.5	26.5	0.72
Prior math test score, *N(0,1)*	−0.05	−0.01	0.10	0.00	−0.07	−0.06	−0.02	−0.01	0.53
Prior English test score, *N(0,1)*	−0.03	−0.02	0.07	0.02	−0.04	−0.05	−0.05	−0.04	0.95
Teacher characteristics									
Years of teaching experience	0.44	0.36	0.54	0.01	0.19	0.43	0.42	0.38	0.87
Has master degree	36%	35%	38%	0.22	10%	19%	23%	24%	0.34
Mentor evaluation, *Sept–Nov, N(0,1)*					−0.32	0.01	−0.09	−0.15	0.44
Mentor evaluation, *Mar–Jun, N(0,1)*					−0.27	−0.08	−0.05	0.01	0.62

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.

New York City. Teachers with evaluations are new and, not surprisingly, considerably younger, less experienced, and less likely to have a master's degree than other teachers. 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 the mentoring program are distinct, there is considerable overlap between them: 27% of mentored teachers were Teaching Fellows and 90% of the Teaching Fellows received evaluations via the mentoring program. However, the descriptive statistics of their students suggest that Teaching Fellows are hired to work in relatively disadvantaged schools, as has been documented in prior studies (Boyd et al., 2008; Kane et al., 2008).

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.[11] While we find very little systematic variation in student characteristics for Fellows who received different evaluations during recruitment, we do find that teachers receiving high evaluations by their mentors at the beginning of the school year are less likely to teach minority 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. If mentor evaluations are valid measures of teaching skills, this suggests that more highly skilled new teachers may be more likely to be hired by schools with higher achieving students. Alternatively, mentors' initial impressions of teacher effectiveness may simply be influenced by factors correlated with students' incoming achievement. This underscores the importance of rigorously testing whether subjective evaluations are related to future student outcomes and the performance of teachers with new groups of students.

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 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).

3. Methodology and regression estimates

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

$$A_{ikt} = \gamma Eval_k + \beta X_{it} + \lambda T_{ikt} + \sum_{g,t} \pi_{gt} D_{it}^{g} + \sum_{z} \pi_z D_{it}^{z} + \varepsilon_{ikt} \qquad (1)$$

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_{it}^{g} is an indicator for whether student i is in grade g in year t, D_{it}^{z} is an indicator for whether student i attends a school located in zip code z in year t, π_{gt} and π_z are grade-year and zip-code fixed effects, and ε_{ikt} is an idiosyncratic error term. We also present results from other specifications, e.g. using school fixed effects, as robustness checks. To gain precision on estimates of fixed effects and other coefficients, the regressions include students taught by other teachers in the same schools, though below we show that our results are similar if we limit the sample to only those teachers who were evaluated. For teachers working alongside those being evaluated, we set their evaluation(s) to zero and we include an indicator variable for missing evaluation(s). For ease of interpretation, we normalize evaluations across all mentored teachers – including those not teaching math or English in grades 4 to 8 – to have mean zero and standard deviation of one, and student test scores are also similarly normalized at the year-grade level.[12] Standard errors are clustered at the teacher level.

[11] 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.

[12] For teachers of math and English in grades 4 to 8, the mean and standard deviation are also quite close to zero and one. Among those teaching math, the mean is 0.03 with a standard deviation of 1.01. For those teaching English, the mean is 0.05 with a standard deviation of 1.03.

Table 3
Subjective evaluations and student achievement in a teacher's first year.

Math	Teaching Fellows		Mentored teachers		Teaching Fellow & Mentored		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
TF Interview evaluation, 4 point scale	0.015 (0.008)*				0.020 (0.008)**		0.019 (0.009)**
Deviation from TF interviewer avg. evaluation		0.016 (0.009)*					
TF Interviewer average evaluation		0.004 (0.025)					
Mentor evaluation, Sept–Nov, N(0,1)			0.016 (0.008)*				
Deviation from mentor's average evaluation				0.022 (0.009)**		0.022 (0.018)	0.020 (0.018)
Mentor average evaluation				0.007 (0.010)		0.006 (0.016)	0.006 (0.016)
Average and deviation coeffs equal (P-value)		0.66		0.08		0.28	0.32
Observations	398,118	398,118	398,118	398,118	398,118	398,118	398,118
Teachers	8260	8260	8260	8260	8260	8260	8260
Teachers with evaluations	516	516	1857	1857	466	466	466
R^2	0.67	0.67	0.67	0.67	0.67	0.67	0.67
English	Teaching Fellows		Mentored teachers		Teaching Fellow & Mentored		
	(8)	(9)	(10)	(11)	(12)	(13)	(14)
TF interview evaluation, 4 point scale	0.007 (0.009)				0.006 (0.010)		0.007 (0.010)
Deviation from TF interviewer avg. evaluation		0.005 (0.009)					
TF interviewer average evaluation		0.024 (0.031)					
Mentor evaluation, Sept–Nov, N(0,1)			0.005 (0.006)				
Deviation from mentor's average evaluation				0.006 (0.007)		0.029 (0.014)**	0.029 (0.014)**
Mentor average evaluation				0.005 (0.007)		0.029 (0.015)**	0.030 (0.015)**
Average and deviation coeffs equal (P-value)		0.55		0.93		0.99	0.92
Observations	351,604	351,604	351,604	351,604	351,604	351,604	351,604
Teachers	8304	8304	8304	8304	8304	8304	8304
Teachers with evaluations	425	425	1879	1879	392	392	392
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. Mentor evaluations from the beginning of the year include only those submitted on or before November 15 of each school year. All regressions control for students' sex, race, cubic polynomials in the previous test scores, prior suspensions and absences, and indicators for English language learner, Special Education, grade retention, and free or reduced price lunch status; each also interacted with grade level. Teacher experience is included as 7 dummies for years of experience and an indicator for missing experience. Class level and school-year demographics consist on averages for race, cubic polynomials on prior math and English scores, absences, English language learner, Special Education, free or reduced price lunch, and class size. In addition, all regressions include year, grade, year–grade, and zip-code fixed effects.
** significant at 5%.
* significant at 10%.

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] The coefficients for regressions of achievement in English (Columns 8 and 10) are positive but 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.

To explore whether the same standards were applied by all evaluators, we test whether variation in evaluations *within* evaluators is a stronger predictor of teacher effectiveness than variation *between* evaluators. If differences in average evaluations across mentors or TF interviewers simply reflected sampling variation in the effectiveness of teachers assigned to them, the coefficient on the average evaluation they give out (variation *between* evaluators) should be equal to the coefficient on the difference between this average and a teacher's own evaluation (variation *within* evaluators). In contrast, if evaluators that gave out higher (lower) average evaluations were simply more lenient (harsh) in applying standards, the *between* coefficient should be smaller than the *within* coefficient.

For evaluations by TF interviewers, we cannot reject that the two coefficients are the same in either subject. In math, the coefficient on the TF interviewer's average evaluation is much smaller than the coefficient on the deviation from that average (Column 2), but in English the coefficient on the TF interviewer's average evaluation is larger than the coefficient on the deviation (Column 9). This is perhaps not surprising, since TF interviewers were given substantial training in order to standardize their evaluations. However, for evaluations by mentors, who were not given such training, we do find evidence of varying standards. In math, the coefficient on the mentor's average evaluation (0.007) is not significantly different than zero and is significantly lower (p-value 0.08) than the coefficient on the deviation of a teacher's evaluation from the mentor average (0.022). In English, the coefficient on the average evaluation is slightly smaller, but not statistically different.

[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 between the middle two groups of teachers is, on average, quite small.

Table 4
Subjective and objective evaluations and student achievement in a teacher's second year.

Math	All teachers	Teaching Fellows		Mentored teachers		
	(1)	(2)	(3)	(4)	(5)	(6)
Objective evaluation (value added year 1)	0.088		0.094			0.085
	(0.006)**		(0.010)**			(0.007)**
TF evaluation, 4 point scale		0.008	0.005			
		(0.012)	(0.010)			
Deviation from mentor's average evaluation				0.033		0.024
Sept–Nov, N(0,1)				(0.009)**		(0.008)**
Mentor's average evaluation				0.014		0.011
Sept–Nov, N(0,1)				(0.011)		(0.010)
Deviation from mentor's average evaluation					0.054	0.031
Mar–Jun, N(0,1)					(0.009)**	(0.008)**
Mentor's average evaluation					0.002	0.000
Mar–Jun, N(0,1)					(0.008)	(0.008)
Average and deviation coeffs equal, Sept–Nov, (P-value)				0.09		0.24
Average and deviation coeffs equal, Mar–Jun, (P-value)					0.00	0.01
Observations	387,916	387,916	387,916	387,916	387,916	387,916
Teachers	7660	7660	7660	7660	7660	7660
Teachers with evaluations	1812	492	492	1747	1747	1747
R^2	0.67	0.67	0.67	0.67	0.67	0.67
English	All teachers	Teaching Fellows		Mentored teachers		
	(7)	(8)	(9)	(10)	(11)	(12)
Objective evaluation (value added year 1)	0.018		0.015			0.018
	(0.004)**		(0.009)*			(0.004)**
TF evaluation, 4 point scale		0.003	0.001			
		(0.009)	(0.009)			
Deviation from mentor's average evaluation				0.007		0.001
Sept–Nov, N(0,1)				(0.007)		(0.007)
Mentor's average evaluation				−0.004		−0.004
Sept–Nov, N(0,1)				(0.008)		(0.009)
Deviation from mentor's average evaluation					0.023	0.020
Mar–Jun, N(0,1)					(0.006)**	(0.006)**
Mentor's average evaluation					−0.009	−0.008
Mar–Jun, N(0,1)					(0.005)	(0.006)
Average and deviation coeffs equal, Sept–Nov, (P-value)				0.18		0.60
Average and deviation coeffs equal, Mar–Jun, (P-value)					0.00	0.00
Observations	340,297	340,297	340,297	340,297	340,297	340,297
Teachers	7803	7803	7803	7803	7803	7803
Teachers with evaluations	1789	398	398	1737	1737	1737
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 5%.
* significant at 10%.

In a final set of specifications examining teachers in their first year, we include both TF evaluations and mentor evaluations, including only teachers with both types of evaluations (Columns 5 to 7 and 12 to 14). Motivated by our findings above, we continue to split mentor evaluations into mentor average and deviation from average. In math, the results are quite similar for this sample. However, in English, we find that variation in evaluations both within and between mentors has significant predictive power for students' test scores. The change in the coefficients across the two samples (from about 0.005 to 0.03) 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.[14] The coefficients on both sets of evaluations are similar whether we estimate them in separate regressions or at the same time, consistent with the weak correlation between them.

We then proceed to examine student achievement in the second year of teachers' careers, which we believe provides a more rigorous test of whether objective and subjective performance metrics provide useful information for decisions such as teacher retention. Consistent with prior research (e.g., Gordon et al., 2006; Kane and Staiger, 2008), first-year value-added estimates are significant predictors of student achievement in the teacher's second year (Table 4, Columns 1 and 7), conditional on prior achievement and other controls.[15]

In both math and English, the relationships between TF evaluations from recruitment and student achievement in the second year are positive but statistically insignificant (Table 4, Columns 2, 3, 8, 9). However,

[14] One concern might be that the relationship between mentor evaluations and English achievement may be different for more disadvantaged student populations, whom Teaching Fellows tend to teach. To examine this we estimated specifications where we drop any school that did not hire any Teaching Fellows during our sample period; we find slightly smaller (but still significant) coefficients for mentor evaluations in math and slightly larger (but still insignificant) coefficients in English.

[15] The coefficient for math (0.09) 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 regression coefficient. For English, the coefficient (0.02) is 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 test format changed in grades five, six, and seven.

evaluations by mentors – and in particular variation in evaluations *within* mentors – bear a substantial positive relationship with student achievement in teachers' second years. In math, the *within* variation in mentors' evaluations both at the beginning and end of the school year have significant positive coefficients (0.033 and 0.054, respectively) and in both cases we can reject that the coefficient on mentors' average evaluations is equally large. Furthermore, the coefficients on these predictors remain significant (0.024 and 0.031, respectively) when we include both of them and the objective evaluation in the same regression. In English, within-mentor variation in the end of year evaluation is a statistically significant predictor of student achievement in a teacher's second year with a coefficient (0.023) that is slightly larger than (and robust to the inclusion of) our objective evaluation of first-year performance.[16] Also, we can reject that the *within* and *between* coefficients on end-of-year evaluations are the same.

Our main findings from the regressions shown in Table 4 are still subject to a number of concerns which we try to address here. First, teachers who perform poorly in their first year may be 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 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 our main findings are not materially affected by endogenous attrition.

Second, our results with respect to mentor evaluations may be driven by the non-random sorting of teachers to students. In particular, mentors directly observe student characteristics that may not be captured by our administrative data, and may rate teachers more highly in the first year if they are assigned with students who are "better" in ways we cannot observe. If this type of "selection on unobservables" is persistent into teachers' second year, this will in turn affect our estimates.[17] We present a series of robustness checks to assess the importance of this issue.

For purposes of comparison, we present estimates of the power of mentor evaluations to predict student achievement in teachers' second year using our original specification in Table 5, Columns 1 and 7. Adding a control for the average achievement of students assigned to the teacher in the first year (Columns 2 and 8) barely changes the coefficients. Restricting the sample to only those teachers receiving evaluations (Columns 3 and 9) also produces very similar results, as does replacing zip-code fixed effects with either mentor fixed effects (Columns 4 and 10) or school fixed effects (Columns 5 and 11).

Finally, we run a specification dropping all control variables and assess whether our results could be driven by selection on unobservables in the spirit of Altonji et al. (2005).[18] Not surprisingly, the omission of control variables increases the coefficients on the evaluation variables – both on the deviations from mentor averages and the average evaluations – confirming that teachers who received better evaluations generally taught higher achieving students. However, selection on unobservables would have to be unusually strong to explain our findings. In math, selection on unobservables would have to be 3.8 times (1.5 times) as strong as selection on observables to drive our results for evaluations made at the beginning (end) of teachers' first years. In English, where our initial results were weaker, our results are less robust; selection on unobservables would have to be 60% as strong as selection on observables to drive the coefficient on end-of-year evaluations. Nevertheless, recent evidence on teacher sorting (Jackson, 2009) suggests that evaluators' perceptions of effective teaching in classrooms serving higher achieving children may reflect real differences in teacher quality, not bias, and our conditional estimates may therefore understate the ability of mentors to observe good teaching.

While ours is the only paper to examine evaluations at the hiring stage (for Teaching Fellows) or by mentors, it is not the first to examine subjective evaluations more broadly. To help gauge the magnitude of our findings, we compare our estimates with those from Jacob and Lefgren (2008) and Harris and Sass (2009), who examine how teacher value-added estimates relate to survey-based evaluations of teachers by school principals. Of course, our studies differ in a number of respects, including the setting (New York vs. small cities in the mid-west and Florida), the teachers (only rookies vs. all teachers), the evaluators (mentors vs. principals), and the data collection process (administrative records vs. a research survey). Jacob and Lefgren find that a one standard deviation higher evaluation is associated with a rise in students' end of year math and reading test scores of 0.10 and 0.05 standard deviations, respectively. In Harris and Sass, the analogous estimates are 0.04 standard deviations for both subjects. We compare these results with our estimates based on variation in evaluation within mentors, since both studies of principals normalize evaluations at the principal level. We estimated that teachers scoring one standard deviation higher on the spring mentor evaluation are expected to raise math and English test scores by 0.054 and 0.023 standard deviations, respectively, the following year (Table 4). Thus, our estimates of the power of mentor evaluations to predict future student performance with particular teachers are similar to those found by Harris and Sass, but somewhat smaller than those found by Jacob and Lefgren, for principal evaluations.[19]

3.1. Interaction of subjective and objective evaluations and the impact of evaluator experience

In this subsection we present extensions to our main results. First, we examine whether subjective and objective evaluations have important interactions. In other words, do subjective evaluations have more power to distinguish effective and ineffective teachers for groups of teachers at different parts of the objective evaluation distribution, and vice versa? To explore this possibility, we run regressions where we include an interaction of a teacher's objective (subjective) evaluation with indicator variables for the tercile of a teacher's subjective (objective) evaluation. We focus on evaluations by mentors made at the end of a teacher's first year, since these were found to have the most consistent predictive power for future student outcomes.

The results (Table 6) indicate that objective evaluations are equally predictive of student achievement in the second year for teachers with subjective evaluations in each tercile. In contrast, the coefficient on the interaction of this subjective evaluation and the middle tercile indicator is larger than interactions with bottom and top tercile for both math and

[16] Notably, in all specifications, the coefficient on the average evaluation given out by mentors at the end of the school year is close to zero and statistically insignificant, suggesting considerable variation in how mentors applied the standards on which they were trained to evaluate teachers.

[17] There is an ongoing debate surrounding the importance of selection on unobservables in estimating teacher effects (see Rothstein, 2009; Kane and Staiger, 2008; Goldhaber and Hansen. 2009; Koedel and Betts, 2011). While this is somewhat beyond the scope of our paper, a number of our robustness checks address the issue of sorting.

[18] We start with the premise that mentors cannot observe test measurement error, which constitutes roughly 15% of the variance in test scores according to both New York State test publishers and an analysis by Boyd et al. (2008). Under this assumption, including factors which are unobservable to us but observable to mentors in our regression could, at most, produce an R^2 of 0.85. Using this upper bound on R^2, the coefficient estimates with controls (β_{with}) and without controls ($\beta_{without}$), and the associated R^2 from these regressions, we can estimate a ratio of the correlation between unobservables and mentor evaluations (ρ_{ue}) to the correlation between observables and mentor evaluations (ρ_{oe}) necessary to reduce our original coefficient to zero. Formally, this ratio equals:

$$\frac{\beta_{with} / \left(0.85 - R^2_{with}\right)}{\beta_{without} / \left(R^2_{with} - R^2_{without}\right)}$$

Ratios larger than one indicate that the correlation of a variable with unobservables must be greater than the correlation with observables in order to drive the estimated coefficient to zero.

[19] One possible explanation (there are many) for the higher estimates in Jacob and Lefgren is that they predict end-of-year test scores using evaluations collected in the middle of the school year, while the evaluations used by Harris and Sass and those that we analyze were collected prior to the school year used for prediction.

Table 5
Robustness checks.

	Math					
	(1)	(2)	(3)	(4)	(5)	(6)
Deviation from mentor's average evaluation	0.026	0.027	0.023	0.027	0.027	0.051
Sept–Nov, N(0,1)	(0.009)**	(0.009)**	(0.009)**	(0.009)**	(0.008)**	(0.023)**
Unobservable/Observable ratio for zero effect						3.8
Mentor's average evaluation	0.011	0.014	−0.001		−0.011	0.187
Sept–Nov, N(0,1)	(0.011)	(0.011)	(0.013)		(0.015)	(0.030)**
Deviation from mentor's average evaluation	0.049	0.050	0.057	0.059	0.055	0.174
Mar–Jun, N(0,1)	(0.009)**	(0.009)**	(0.009)**	(0.008)**	(0.008)**	(0.023)**
Unobservable/Observable ratio for zero effect						1.5
Mentor's average evaluation	0.003	0.003	0.007		−0.002	0.034
Mar–Jun, N(0,1)	(0.008)	(0.008)	(0.010)		(0.013)	(0.020)*
Teachers in schools w/1+ evaluated teachers	√	√				√
Zip-code fixed effects	√	√	√			
Only evaluated teachers			√	√	√	
Mentor fixed effects				√		
School fixed effects					√	
Added control for prior class achievement in year 1		√				
Drop controls for observables (Altonji-Elder-Taber)						√
Observations	387,916	387,916	61,319	61,319	61,319	387,920
Teachers	7660	7660	1747	1747	1747	7660
Teachers with evaluations	1747	1747	1747	1747	1747	1747
R^2	0.67	0.67	0.66	0.67	0.69	0.01
	English					
	(7)	(8)	(9)	(10)	(11)	(12)
Deviation from mentor's average evaluation	0.002	0.002	−0.003	−0.001	−0.012	0.040
Sept–Nov, N(0,1)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.024)*
Unobservable/Observable ratio for zero effect						0.1
Mentor's average evaluation	−0.003	−0.003	−0.001		−0.017	0.159
Sept–Nov, N(0,1)	(0.009)	(0.009)	(0.009)		(0.011)	(0.028)**
Deviation from mentor's average evaluation	0.023	0.022	0.025	0.024	0.024	0.134
Mar–Jun, N(0,1)	(0.006)**	(0.007)**	(0.006)**	(0.007)**	(0.007)**	(0.021)**
Unobservable/Observable ratio for zero effect						0.5
Mentor's average evaluation	−0.007	−0.008	−0.006		0.005	−0.042
Mar–Jun, N(0,1)	(0.006)	(0.006)	(0.007)		(0.008)	(0.018)**
Teachers in schools w/1+ evaluated teachers	√	√				√
Zip-code fixed effects	√	√	√			
Only evaluated teachers			√	√	√	
Mentor fixed effects				√		
School fixed effects					√	
Added control for prior class achievement in year 1		√				
Drop controls for observables (Altonji-Elder-Taber)						√
Observations	340,297	340,297	53,271	53,271	53,271	340,297
Teachers	7803	7803	1737	1737	1737	7803
Teachers with evaluations	1737	1737	1737	1737	1737	1737
R^2	0.61	0.61	0.60	0.60	0.61	0.01

Notes: Standard errors (in parentheses) are clustered at the teacher level. Observable controls include 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, and additional controls include 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, class size, and year-grade fixed effects. See text for a detailed explanation of the Unobservable/Observable Ratio.
** significant at 5%.
* significant at 10%.

English achievement, and in English achievement we can reject equality of the three coefficients at conventional levels. In other words, mentor evaluations appear to have greater power to distinguish effective and ineffective teachers among those whose first year value-added does not put them either at the lower or upper tail of the distribution.[20]

These results are somewhat reminiscent of Jacob and Lefgren (2008) finding a non-linear relation between contemporaneous value-added added measures and principals' subjective opinions of teacher effectiveness. Specifically, they find most of the power is in the tails of the subjective evaluation distribution. The analysis presented in Table 6 speaks to a different issue—whether the relation between subjective evaluation and *future* performance is similar at different parts of the objective evaluation distribution. To provide a better comparison with Jacob and Lefgren, we ran regressions of teachers' first-year value added on cubic polynomials of the mentor's subjective evaluation. We find no evidence of significant non-linearity; the linear terms are always large and highly significant but the higher order terms are always small and insignificant. As mentioned above, our setting differs from Jacob and Lefgren in a number of ways, and locating the cause for the divergence in our results is beyond the scope of this paper.

Our second extension is to investigate whether evaluations by TF interviewers and mentors who have more evaluation experience are more powerful predictors of student achievement. Specifically, to our main regression specification we add a control for the number of interviews conducted by each TF interviewer prior to their interview

[20] Importantly, these coefficients are dependent on scaling and one should interpret them with caution. We base our analysis on scale scores from standardized exams and (normalized) evaluations submitted by the mentors on the 1–5 scale, but there are other reasonable scales one could use (e.g., percentiles). Given this caveat, one potential explanation for our finding is that the mentor evaluation process was geared toward "typical" experiences of first-year teachers and that mentors were less adept at evaluating teachers having major problems or performing far above their peers.

Table 6
Interactions of subjective and objective evaluations for mentored teachers.

	Math		English	
	(1)	(2)	(3)	(4)
Objective evaluation Mentor evaluation in…				
Bottom tercile	0.080		0.016	
	(0.017)**		(0.007)**	
Middle tercile	0.080		0.017	
	(0.010)**		(0.008)**	
Top tercile	0.090		0.019	
	(0.009)**		(0.007)**	
Mentor evaluation Objective evaluation in…				
Bottom tercile		0.028		0.002
		(0.019)		(0.013)
Middle tercile		0.041		0.047
		(0.022)*		(0.015)**
Top tercile		0.029		0.018
		(0.019)		(0.013)
Observations	387,916	387,916	340,297	340,297
Teachers	7660	7660	7803	7803
Teachers with subjective and objective evaluation	1078	1078	1113	1113
R^2	0.67	0.67	0.61	0.61

Notes: Standard errors (in parentheses) are clustered at the teacher level. Mentor evaluations are those submitted from March through June. Regressions in odd columns control for objective evaluation tercile indicators and regressions in even columns control for subjective evaluation tercile indicators. Other control variables are the same as those enumerated in Tables 3 and 4.
** significant at 5%.
* significant at 10%.

with each TF candidate, a control for the number of teachers with whom a mentor has worked, and interactions of these variables with subjective evaluations (Table 7). For math scores, we do find a positive interaction of experience and evaluations given by mentors at the start of the school year. This provides some suggestive evidence that experienced mentors have more accurate "first impressions" of teacher effectiveness, but evaluations made after a full year of observation are no more predictive for experienced mentors than their less experienced colleagues.

4. Conclusion

We use data from New York City to examine the power of subjective and objective evaluations to identify effective and ineffective teachers early in their careers. We find evidence that teachers who receive better subjective evaluations of teaching ability prior to hire or in their first year of teaching also produce greater gains in achievement, on average, with their future students. Consistent with prior research, our results support the idea that teachers who produce greater achievement gains in the first year of their careers also produce greater gains, on average, in future years with different students. More importantly, subjective evaluations present significant and meaningful information about a teacher's future success in raising student achievement even conditional on objective data on first year performance. This is an especially noteworthy finding, considering that variation in subjective evaluations likely also captures facets of teaching skill that may affect outcomes not captured by standardized tests.

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 states and school districts will begin to implement policies that put greater stress on teacher effectiveness.

Table 7
Predictive power of evaluations and evaluator experience.

	Teaching Fellows	Mentored teachers	
Math	(1)	(2)	(3)
TF evaluator experience, *N(0,1)*	0.001		
	(0.009)		
TF evaluation, *4 point scale*	0.005		
	(0.013)		
TF evaluator experience*TF Evaluation	0.001		
	(0.004)		
Mentor experience, *N(0,1)*		0.008	0.008
		(0.005)	(0.005)
Mentor evaluation, *Sept–Nov, N(0,1)*		0.029	
		(0.009)**	
Mentor experience*Mentor evaluation *(Sept–Nov)*		0.013	
		(0.007)*	
Mentor evaluation, *Mar–Jun, N(0,1)*			0.054
			(0.009)**
Mentor experience*Mentor evaluation *(Mar–Jun)*			−0.005
			(0.006)
Observations	387,916	387,916	387,916
Teachers	7660	7660	7660
Teachers with evaluations	492	1747	1747
R^2	0.67	0.67	0.67
	Teaching Fellows	Mentored teachers	
English	(4)	(5)	(6)
TF evaluator experience, *N(0,1)*	0.007		
	(0.005)		
TF evaluation, *4 point scale*	−0.000		
	(0.010)		
TF evaluator experience*TF evaluation	0.003		
	(0.003)		
Mentor experience, *N(0,1)*		0.005	0.005
		(0.004)	(0.004)
Mentor evaluation, *Sept–Nov, N(0,1)*		0.009	
		(0.007)	
Mentor experience*Mentor evaluation *(Sept–Nov)*		−0.002	
		(0.006)	
Mentor evaluation, *Mar–Jun, N(0,1)*			0.026
			(0.007)**
Mentor experience*Mentor evaluation *(Mar–Jun)*			−0.006
			(0.004)
Observations	340,297	340,297	340,297
Teachers	7803	7803	7803
Teachers with evaluations	398	1737	1737
R^2	0.61	0.61	0.61

Notes: Standard errors (in parentheses) are clustered at the teacher level. Mentor evaluations from the beginning of the year include only those submitted on or before November 15 of each school year. Experience represents the number of teachers the TF evaluator/mentor has rated when evaluating each teacher and is normalized to have mean zero and standard deviation of one. All regressions controls for the same additional variables as in Tables 3 and 4.
** significant at 5%.
* significant at 10%.

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, Barrow, Lisa, Sander, William, 2007. Teachers and student achievement in the Chicago public high schools. Journal of Labor Economics 25 (1), 95–135.

Altonji, Joseph G., Elder, Todd E., Taber, Christopher R., 2005. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. Journal of Political Economy 113 (1), 151–184.
Anderson, Harold M., 1954. A study of certain criteria of teaching effectiveness. Journal of Experimental Education 23 (1), 41–71.
Armor, David, Conry-Oseguera, Patricia, Cox, Millicent, King, Nicelma, McDonnell, Lorraine, Pascal, Anthony, Pauly, Edward, Zellman, Gail, 1976. Analysis of the School Preferred Reading Program in Selected Los Angeles Minority Schools. Rand Corp, Santa Monica, CA.
Boyd, Donald, Grossman, Pamela, Lankford, Hamilton, Loeb, Susanna, Wyckoff, James, 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. Education Finance and Policy 1 (2), 176–216.
Boyd, Donald, Lankford, Hamilton, Loeb, Susanna, Rockoff, Jonah E., Wyckoff, James, 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.
Brookover, Wilbur B., 1945. The relation of social factors to teaching ability. Journal of Experimental Education 13 (4), 191–205.
Brophy, Jere, Good, Thomas L., 1986. Teacher behavior and student achievement, In: Wittrock, M.C. (Ed.), Handbook of Research on Teaching, 3rd ed. Simon and Schuster, New York, pp. 238–375.
Cantrell, Steven, Jon Fullerton, Kane, Thomas J., and Staiger, Douglas O. 2007. "National Board Certification and Teacher Effectiveness: Evidence from a Random Assignment Experiment." Unpublished Manuscript.
Cavalluzzo, Linda., 2004. Is national board certification an effective signal of teacher quality? CNA Corporation Working Paper.
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.
Goldhaber, Dan, Anthony, Emily, 2007. Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. The Review of Economics and Statistics 89 (1), 134–150.
Goldhaber, Dan, Hansen., Michael, 2009. Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions. Center on Reinventing Public Education Working Paper #2009_2.
Gordon, Robert, Kane, Thomas J., Staiger, Douglas O., 2006. The Hamilton Project: Identifying Effective Teachers Using Performance on the Job. The Brookings Institution, Washington, DC.
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 Sass, Tim R. 2006. "Value-Added Models and the Measurement of Teacher Quality." Unpublished Manuscript, Florida State University.
Harris, Douglas N., and Sass, Tim R. 2009. "What Makes for a Good Teacher and Who Can Tell?" Calder Center Working Paper #30.
Harris, Douglas N., Sass, Tim R., 2007. The effects of NBPTS-certified teachers on student achievement. National Center for Analysis of Longitudinal Data in Education Research Working Paper #4.
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.
Jackson, C. Kirabo, 2009. Student demographics, teacher sorting, and teacher quality: evidence from the end of school desegregation. Journal of Labor Economics 27 (2), 213–256.
Jacob, Brian A., Lefgren, Lars J., 2008. Can principals identify effective teachers? Evidence on subjective evaluation in education. Journal of Labor Economics 26 (1), 101–136.
Kane, Thomas J., Staiger, Douglas O., 2008. Estimating teacher impacts on student achievement: an experimental evaluation. NBER Working Paper 14607.
Kane, Thomas J., Rockoff, Jonah E., Staiger, Douglas O., 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review 27 (6), 615–631.
Kimball, Steven M., White, Brad, Milanowski, Anthony T., Borman, Geoffrey, 2004. Examining the relationship between teacher evaluation and student assessment results in Washoe County. Peabody Journal of Education 79 (4), 54–78.
Koedel, Cory, Betts, Julian R., 2011. Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. Education Finance and Policy 6 (1), 18–42.
Manatt, Richard P., Daniels, Bruce, 1990. Relationships between principals' ratings of teacher performance and student achievement. Journal of Personnel Evaluation in Education 4 (2), 189–201.
Medley, Donald M., Coker, Homer, 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. Balinger, Cambridge, MA.
Peterson, Kenneth D., 1987. Teacher evaluation with multiple and variable lines of evidence. American Educational Research Journal 24 (2), 311–317.
Rivkin, Steven G., Hanushek, Eric A., Kain, John, 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., Staiger, Douglas O., Eric, Taylor, Kane, Thomas J., 2010. Information and employee evaluation: evidence from a randomized intervention in public schools. NBER Working Paper 16240.
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.
Rothstein, Jesse, 2009. Student sorting and bias in value-added estimation: selection on observables and unobservables. Education Finance and Policy 4 (4), 537–571.
Sanders, William L., Rivers, June C., 1996. Cumulative and residual effects of teachers on future student academic achievement. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
Schacter, John, Thum, Yeow M., 2004. Paying for high- and low-quality teaching. Economics of Education Review 23 (4), 411–440.
Tyler, John H., Taylor, Eric S., Kane, Thomas J., Wooten, Amy L., 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, Sexton, Susan, Mulhern, Jennifer, Keeling, David, 2009. The Widget Effect. The New Teacher Project, Brooklyn, NY.
Wilkerson, David J., Manatt, Richard P., Rogers, Mary A., Maughan, Ron, 2000. Validation of student, principal, and self-ratings in 360° Feedback® for teacher evaluation. Journal of Personnel Evaluation in Education 14 (2), 179–192.