American Economic Review: Papers & Proceedings 100 (May 2010): 261–266
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.261

Subjective and Objective Evaluations of Teacher Effectiveness

By Jonah E. Rockoff and Cecilia Speroni*

Research on the impact of teachers on student achievement (e.g., Jonah E. Rockoff 2004; Steven G. Rivkin, Hanushek, and John Kain 2005) has established two stylized facts: (1) teacher effectiveness varies widely, and (2) outside of experience, qualifications that determine a teacher's certification and salary bear little relation to outcomes. This provides motivation to understand how to identify effective and ineffective teachers, particularly early in their careers.

Studies that examine how student achievement data can predict teachers' impacts on student outcomes in the future (e.g., Robert Gordon, Thomas J. Kane, and Douglas O. Staiger 2006; Dan Goldhaber and Michael Hansen 2010) conclude that using such data to selectively retain teachers could yield large benefits. However, "value-added" measures of effectiveness are noisy and can be biased if some teachers are persistently given students that are difficult to teach in ways that are hard to observe. Thus, using other information may achieve more stability and accuracy in teacher evaluations.

There is also a literature on subjective teaching evaluations (i.e., evaluations by the school principal or evaluations based on classroom observation protocols or "rubrics"), which also finds significant relationships between evaluations and achievement gains.[1] However, these studies typically investigate how evaluations predict the exam performance of current, not future, students. A stronger test would be to examine a new group of students assigned to the teacher in another year (as done by Gordon, Kane, and Staiger 2006). Also, teachers in these studies are usually experienced, and these results may not generalize to new teachers.

In this paper, we measure how subjective and objective evaluations of new teachers in New York City predict their future impacts on student achievement. Specifically, we examine evaluations of applicants to an alternative certification program, evaluations of new teachers by professional mentors that work with them during their first year, and evaluations based on student achievement data from their first year of teaching. We use a large sample, relative to prior work, and, unlike other studies (with the exception of John H. Tyler et al. 2010), we examine subjective evaluations made by professionals as part of their jobs, not survey responses.

Examined separately, both subjective and objective evaluations bear significant relationships with the achievement of the teachers' future students. Moreover, when both types of evaluations are entered in a regression of future students' test scores, their coefficients are only slightly attenuated—each evaluation contains information distinct from the other. We also find evidence of variation in the leniency with which standards were applied by some evaluators. Specifically, for evaluations by mentors, variation in evaluations *within* evaluators is a much stronger predictor of student outcomes than variation *between* evaluators. This highlights the importance of reliability in the procedures used to generate subjective evaluations.

I. Data and Descriptive Statistics

We primarily use data on the characteristics and achievement of grade 3 to 8 students in New York City during the school years 2003–2004

*Rockoff: Columbia Business School, Uris Hall 603, 3022 Broadway, New York, NY 10027–6902 (e-mail: jonah.rockoff@columbia.edu); Speroni: Columbia University, Teachers College, 525 West 120th Street, Box 174, New York, NY 10027 (e-mail: cs2456@columbia.edu). 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.

[1] Early studies of principal evaluations were done by educators (e.g., C.W. Hill 1921, Harold M. Anderson 1954), but economists have made recent contributions (e.g., Brian A. Jacob and Lars J. Lefgren 2008). An example of a study of evaluations based on classroom observation rubrics is Anthony Milanowski (2004).

through 2007–2008, as well as information on their math and English teacher(s). We evaluate teachers' impacts on student test scores in their first year using an empirical Bayes's method. We avoid using data from teachers' second years to evaluate first-year performance.[2]

One set of data on subjective evaluations comes from the New York City Teaching Fellows (TF), an alternative path to certification taken by about a third of new teachers in New York City.[3] We use data on TF applicants who began teaching in the school years 2004–2005 through 2006–2007, and were evaluated on a five-point scale during an interview process.[4] To be accepted into TF, applicants must receive one of the top three evaluations; after a committee review, about five percent of applicants receiving lower evaluations are accepted. Very few applicants received the second-lowest evaluation, and, in our analysis, we combine them with Fellows receiving the lowest evaluation.

The second source of subjective evaluations data is a mentoring program for new teachers which operated during the school years 2004–2005 through 2006–2007.[5] Starting between late September and mid-October, a trained, full time mentor would meet with each teacher every one or two weeks and work on improving his/her teaching skills. Mentors submitted monthly summative evaluations and bimonthly formative evaluations of teachers on a five-point scale, based on a detailed set of teaching standards.[6] Summative and formative evaluations are highly correlated (coefficient of correlation 0.84), and we therefore average them into a single measure of teacher effectiveness. While evaluations by mentors may have been affected by the students assigned to teachers in their first year, it is 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 up until November 15. We use evaluations from March through June to examine teacher effectiveness the following year.

Some mentors and TF interviewers may have been "tougher" than others in applying the evaluation standards on which they were trained.

[2] Our method follows Kane, Rockoff, and Staiger (2008). The empirical Bayes estimator requires an estimate of the correlation across years in the average residuals across classrooms taught by the same teacher. However, rather than obtain a single estimate for all years, we run a series of regressions, each of which uses two years of data and produces objective evaluations for a single cohort of first-year teachers (e.g., data from 2004–2005 and 2005–2006 are used to estimate value added for teachers who began their careers in school year 2005–2006). Some teachers received subjective evaluations and were linked to students in their second year, but not their first year. To include them in our regressions, we set their value-added estimates to zero and include a variable indicating a missing estimate.

[3] Fellows attend an intensive pre-service training program to prepare them to teach and study for a master's degree in education while teaching. Approximately 60 percent of Teaching Fellows applicants are invited for an interview, which includes a mock teaching lesson, a written essay, a discussion with other applicants, and a personal interview. Kane, Rockoff, and Staiger (2008) provide a more detailed description and analysis of this program.

[4] The first evaluations on a five-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 five-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.

[5] See Rockoff (2008) for a detailed analysis of this program. It 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 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.

[6] Formative evaluations were more detailed than summative evaluations. Teachers were rated on six competencies, and each of these competencies had between five and eight items. However, evaluations were highly correlated (and often identical) across competencies. Factor analysis (results available upon request) reveals 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 nonmissing subcategory evaluations. 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 evaluations were not completed every month for every teacher, we account for the timing of teachers' evaluations by normalizing evaluations by the month and year they were submitted.

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's degree	36%	21%	76%
Student characteristics			
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.

Fortunately, each TF interviewer typically saw dozens of applicants, and each mentor worked with roughly 15 teachers per year. In our analysis, we separate overall variation in evaluations from relative variation within evaluators.

We examine teachers of math and/or English to students in grades 4 to 8 during the school years 2004–2005 through 2007–2008.[7] Table 1 provides descriptive statistics for these teachers, separately for those who did and did not receive subjective evaluations. Teachers with evaluations are younger, less likely to have a master's degree, and have little experience. Their students 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.

II. Methodology and Regression Estimates

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

$$A_{ikt} = \gamma \mathbf{Eval}_k + \beta \mathbf{X}_{it} + \lambda \mathbf{T}_{ikt} + \sum_{g,t} \pi_{gt} D_{it}^{g} + \sum_{z} \pi_z D_{it}^{z} + \varepsilon_{ikt} \quad (1)$$

where A_{ikt} is the standardized achievement test score for student i taught by teacher k in year t, $\mathbf{Eval}_k$ is a vector of (subjective and/or objective) evaluations of teacher effectiveness, $\mathbf{X}_{it}$ are student level control variables (including prior achievement), $\mathbf{T}_{ikt}$ are controls for teacher and classroom 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 error term. To gain precision on estimates of fixed effects and other coefficients, the sample includes students taught by other teachers in the same schools. For these teachers, evaluation(s) variable(s) are set to zero and we include an indicator variable for missing evaluation(s). Standard errors are clustered by teacher.

Estimates of the power of subjective evaluations to predict student achievement in a teacher's first year are shown in panel A of Table 2. Due to space constraints, only results for math are shown, though we discuss the results for English achievement. Evaluations are normalized to have mean zero and standard deviation of one, and student test scores are also normalized at the year-grade level. 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). Notably,

[7] We also implement a few additional sample restrictions, following Kane, Rockoff, and Staiger (2008). For example, we drop classrooms with more than 25 percent special education students.

if we add a control for the average evaluation given out by mentors, we find it has a negative significant coefficient, indicating that mentors varied in their application of evaluation standards (column 4). Coefficients on both types of evaluations for English achievement are positive but quite small and statistically insignificant. However, estimates of variance in teacher effectiveness are smaller for English than math, both in New York and elsewhere. Thus, we need more power to identify statistically significant effects in English of the same *proportional* magnitude as the effects we find for math.

In specifications that include TF and mentor evaluations—where coefficients are identified from variation across teachers with both evaluations—the estimates are quite similar. Interestingly, the coefficient on mentor evaluations from the start of the year is considerably larger in English for this subsample of teachers (i.e., Teaching Fellows who receive mentoring services) than for all mentored teachers (0.03 versus 0.005) and statistically significant, suggesting a stronger relationship between achievement and mentor evaluations for Teaching Fellows.

We then examine student achievement in a teacher's second yearg. First, we show that the value-added estimates are highly significant predictors of student achievement in a teacher's second year (panel B, Table 2, column 1), with more variation in achievement predicted in math (0.09) than English (0.02).[8] This is consistent with prior research (e.g., Gordon et al. 2006, Kane and Staiger 2008).

In both subjects, TF evaluations from recruitment and student achievement in the second year are not significantly related (panel B, Table 2, columns 2 and 3). 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.[9]

III. Conclusion

We find that teachers who receive higher subjective evaluations either prior to hire or in their first year of teaching produce greater average gains in achievement with their future students. Consistent with prior work, we also find that teachers who produce greater test score gains in their first year also produce greater average gains in their second year. Importantly, we find that—conditional on objective data on first year performance—subjective evaluations present 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. Our results 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.

[8] 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 5, 6, and 7.

[9] 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.

Table 2—Evaluations and Student Achievement in a Teacher's First and Second Year

	Teaching fellows		Mentored teachers		TF and mentored		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. First year math outcomes							
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*			0.016	0.021		0.021	0.018
			(0.008)*	(0.009)**		(0.018)	(0.017)
Mentor average evaluation, *Sept–Nov*				−0.014		−0.012	−0.011
				(0.008)*		(0.014)	(0.014)

	All teachers	Teaching fellows		Mentored teachers			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel B. Second year math outcomes							
Objective evaluation, *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*				0.032		0.026	0.024
				(0.009)**		(0.009)**	(0.008)**
Mentor average evaluation, *Sept–Nov*				−0.018		−0.015	−0.014
				(0.011)*		(0.011)	(0.011)
Mentor evaluation, *Mar–Jun*					0.054	0.050	0.032
					(0.009)**	(0.008)**	(0.008)**
Mentor average evaluation, *Mar–Jun*					−0.052	−0.047	−0.031
					(0.012)**	(0.012)**	(0.011)**

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, class size, year-grade, and zip code fixed effects. Panel A regressions are based on a sample of 399,982 student-year observation and 8,287 teachers. In panel A, teachers with evaluations number 529 in columns 1 and 2, 1,868 in columns 3 and 4, and 477 in columns 5 to 7. Panel B regressions are based on a sample of 389,530 student-year observations and 7,678 teachers. In panel B, teachers with evaluations number 1,821 in column 1, 501 in columns 2 and 3 and 1,755 in columns 4 to 7.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

REFERENCES

Goldhaber, Dan, and Michael Hansen. 2010. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review*, 100(2): 388–392.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Hamilton Project Discussion Paper 2006-01.

Hill, C.W. 1921. "The Efficiency Ratings of Teachers." *The Elementary School Journal*, 21(6): 438–43.

Jacob, Brian A., and Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26(1): 101–36.

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): 615–31.

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

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–58.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review*, 94(2): 247–52.

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

Tyler, John H., Eric S. Taylor, Thomas J. Kane, and Amy L. Wooten. 2010. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review*, 100(2): 393–397.