Can Teacher Evaluation Improve Teaching?

Evidence of systematic growth in the effectiveness of midcareer teachers

By Eric S. Taylor and John H. Tyler

| | |

FALL 2012 / VOL. 12, NO. 4

The modernization of teacher evaluation systems, an increasingly common component of school reform efforts, promises to reveal new, systematic information about the performance of individual classroom teachers. Yet while states and districts race to design new systems, most discussion of how the information might be used has focused on traditional human resource–management tasks, namely, hiring, firing, and compensation. By contrast, very little is known about how the availability of new information, or the experience of being evaluated, might change teacher effort and effectiveness.

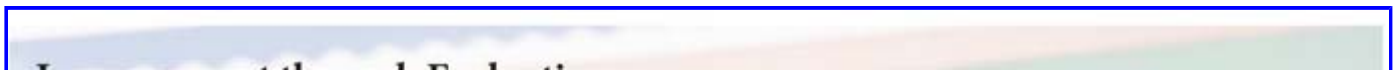
In the research reported here, we study one approach to teacher evaluation: practice-based assessment that relies on multiple, highly structured classroom observations conducted by experienced peer teachers and administrators. While this approach contrasts starkly with status quo "principal walk-through" styles of class observation, its use is on the rise in new and proposed evaluation systems in which rigorous classroom observation is often combined with other measures, such as teacher value-added based on student test scores.

Proponents of evaluation systems that include high-quality classroom observations point to their potential value for improving instruction (see "Capturing the Dimensions of Effective Teaching," *Features, Fall 2012*). Individualized, specific information about performance is especially scarce in the teaching profession, suggesting that a lack of information on *how* to improve could be a substantial barrier to individual improvement among teachers. Well-designed evaluation might fill that knowledge gap in several ways. First, teachers could gain information through the formal scoring and feedback routines of an evaluation program. Second, evaluation could encourage teachers to be generally more self-reflective, regardless of the evaluative criteria. Third, the evaluation process could create more opportunities for conversations with other teachers and administrators about effective practices.

In short, there are good reasons to expect that well-designed teacher-evaluation programs could have a direct and lasting effect on individual teacher performance. To our knowledge, however, ours is the first study to test this hypothesis directly. We study a sample of midcareer elementary and middle school teachers in the Cincinnati Public Schools, all of whom were evaluated in a yearlong program, based largely on classroom observation, sometime between the 2003–04 and 2009–10 school years. The specific school year of each teacher's evaluation was determined years earlier by a district planning process. This policy-based assignment of *when* evaluation occurred permits a quasi-experimental analysis. We compare the achievement of individual teachers' students before, during, and after the teacher's evaluation year.

We find that teachers are more effective at raising student achievement during the school year when they are being evaluated than they were previously, and even more effective in the years after evaluation. A student instructed by a teacher after that teacher has been through the Cincinnati evaluation will score about 11 percent of a standard deviation (4.5 percentile points for a median student) higher in math than a similar student taught by the same teacher before the teacher was evaluated.

Our data do not allow us to identify the exact mechanisms driving these improvements. Nevertheless, the results contrast sharply with the view that the effectiveness of individual teachers is essentially fixed after the first few years on the job. Indeed, we find that postevaluation improvements in performance were largest for teachers whose performance was weakest prior to evaluation, suggesting that rigorous teacher evaluation may offer a new way to think about teacher professional development.



Evaluation in Cincinnati

The data for our analysis come from the Cincinnati Public Schools. In the 2000–01 school year, Cincinnati launched the Teacher Evaluation System (TES) in which teachers' performance in and out of the classroom is assessed through classroom observations and a review of work products. During the yearlong TES process, teachers are typically observed in the classroom and scored four times: three times by an assigned peer evaluator—a high-performing, experienced teacher who previously taught in a different school in the district—and once by the principal or another school administrator. Teachers are informed of the week during which the first observation will occur, with all other observations unannounced. Owing mostly to cost, tenured teachers are typically evaluated only once every five years.

The evaluation measures dozens of specific skills and practices covering classroom management, instruction, content knowledge, and planning, among other topics. Evaluators use a scoring rubric based on Charlotte Danielson's *Enhancing Professional Practice: A Framework for Teaching*, which describes performance of each skill and practice at four levels: "Distinguished," "Proficient," "Basic," and "Unsatisfactory." (See Table 1 for a sample standard.)

Both the peer evaluators and administrators complete an intensive TES training course and must accurately score videotaped teaching examples. After each classroom observation, peer evaluators and administrators provide written feedback to the teacher and meet with the teacher at least once to discuss the results. At the end of the evaluation school year, a final summative score in each of four domains of practice is calculated and presented to the evaluated teacher. Only these final scores carry explicit consequences. For beginning teachers (those evaluated in their first and fourth years), a poor evaluation could result in nonrenewal of their contract, while a successful evaluation is required before receiving tenure. For tenured teachers, evaluation scores determine eligibility for some promotions or additional tenure protection, or, in the case of very low scores, placement in a peer assistance program with a small risk of termination.

Despite the training and detailed rubric provided to evaluators, the TES program experiences some of the leniency bias typical of other teacher-evaluation programs. More than 90 percent of teachers receive final overall TES scores in the highest two categories. Leniency is much less frequent in the individual rubric items and individual observations. We hypothesize that this microlevel evaluation feedback is more important to lasting performance improvements than the final, overall TES scores.

Previous research has found that the scores produced by TES predict student achievement gains (see "Evaluating Teacher Effectiveness," *research*, Summer 2011). Student math achievement was 0.09 standard deviations higher for teachers whose overall evaluation score was 1 standard deviation higher (the estimate for reading was 0.08). This relationship suggests that Cincinnati's evaluation program provides feedback on teaching skills that are associated with larger gains in student achievement.

As mentioned above, teachers only undergo comprehensive evaluation periodically. All teachers newly hired by the district, regardless of experience, are evaluated during their first year working in Cincinnati schools. Teachers are also evaluated just prior to receiving tenure, typically their fourth year after being hired, and every fifth year after achieving tenure.

Teachers hired before the TES program began in 2000–01 were not initially evaluated until some years into the life of the program. Our analysis only includes these pre-TES hires: specifically, teachers hired by the district in the school years from 1993–94 through 1999–2000. We further focus, given available data, on those who were teaching 4th through 8th grade in the years 2003–04 through 2009–10. We limit our analysis to this sample of midcareer teachers for three reasons. First, for teachers hired before the new TES program began in 2000–01, the timing of their first TES review was determined largely by a "phase-in" schedule devised during the program's planning stages. This schedule set the year of first evaluation based on a teacher's year of hire, thus reducing the potential for bias that would arise if the timing of evaluation coincided with, for example, a favorable class assignment. Second, because the timing of evaluation was determined by year of hire, and not experience level, teachers in our sample were evaluated at different points in their careers. This allows us to measure the effect of evaluation on performance separate from any gains that come from increased experience. Third, the delay in first evaluation allows us to observe the achievement gains of these teachers' students in classes the teachers taught before the TES assessment so that we can make before-and-after comparisons of the same teacher.

Additionally, our study focuses on math test scores in grades 4–8. For most other subjects and grades, student achievement measures are simply not available. Students are tested in reading, but empirical research frequently finds less teacher-driven variation in reading achievement than in math, and ultimately this is the case for the present analysis as well. While not the focus of our research, we briefly discuss reading results below.

Data provided by the Cincinnati Public Schools identify the year(s) in which a teacher was evaluated by TES, the dates when each observation occurred, and the scores. We combine these TES data with additional administrative data provided by the district that allow us to match teachers to students and student test scores. As we would expect, the 105 teachers in our analysis sample are a highly experienced group: 66.5 percent have 10 to 19 years of experience, compared to 29.3 percent for the rest of the district. Teachers in our analysis are also more likely to have a graduate degree and be certified by the National Board for Professional Teaching Standards, two characteristics correlated with experience.

How Teachers Are Evaluated in Cincinnati: A Sample (Table 1)

Standard 3.4: The teacher engages students in discourse and uses thought-provoking questions aligned with the lesson objectives to explore and extend content knowledge.

	Distinguished	Proficient	Basic	Unsatisfactory
Discourse	Teacher structures and facilitates discourse at the evaluative, synthesis, and/or analysis levels between teacher and students and among students to explore and extend content knowledge.	Teacher initiates and leads discourse at the evaluative, synthesis, and/or analysis levels to explore and extend the content knowledge.	Teacher frames content-related discussion that is limited to a question and answer session.	Teacher permits off-topic discussions, or does not elicit student responses.
Thought-Provoking Questions	Teacher routinely asks thought -provoking questions at the evaluative, synthesis, and/ or analysis levels that focus on the objectives of the lesson. Teacher seeks clarification and elaboration through additional questions. Teacher provides appropriate wait time.	Teacher asks thought-provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. Teacher seeks clarification through additional questions. Teacher provides appropriate wait time.	Teacher asks questions that are relevant to the objectives of the lesson. Teacher asks follow-up questions. Teacher is inconsistent in providing appropriate wait time.	Teacher frequently asks questions that are inappropriate to objectives of the lesson. Teacher frequently does not ask follow-up questions. Teacher answers own questions. Teacher frequently does not provide appropriate wait time.

SOURCE: Cincinnati Public Schools Teacher Evaluation System 2005. The complete rubric is available at http://www.cps-k12.org/employment/tchreval/stndsrubrics.pdf.

Methodology

Our objective is to measure the impact of practice-based performance evaluation on teacher effectiveness. Simply comparing the test scores of students whose teachers are evaluated in a given year to the scores of other teachers' students would produce misleading results because, among other methodological issues, less-experienced teachers are more likely to be evaluated than more-experienced teachers.

Instead, we compare the achievement of a teacher's students during the year that she is evaluated to the achievement of

the *same teacher's students* in the years before and after the evaluation year. As a result, we effectively control for any characteristics of the teacher that do not change over time. In addition, we control for determinants of student achievement that may change over time, such as a teacher's experience level, as well as for student characteristics, such as prior-year test scores, gender, racial/ethnic subgroup, special education classification, gifted classification, English proficiency classification, and whether the student was retained in the same grade.

Our approach will correctly measure the effect of evaluation on teacher effectiveness as long as the timing of a teacher's evaluation is unrelated to any student characteristics that we have not controlled for in the analysis but that affect achievement growth. This key condition would be violated, for example, if during an evaluation year or in the years after, teachers were systematically assigned students who were better (or worse) in ways we cannot determine and control for using the available data. It would also be violated if evaluation coincided with a change in individual teacher performance unrelated to evaluation per se. Below, we discuss evidence that our results are not affected by these kinds of issues. We also find no evidence that teachers are systematically assigned students with better (or worse) observable characteristics in their evaluation year compared to prior and subsequent years.

Results

We find suggestive evidence that the effectiveness of individual teachers improves during the school year when they are evaluated. Specifically, the average teacher's students score 0.05 standard deviations higher on end-of-year math tests during the evaluation year than in previous years, although this result is not consistently statistically significant across our different specifications.

These improvements persist and, in fact, increase in the years after evaluation (see Figure 1). We estimate that the average teacher's students score 0.11 standard deviations higher in years after the teacher has undergone an evaluation compared to how her students scored in the years before her evaluation. To get a sense of the magnitude of this impact, consider two students taught by the same teacher in different years who both begin the year at the 50th percentile of math achievement. The student taught after the teacher went through the TES process would score about 4.5 percentile points higher at the end of the year than the student taught before the teacher went through the evaluation.

We also find evidence that the effects of going through evaluation in the TES system are not the same for all teachers. The improvement in teacher performance from before to after evaluation is larger for teachers who received relatively low TES scores, teachers whose TES scores improved the most during the TES year, and especially for teachers who were relatively ineffective in raising student test scores prior to TES. The fact that the effects were largest for teachers who, presumably, received more critical feedback and for those with the most room for improvement strengthens our confidence in the causal interpretation of the overall results.

Our findings remain similar when we make changes to our methodological choices, such as varying the way we control for teacher experience, not controlling for teacher experience, and not controlling for student characteristics. We also examine whether our results could be biased by a preexisting upward trend in each teacher's performance unrelated to experience or evaluation, and find no evidence of such a trend. Finally, we find no evidence that our results reflect teacher turnover from school to school or from grade to grade that causes them not to appear in our data in later years (for example, by moving to a nontested grade or leaving the Cincinnati Public Schools).

In contrast to the results for math achievement, we do not find any evidence that being evaluated increases the impact that teachers have on their students' reading achievement. Many studies find less variation in teachers' effect on reading achievement compared to teachers' effect on math achievement, a pattern that is also evident in our data from Cincinnati. Some have hypothesized that the smaller differences in effectiveness among reading teachers could arise because students learn reading in many in- and out-of-school settings (e.g., reading with family at home) that are outside of a formal reading class. If teachers have less influence on reading achievement, then even if evaluation induces changes in teacher practices, those changes would have smaller effects on achievement growth.

Discussion

The results presented here—greater teacher performance as measured by student achievement gains in years following

TES review—strongly suggest that teachers develop skills or otherwise change their behavior in a *lasting* manner as a result of undergoing subjective performance evaluation in the TES process. A potential explanation for these results is that teachers learn new information about their own performance during the evaluation and subsequently develop new skills. New information is potentially created by the formal scoring and feedback routines of TES, as well as increased opportunities for self-reflection and for conversations regarding effective teaching practice in the TES environment.

Moreover, two features of this study—the analysis sample of experienced teachers and Cincinnati's use of peer evaluators—may increase the saliency of these hypothesized mechanisms. First, the teachers we study experienced their first rigorous evaluation after 8 to 17 years on the job. Thus they may have been particularly receptive to and in need of information on their performance. If, by contrast, teachers were evaluated every school year (as they are in a new but similar program in Washington, D.C.), the effect resulting from each subsequent year's evaluation might well be smaller. Second, Cincinnati's use of peer evaluators may result in teachers being more receptive to feedback from their subjective evaluation than they would be were the feedback to come solely from their supervising principals.

Teachers also appear to generate higher test-score gains during the year they are being evaluated, though these estimates, while consistently positive, are smaller. These improvements during the evaluation could represent the beginning of the changes seen in years following the review, or they could be the result of simple incentives to try harder during the year of evaluation, or some combination of the two.

A remaining question is whether the effects we find are small or large. A natural comparison would be to the estimated effects of different teacher professional-development programs (in-service training often delivered in formal classroom settings). Unfortunately, despite the substantial budgets allocated to such programs, there is little rigorous evidence on their effects. There are, however, other results from research on teacher effectiveness that can be used for comparison. First, the largest gains in teacher effectiveness appear to occur as teachers gain on-the-job experience in the first three to five years. Jonah Rockoff reports gains of about 0.10 student standard deviations over the first two years of teaching when effectiveness is measured by improvements in math computation skills; when using an alternative student math test measuring conceptual understanding, the gains are about half as large. Second, Kirabo Jackson and Elias Bruegmann find that having more effective teacher peers improves a teacher's own performance; a 1-standard-deviation increase in teacher-peer quality is associated with a 0.04-standard-deviation increase in student math achievement. Compared to these two findings, the sustained effect of TES assessment is large.

But are these benefits worth the costs? The direct expenditures for the TES program are substantial, which is not surprising given its atypically intensive approach. From 2004–05 to 2009–10, the Cincinnati district budget directly allocated between $1.8 and $2.1 million per year to the TES program, or about $7,500 per teacher evaluated. More than 90 percent of this cost is associated with evaluator salaries.

A second, potentially larger "cost" of the program is the departure from the classroom of the experienced and presumably highly effective teachers selected to be peer evaluators. The students who would otherwise have been taught by the peer evaluators will likely be taught by less-effective, less-experienced teachers; in those classrooms, the students' achievement gains will be smaller on average. (The peer evaluator may in practice be replaced by an equally effective or more effective teacher, but that teacher must herself be replaced in the classroom she left.)

While this second cost is more difficult to calculate, it is certainly offset by the larger gains made by students in the evaluated teachers' classrooms. Those students are scoring, on average, 10 percent of a standard deviation better than they would have otherwise, and since each peer evaluator evaluates 10 to 15 teachers each year, those gains are occurring in multiple teachers' classrooms for a number of years.

The results of our study provide evidence that subjective evaluation can improve employee performance, even after the evaluation period ends. This is particularly encouraging for the education sector. In recent years, the consensus among policymakers and researchers has been that after the first few years on the job, teacher performance, at least as measured by student test-score growth, cannot be improved. In contrast, we demonstrate that, at least in this setting, experienced teachers provided with unusually detailed information on their performance improved substantially.

American public schools have been under new pressure from regulators and constituents to improve teacher performance.

To date, the discussion has focused primarily on evaluation systems as sorting mechanisms, a means to identify the lowest-performing teachers for selective termination. Our work suggests optimism that, while costly, well-structured evaluation systems can not only serve this sorting purpose but can also enhance education through improvements in teacher effectiveness. In other words, if done well, performance evaluation can be an effective form of teacher professional development.

Eric S. Taylor is a doctoral student at Stanford University. John H. Tyler is professor of education, economics, and public policy at Brown University. This article is based in part on a forthcoming study in the American Economic Review.

Business Office

Program on Education Policy and Governance

Harvard Kennedy School

79 JFK Street, Cambridge, MA 02138

Phone (877) 476-5354 Fax (617) 496-1507