

Evaluating Teacher Effectiveness

Can classroom observations identify practices that raise achievement?

By [Thomas J. Kane](#), [Eric S. Taylor](#), [John H. Tyler](#) and [Amy L. Wooten](#)



Summer 2011 / Vol. 11, No. 3

“The Widget Effect,” a widely read 2009 report from The New Teacher Project, surveyed the teacher evaluation systems in 14 large American school districts and concluded that status quo systems provide little information on how performance differs from teacher to teacher. The memorable statistic from that report: 98 percent of teachers were evaluated as “satisfactory.” Based on such findings, many have characterized classroom observation as a hopelessly flawed approach to assessing teacher effectiveness.

The ubiquity of “satisfactory” ratings stands in contrast to a rapidly growing body of research that examines differences in teachers’ effectiveness at raising student achievement. In recent years, school districts and states have compiled datasets that make it possible to track the achievement of individual students from one year to the next, and to compare the progress made by similar students assigned to different teachers. Careful statistical analysis of these new datasets confirms the long-held intuition of most teachers, students, and parents: teachers vary substantially in their ability to promote student achievement growth.

The quantification of differences has generated a flurry of policy proposals to promote teacher quality over the past decade, and the Obama administration’s recent Race to the Top program only accelerated interest. Yet, so far, little has changed in the way that

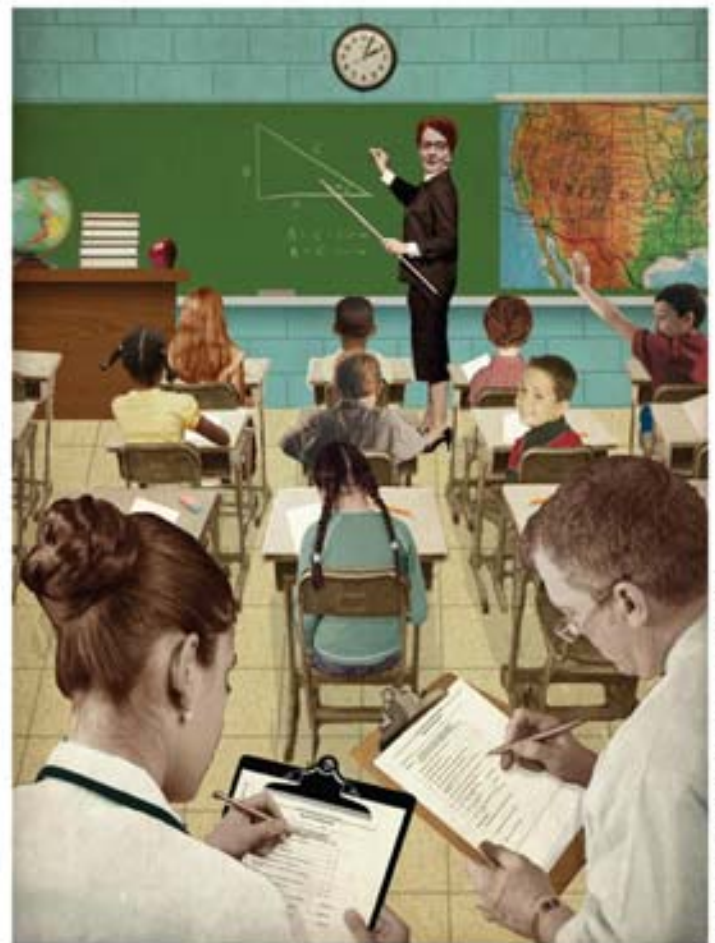


ILLUSTRATION / MICHAEL WARAKSA

teachers are evaluated, in the content of pre-service training, or in the types of professional development offered. A primary stumbling block has been a lack of agreement on how best to identify and measure effective teaching.

A handful of school districts and states—including Dallas, Houston, Denver, New York, and Washington, D.C.—have begun using student achievement gains as indicated by annual test scores (adjusted for prior achievement and other student characteristics) as a direct measure of individual teacher performance. These student-test-based measures are often referred to as “value-added” measures. Yet even supporters of policies that make use of value-added measures recognize the limitations of those measures. Among the limitations are, first, that these performance measures can only be generated in the handful of grades and subjects in which there is mandated annual testing. Roughly one-quarter of K–12 teachers typically teach in grades and subjects where obtaining such measures is currently possible. Second, test-based measures by themselves offer little guidance for redesigning teacher training or targeting professional development; they allow one to identify particularly effective teachers, but not to determine the specific practices responsible for their success. Third, there is the danger that a reliance on test-based measures will lead teachers to focus narrowly on test-taking skills at the cost of more valuable academic content, especially if administrators do not provide them with clear and proven ways to improve their practice.

Student-test-based measures of teacher performance are receiving increasing attention in part because there are, as yet, few complementary or alternative measures that can provide reliable and valid information on the effectiveness of a teacher’s classroom practice. The approach most commonly in use is to evaluate effectiveness through direct observation of teachers in the act of teaching. But as “The Widget Effect” reports, such evaluations are a largely perfunctory exercise.

In this article, we report a few results from an ongoing study of teacher classroom observation in the Cincinnati Public Schools. The motivating research question was whether classroom observations—when performed by trained professionals external to the school, using an extensive set of standards—could identify teaching practices likely to raise achievement.

We find that evaluations based on well-executed classroom observations do identify effective teachers and teaching practices. Teachers’ scores on the classroom observation components of Cincinnati’s evaluation system reliably predict the achievement gains made by their students in both math and reading. These findings support the idea that teacher evaluation systems need not be based on test scores alone in order to provide useful information about which teachers are most effective in raising student achievement.

The Cincinnati Evaluation System

Jointly developed by the local teachers union and district more than a decade ago, the Cincinnati Public Schools’ Teacher Evaluation System (TES) is often cited as a rare

example of a high-quality evaluation program based on classroom observations. At a minimum, it is a system to which the district has devoted considerable resources. During the yearlong TES process, teachers are typically observed and scored four times: three times by a peer evaluator external to the school and once by a local school administrator. The peer evaluators are experienced classroom teachers chosen partly based on their own TES performance. They serve as full-time evaluators for three years before they return to the classroom. Both peer evaluators and administrators must complete an intensive training course and accurately score videotaped teaching examples.

The system requires that all new teachers participate in TES during their first year in the district, again to receive tenure (usually in their fourth year), and every fifth year thereafter. Teachers tenured before 2000–01 were gradually phased into the five-year rotation. Additionally, teachers may volunteer to be evaluated; most volunteers do so to post the high scores necessary to apply for selective positions in the district (for example, lead teacher or TES evaluator).

The TES scoring rubric used by the evaluators, which is based on the work of educator Charlotte Danielson, describes the practices, skills, and characteristics that effective teachers should possess and employ. We focus our analysis on the two (out of four total) domains of TES evaluations that directly address classroom practices: “Creating an Environment for Student Learning” and “Teaching for Student Learning.” (The other two TES domains assess teachers’ planning and professional contributions outside of the classroom; scores in these areas are based on lesson plans and other documents included in a portfolio reviewed by evaluators.) These two domains, with scores based on classroom observations, contain more than two dozen specific elements of practice that are grouped into eight “standards” of teaching. Table 1 provides an example of two elements that comprise one standard. For each element, the rubric provides language describing what performance looks like at each scoring level: Distinguished (a score of 4), Proficient (3), Basic (2), or Unsatisfactory (1).

Data and

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

One sample standard from the Cincinnati evaluation rubric

Standard 3.2: Teacher demonstrates knowledge...

	Distinguished	Proficient	Basic	Unsatisfactory
Instructional Strategies	Teacher routinely uses a broad range of multiple instructional strategies that are effective and appropriate to the content.	Teacher uses instructional strategies that are effective and appropriate to the content.	Teacher uses a limited range of instructional strategies that are effective and appropriate to the content.	Teacher uses instructional strategies that are ineffective and/or inappropriate to the content.
Content Knowledge	Teacher conveys accurate content knowledge, including standards-based content knowledge.	Teacher conveys accurate content knowledge, including standards-based content knowledge.	Teacher conveys some minor content inaccuracies that do not contribute to making the content incomprehensible to the students.	Teacher conveys content inaccuracies that contribute to making the content incomprehensible to the students.

SOURCE: Cincinnati Public Schools Teacher Evaluation System 2005

Methodology

Cincinnati provided us with records of each classroom observation conducted between the 2000–01 and 2008–09 school years, including the scores that evaluators assigned for each specific practice element as a result of that observation. Using these data, we calculated a score for each teacher on the eight TES “standards” by averaging the ratings assigned during the different observations of that teacher in a given year on each element included under the standard. We then collapsed these eight standard-level scores into three summary indexes that measure different aspects of a teacher’s practice:

- The first, which we call Overall Classroom Practices, is simply the teacher’s average score across all eight standards. This index captures the general importance of the full set of teaching practices measured by the evaluation.
- The second, Classroom Management vs. Instructional Practices, measures the difference in a teacher’s rating on standards that evaluate classroom management and that same teacher’s rating on standards that assess instructional practices. A teacher who is more skilled at managing the classroom environment, as compared to her ability to engage in desired instructional activities, will receive a higher score on this index than a teacher who engages in these instructional practices but who is less skilled at managing the classroom.
- The third, Questions/Discussion vs. Standards/Content, measures the difference between a teacher’s rating on a single standard that evaluates the use of questions and classroom discussion as an instructional strategy, and that same teacher’s average rating on three standards that assess teaching practices that focus on classroom management routines, on conveying standards-based instructional objectives to students, and on demonstrating content-specific knowledge in teaching these objectives.

Our main analysis below examines the degree to which these summary indices predict a teacher’s effectiveness in raising student achievement. Note, however, that we did not construct the indices based on any hypotheses of our own about which aspects of teaching practice measured by TES were most likely to influence student achievement. Rather, we used a statistical technique known as principal components analysis, which identifies the smaller number of underlying constructs that the eight different dimensions of practice are trying to capture. As it turns out, scores on these three indices explain 87 percent of the total variation in teacher performance across all eight standards.

For all teachers in our sample, the average score on the Overall Classroom Practices index was 3.21, or between the “Proficient” and “Distinguished” categories. Yet one-quarter of teachers received an overall score higher than 3.53 and one-quarter received a score lower than 2.94. In other words, despite the fact that TES evaluators tended to assign relatively high scores on average, there is a fair amount of variation from teacher to teacher that we can use to examine the relationship between TES ratings and classroom effectiveness.

In addition to TES observation results, Cincinnati provided student data for the 2003–04 through 2008–09 school years, including information on each student’s gender, race/ethnicity, English proficiency status, participation in special education or gifted and

talented programs, class and teacher assignments by subject, and state test scores in math and reading. This rich dataset allows us to study students' math and reading test-score growth from year to year in grades four through eight (where end of year and prior year tests are available), while also taking account of differences in student backgrounds.

Our primary goal was to examine the relationship between teachers' TES ratings and their assigned students' test-score growth. This task is complicated, however, by the possibility that factors not measured in our data, such as the level of social cohesion among the students or unmeasured differences in parental engagement, could independently affect both a TES observer's rating and student achievement. To address this concern, we use observations of student achievement from teachers' classes in the one or two school years prior to and following TES measurement, but we do not use student achievement gains from the year in which the observations were conducted. (If some teachers are assigned particularly engaged or cohesive classrooms year after year, the results could still be biased; this approach, however, does eliminate bias due to year-to-year differences in unmeasured classroom traits being related to classroom observation scores.)

We restrict our comparisons to teachers and students within the same schools in order to eliminate any potential influence of differences between schools on both TES ratings and student achievement. In other words, we ask whether teachers who receive higher TES ratings than other teachers in their school produce larger gains in student achievement than their same-school colleagues.

Results

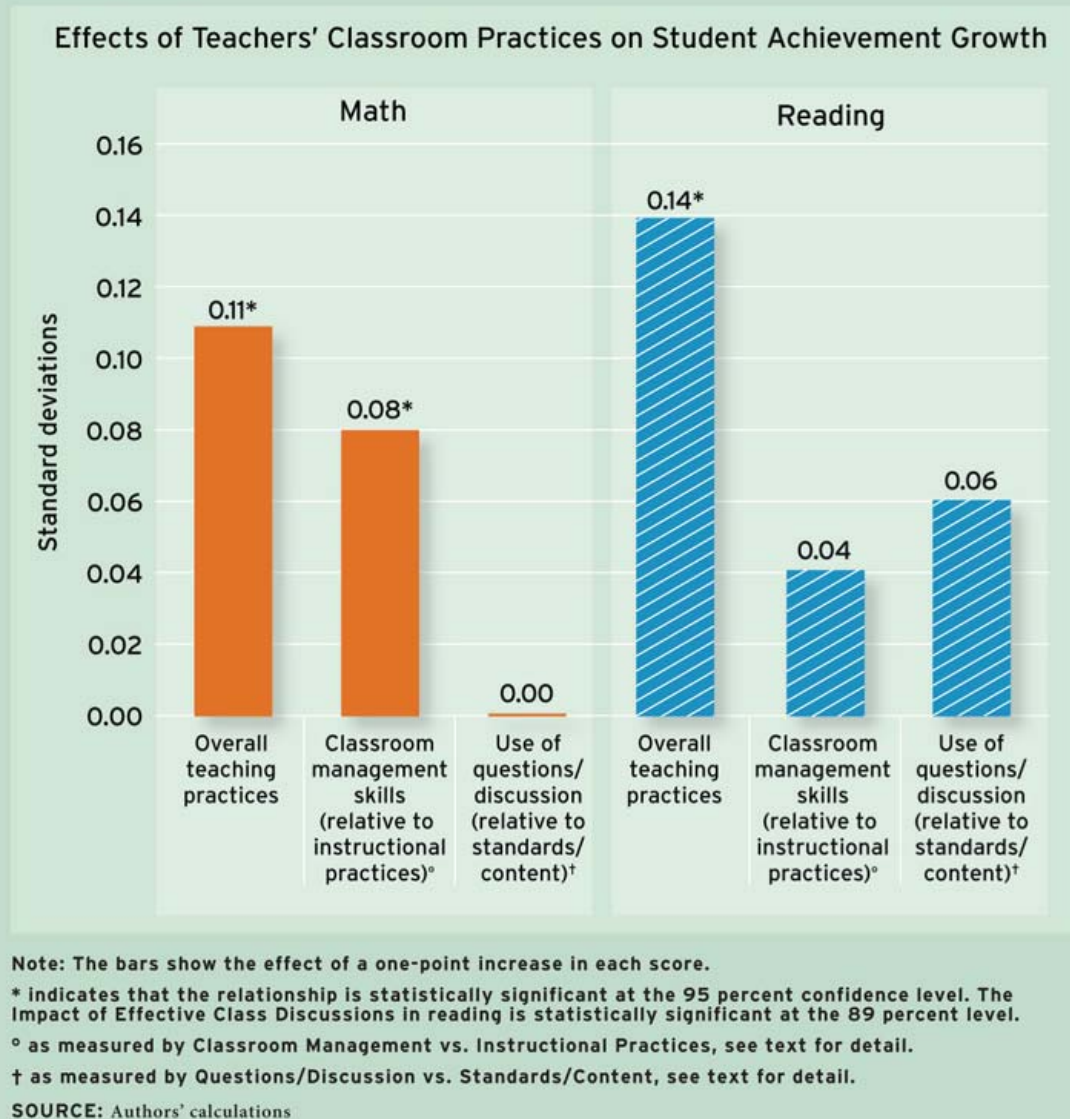
We find that teachers' classroom practices, as measured by TES scores, do predict differences in student achievement growth. Our main results, which are based on a sample of 365 teachers in reading and 200 teachers in math, indicate that improving a teacher's Overall Classroom Practices score by one point (e.g., moving from an overall rating of "Proficient" [3] to

"Distinguished" [4]) is associated with one-seventh of a standard deviation increase in reading achievement, and one-tenth of a standard deviation increase in math (see Figure 1).

The specific point system that TES uses to rate teachers as Proficient and Distinguished is somewhat arbitrary. For a better sense of the magnitude of these estimates, consider a student who begins the year at the 50th percentile and is assigned to a top-quartile teacher as measured by the Overall Classroom Practices score; by the end of the school year, that student, on average, will score about three percentile points higher in reading and about two points higher in math than a peer who began the year at the same achievement level but was assigned to a bottom-quartile teacher.

Evaluations Identify Good Teachers (Figure 1)

In Cincinnati, teachers' ratings by classroom observers predict how much their students learn.



This difference might not seem large but, of course, a teacher is just one influence on student achievement scores (and classroom observations are only one way to assess the quality of a teacher's instruction). By way of comparison, we can estimate the total effect a given teacher has on her students' achievement growth; that total effect includes the practices measured by the TES process along with everything else a teacher does. The difference between being taught by a top-quartile total-effect teacher versus a bottom-quartile total-effect teacher would be about seven percentile points in reading and about six points in math (see Figure 2). This total-effect measure is one example of the kind of "value-added" approach taken in current policy proposals.

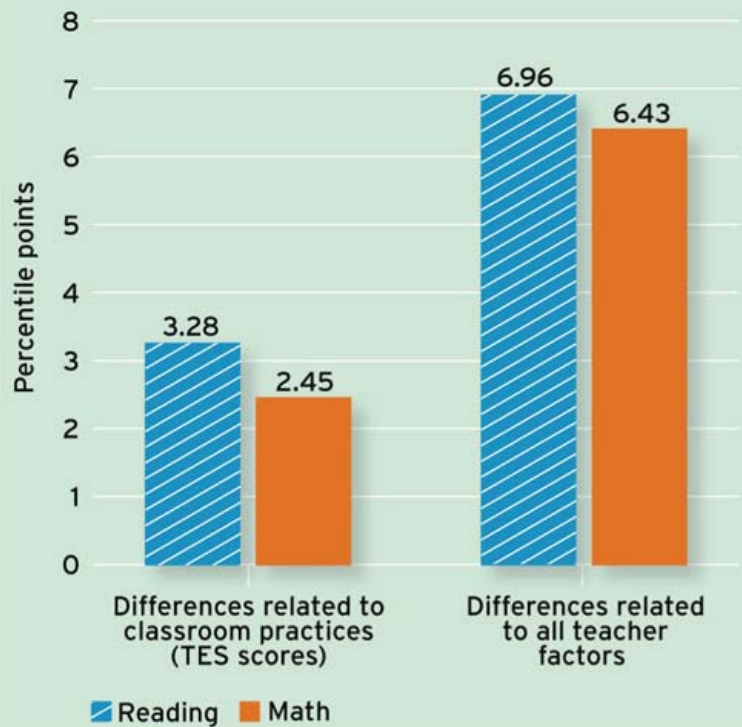
From these data, we can also discern relationships between more specific teaching practices and student outcomes across academic subjects (see Figure 1). Among students assigned to different teachers with the same

Overall Classroom Practices score, math achievement will grow more for students whose teacher is better than his peers at classroom management (i.e., has a higher score on our Classroom Management vs. Instructional Practices measure). We also find that reading scores increase more among students whose teacher is relatively better than his peers at engaging students in questioning and discussion (i.e., has a high score on Questions/Discussion vs. Standards/Content). This does not mean, however, that students' math achievement would rise if their teachers were to become worse at a few carefully selected instructional practices. Although this might raise their Classroom Environment vs. Instructional Practices score it would also lower the Overall Classroom Practices score, and any real teacher is the combination of these three scores.

How Much Can Classroom Observations Explain? (Figure 2)

Observed classroom practices capture just under half of the measurable differences in teacher effectiveness.

Differences in the achievement of students assigned to a top-quartile versus a bottom-quartile teacher



Bars show the estimated effect on the median student of being assigned to a top-quartile rather than a bottom-quartile teacher. Classroom practices differences based only on Overall Classroom Practices score.

SOURCE: Authors' calculations

Do these statistics provide any insight that teachers can use to focus their efforts? First, our finding that Overall Classroom Practices is the strongest predictor of student achievement in both subjects indicates that improved practice in any of the areas considered in the TES process should be encouraged. In other words, the practices captured by the TES rubric do predict better outcomes for students. If, however, teachers must choose a smaller number of practices on which to focus their improvement efforts (for example, because of limited time or professional development opportunities), our results suggest that math achievement would likely benefit most from improvements in classroom management skills before turning to instructional issues. Meanwhile, reading achievement would benefit most from time spent improving the practice of asking thought-provoking questions and engaging students in discussion.

Can we be confident that the various elements of practice measured by TES are the reasons that students assigned to highly rated teachers make larger achievement gains? Skeptical readers may worry that better teachers engage in more of the practices encouraged by TES, but that these practices are not what make the teacher more effective. To address this concern, we take advantage of the fact that some teachers were evaluated by TES multiple times. For these teachers, we can test whether improvement over time in the practices measured by TES is related to improvement in the achievement gains made by the teachers' students. This is exactly what we find. Since this exercise compares each teacher only to his own prior performance, we can be more confident that it is differences in the use of the TES practices themselves that promote student achievement growth, not just the teachers who employ these strategies.

Conclusion

Is TES worth the considerable effort and cost? Does the intensive TES process (with its multiple observations and trained peer evaluators) produce more accurate information on teachers' effectiveness in raising student achievement gains than do more-subjective evaluations? In fact, studies of informal surveys of principals (see "[When Principals Rate Teachers](#)," research, Spring 2006) and teacher ratings by mentor teachers find that these more-subjective evaluation methods have similar power to detect differences in teacher effectiveness as the TES ratings. These studies may lead some to question the need for the more detailed TES process. We contend, however, that evaluations based on observations of classroom practice are valuable, even if they do not *predict* student achievement gains considerably better than more subjective methods like principal ratings of teachers.

The additional information the TES system provides can be used in several important ways. First, the data gleaned from the observations allow researchers to connect specific teaching practices with student achievement outcomes, providing evidence of effective teaching practices that can be widely shared.

The TES program also has the advantage of furnishing teachers and administrators with details about the specific practices that contributed to each teacher's score. The descriptions of practices, and different performance levels for each practice, that comprise

the TES rubric can help teachers and administrators map out professional development plans. A school administrator who desires to differentiate the support she provides to individual teachers would benefit from knowing the components of each teacher's overall scores. A teacher who would like to improve his classroom management skills may find that he has scored relatively low in a particular standard, and then take steps to improve his practice in response to that information.

Finally, scoring individual practices allows for understanding of more fine-grained variations in skill among teachers with similar overall ratings. It is notable, especially given "The Widget Effect" study, that nearly 90 percent of teachers in our sample received an overall "Satisfactory" rating (i.e., "Distinguished" or "Proficient" in Cincinnati's terms). Still, there are readily discernible differences in mastery of specific skills within that 90 percent, and those differences in skills predict differences in student achievement.

There are other aspects of the Cincinnati system that may or may not account for the results we observed. First, the observers were external to the school and, in most cases, had no personal relationship with the person they were observing. Second, the observers were trained beforehand and were required to demonstrate their ability to score some sample videos in a manner consistent with expert scores. Simply handing principals a checklist with the same set of standards may not lead to a similar outcome.

The results presented here constitute the strongest evidence to date on the relationship between teachers' observed classroom practices and the achievement gains made by their students. The nature of the relationship between practices and achievement supports teacher evaluation and development systems that make use of multiple measures. Even if one is solely interested in raising student achievement, effectiveness measures based on classroom practice provide critical information to teachers and administrators on what actions they can take to achieve this goal.

Thomas J. Kane is professor of education and economics at the Harvard Graduate School of Education. Eric S. Taylor is a doctoral student at the Stanford University School of Education. John H. Tyler is associate professor of education, economics, and public policy at Brown University. Amy L. Wooten is a doctoral student at the Harvard Graduate School of Education. Reflecting equal contributions to this work, authors are listed alphabetically. This article is based in part on a larger study which is forthcoming in the Journal of Human Resources.

10
tweets

retweet

[User Agreement](#) | [Privacy Policy](#)

[Reporting Copyright Infringement](#) | [Guidelines for Submissions](#) | [Permissions](#) | [FAQ](#)

4/26/2011

Evaluating Teacher Effectiveness : Edu...

Web-only content Copyright © 2011 President & Fellows of Harvard College

Journal content Copyright © 2011 by the Board of Trustees of Leland Stanford Junior University

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