National Institute for Excellence in Teaching

RESEARCH BRIEF | August | 2010

National Institute for Excellence in Teaching®

A Teacher Evaluation System That Works

Summary

Teachers are the most important school-related factor for student achievement gains, but evaluation of teacher performance is seldom conducted in any rigorous way. As policymakers call for a better approach to teacher evaluation, the 10-year history of TAP™: The System for Teacher and Student Advancement provides an example of an integrated system for teacher evaluation and support. TAP teachers are evaluated every year through multiple classroom observations by trained and certified raters and through their contributions to student achievement growth. Based on data from TAP schools, research shows that:

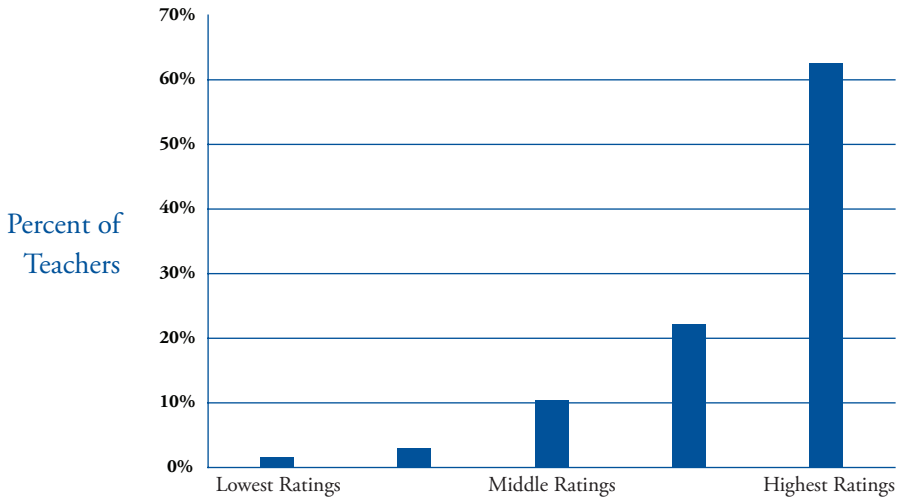
- TAP teacher evaluations provide differentiated feedback on teacher performance.
- TAP classroom evaluations are aligned with value-added student achievement outcomes.
- TAP teachers become more effective over time.
- TAP schools show higher retention of more effective teachers, and higher turnover of less effective teachers.

Creating the capacity for evaluation and evaluation-guided improvement in schools requires the right tools as well as the sustained engagement of teachers and leaders. The example of TAP implies that teacher evaluation should not be pursued as a one-time, one-size-fits-all policy prescription, but should be integrated within a comprehensive, site-based system with specific practical elements to support teachers and improve teaching and learning in the classroom.

Background

Teachers are the most important school-related factor impacting student achievement gains. However, evaluation of teacher performance is seldom conducted in any rigorous way. As shown in Figure 1, evaluations commonly rate most teachers at the highest level of performance despite the fact that schools are not educating their students at the highest levels.[1]

Figure 1. Teacher Evaluations in Urban School Districts



Based on data from Weisberg et al., (2009). Scores on 3-point and 4-point scales have been interpolated to a 5-point scale using a cumulative probability density function based on the reported data.

In this context, the Obama administration has made the evaluation of teacher effectiveness a key piece of education reform. The American Recovery and Reinvestment Act of 2009 (P.L. 111-5) requires states to take actions to improve teacher effectiveness by 2011. In a letter from Secretary of Education Arne Duncan to state governors, he specified that states must report the number and percent of teachers and principals rated at each performance level in each local educational agency's evaluation system, and the number and percent of those teacher and principal evaluation systems that require evidence of student achievement outcomes.[2]

As policymakers call for a better approach to teacher evaluation, the 10-year history of TAP™: The System for Teacher and Student Advancement provides an example of an integrated system for teacher evaluation and support to improve teacher effectiveness and student achievement. The TAP system consists of four interrelated elements:

- Multiple Career Paths
- Ongoing Applied Professional Growth
- Instructionally Focused Accountability
- Performance-Based Compensation

1. Weisberg, D., Sexton, S., Mulhern, J., Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* Brooklyn: The New Teacher Project. Available online at http://widgeteffect.org/
2. Duncan, Arne. (2009). Letter to State Governors regarding the American Recovery and Reinvestment Act. Accessed online at http://www2.ed.gov/programs/statestabilization/2009-394-cover.pdf

The accountability component of TAP, i.e., teacher evaluation, is aligned to each of the other elements. TAP evaluations provide feedback to guide professional development and serve as the basis for determining performance-pay awards. TAP relies on trained, expert master and mentor teachers as well as principals to carry out multiple classroom evaluations a year using a research-based rubric, and to provide personalized support for improvement based on these assessments.

Research Findings

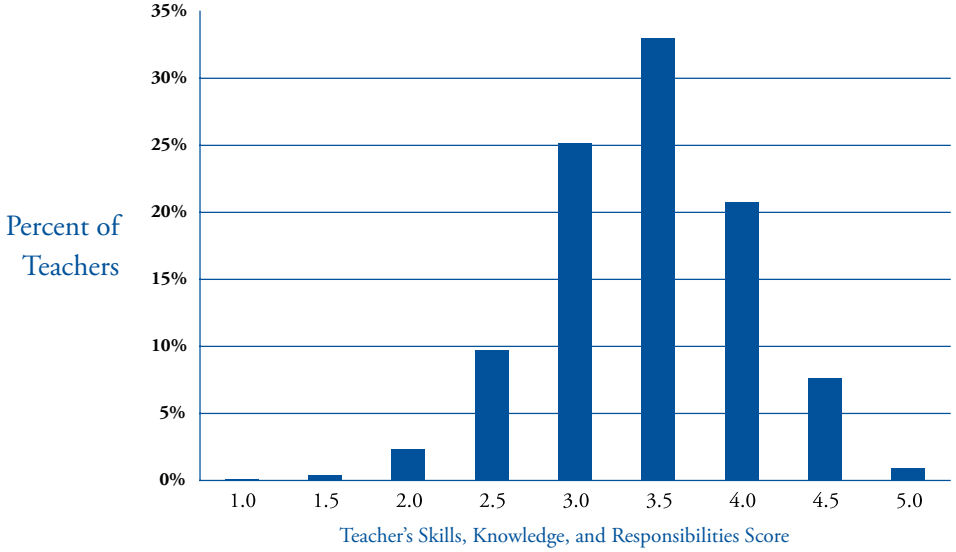
Researchers at the National Institute for Excellence in Teaching (NIET), which manages TAP, have reported on the TAP evaluation structure and its statistical properties in an NIET Working Paper.[3] The study used nationwide TAP evaluation data including value-added scores from the 2006-07 and 2007-08 school years,[4] for a sample of 1,830 teacher-level records. Analysis of classroom evaluation scores and retention outcomes drew from a sample of 7,377 teacher-level records from the 2004-05 to 2008-09 school years.

The findings of the study are as follows:

1. TAP teacher evaluations provide differentiated feedback on teacher performance.

The TAP teacher evaluation structure includes four or more classroom evaluations each year by trained and certified observers using research-based instructional quality rubrics. These evaluations result in a Skills, Knowledge, and Responsibilities (SKR) score on a ***1-5*** scale, with ***3*** representing proficient performance that still has room for improvement. The scores are averaged over the year for a final SKR score for each teacher. The mean SKR score for TAP teachers nationwide is ***3.5*** out of ***5***, significantly different from other evaluation systems nationwide that rate few teachers below the top level of performance. The scores of TAP teachers follow a mound-shaped distribution, as shown in Figure 2, which much more closely matches what we know about how teachers vary in effectiveness than does the inflated distribution shown in Figure 1. The TAP teacher evaluation system offers more useful feedback to teachers and administrators than evaluation structures that uniformly assign high ratings irrespective of a teacher's actual performance.

Figure 2: Differentiated Teacher Evaluations in TAP



n=7,377 teacher-level records from 2005-05 through 2008-09.

The next finding addresses whether these differentiated results are connected with measures of student learning.

3. Daley, G., and Kim, L. (2010). A teacher evaluation system that works. NIET Working Paper. Santa Monica, CA: The National Institute for Excellence in Teaching. Available online at http://www.tapsystem.org/publications/wp_eval.pdf
4. Teacher-level value-added data for later years were incomplete at the time of doing the analysis.

2. TAP classroom evaluations are aligned with value-added student achievement outcomes.

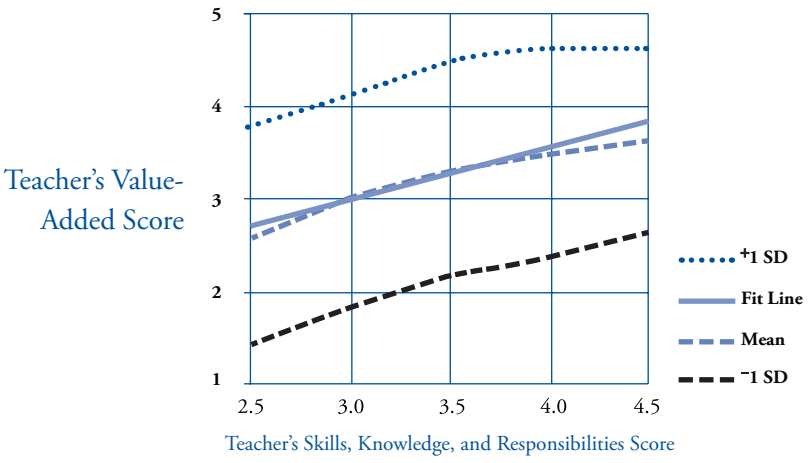
A higher quality of instruction in the classroom would be expected to lead to greater student gains on standardized achievement tests. Our analysis reveals a strong relationship between observed teacher evaluation ratings (SKR) and value-added measures of student learning.

Value-added assessment is a method for measuring the contribution of teachers or schools to the growth in their students' academic achievement during a school year. This method involves matching each student's test scores to his or her own previous scores in order to measure individual growth. Through value-added assessment, the impact of a school year on a student's learning can be separated from the student's prior experiences in and out of school, as well as the student's individual characteristics such as demographics, socioeconomic status, and family conditions.[5]

The TAP value-added component provides each teacher with a classroom score showing the teacher's average student gain during the school year. The majority of value-added calculations used by TAP schools are performed by a single, independent value-added provider. For TAP teacher evaluations, these statistics are converted to a 5-point scale: a ***1*** represents significantly lower than one year of classroom-average student growth as compared to classrooms of students with similar previous achievement, a ***3*** represents one year of expected academic growth for similar students, and a ***5*** represents significantly higher than one year of growth for similar students.

There is a wide distribution of these value-added scores at each point on the SKR scale, as shown by the dashed lines in Figure 3, representing the mean value added and one standard deviation (SD) above and below the mean. This is one reason to combine complementary measures of performance rather than relying entirely on either observations or value-added results to determine teacher effectiveness or performance-based compensation. However, the overall relationship between teachers' value-added scores and their SKR scores is significant and positive, as shown by the slope of the straight line ("fit line") in Figure 3, representing the best statistical estimate of the true relationship, i.e. the model that best fits the data. In other words, higher SKR scores for teachers during the school year are associated with higher value-added scores for their students at the end of the year. This confirms that TAP classroom evaluations are aligned with value-added assessments of teacher performance in terms of student learning.

Figure 3. Relationship between SKR and Value-Added, Simple Regression Model

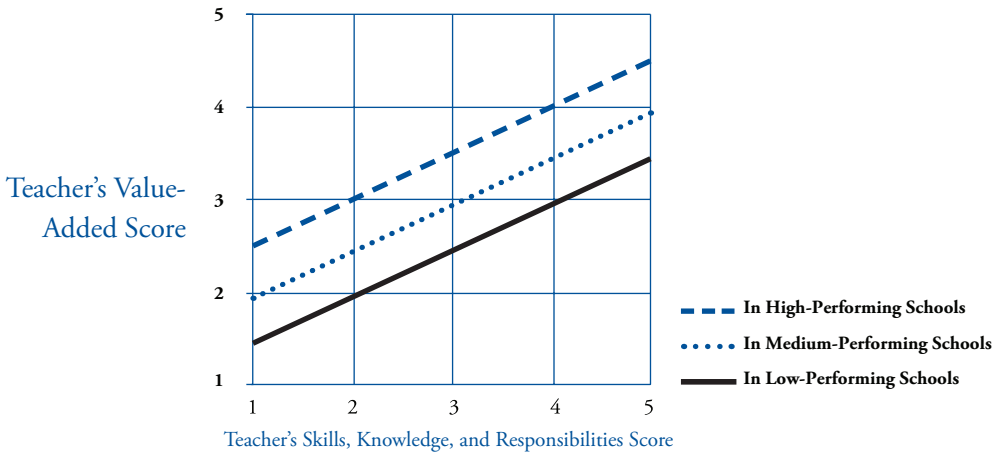


Scores are from TAP schools for the 2006-2007 and 2007-2008 school years. n = 1780 teachers

5. For more information about value-added assessment, please see http://www.tapsystem.org/policyresearch/policyresearch.taf?page=valueadded

This finding of a strong relationship between SKR and value-added measures holds up through a variety of statistical models controlling for school characteristics and schoolwide performance. A hierarchical linear model that includes schoolwide value added (i.e., the average student growth for the whole school as opposed to a teacher's classroom) shows that teachers in high-performing schools are more likely to have higher individual value added than others with the same SKR at low-performing schools (Figure 4). This result is meaningful because schoolwide value added is *not* simply the aggregate of teacher value added. In value-added modeling, teachers are compared with other teachers who have similar students, and schools are compared with other schools attended by similar students. Furthermore, only teachers in tested grades and subjects, with enough students who have previous test score histories in the data set, can receive classroom value-added scores. The schoolwide value-added score is more inclusive. The fact that this analysis yields significant positive results suggests that there is indeed a schoolwide effect separate from the aggregate of teacher effects. This point is consistent with the TAP concept that teachers' overall impact on student learning is raised by site-based collaborative professional growth and accountability under the instructional leadership of master and mentor teachers as well as school principals.

Figure 4. Relationship between SKR and Value-Added, Hierarchical Model Fit Lines



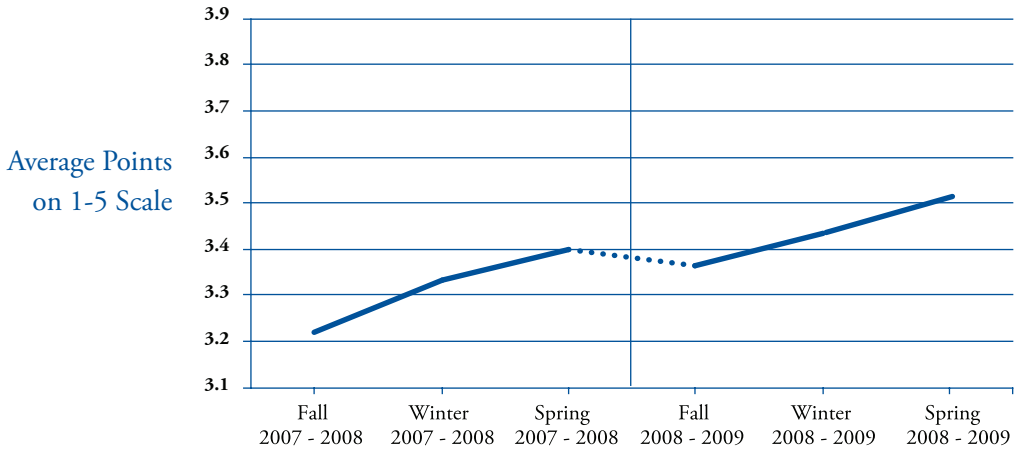
High-Performing Schools: n=682 teachers, with schoolwide value added 4 or 5
Medium-Performing Schools: n=649 teachers, with schoolwide value added 3
Low-Performing Schools: n=449 teachers, with schoolwide value added 1 or 2
Scores are from 2006-2007 and 2007-2008 school years.

These results provide an important validation of TAP teacher evaluations. When teachers demonstrate strong instructional skills as measured by classroom observations, their students show higher academic growth regardless of previous achievement and socioeconomic status.

3. TAP teachers become more effective over time.

The study also investigated evidence regarding whether teachers' performance in TAP schools improves over time, both individually as a result of professional growth and across schools as a result of retention of more effective teachers. TAP teachers demonstrate steady improvement in observed skills during the course of the school year. Figure 5 shows the improvement in instructional quality scores over a two-year period. In the data shown, despite a slight dip over the summer while away from school, teachers demonstrated an overall path of improvement that continued over both years.

Figure 5. Improvement in Observed Teacher Skills, 2007-08 and 2008-09



Average of Instructional SKR indicators for 2007-2009 cohort (n = 650 teachers)

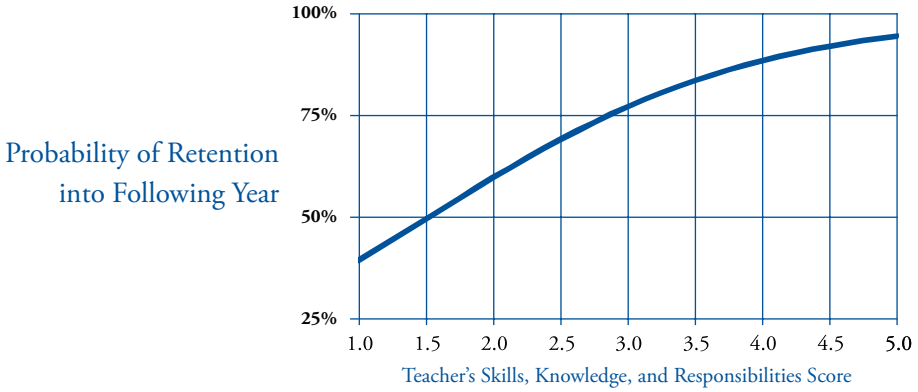
This graph is based on a sample including all TAP schools during the years 2007-09. We tracked a cohort of 650 teachers through observations grouped into six periods in the fall, winter and spring of the 2007-08 and 2008-09 school years. The cohort was composed of teachers working in TAP schools both years, with observations in each of the observation periods. Teachers present in only one school year or whose observations clustered around the same time frame during a year were excluded from the sample. The graph shows consistent growth during each school year as well as growth from one year to the next.

This result for SKR scores taken at multiple points during a two-year period is confirmed by also looking at annual SKR scores taken over multiple years. For the school years 2005-06 through 2008-09, for teachers with matched records over consecutive years (n=4,882 teacher-level records by year), TAP teachers' SKR scores improved on average by a significant sixth of a point per year on the ***1-5*** SKR scale. Importantly, teachers with previous-year SKR scores less than ***3*** demonstrated the most growth, averaging more than a half-point increase per year on the ***1-5*** SKR scale. TAP helps develop less effective teachers into more effective teachers through ongoing, applied professional growth informed by rigorous evaluations.

4. TAP schools show higher retention of more effective teachers, and higher turnover of less effective teachers.

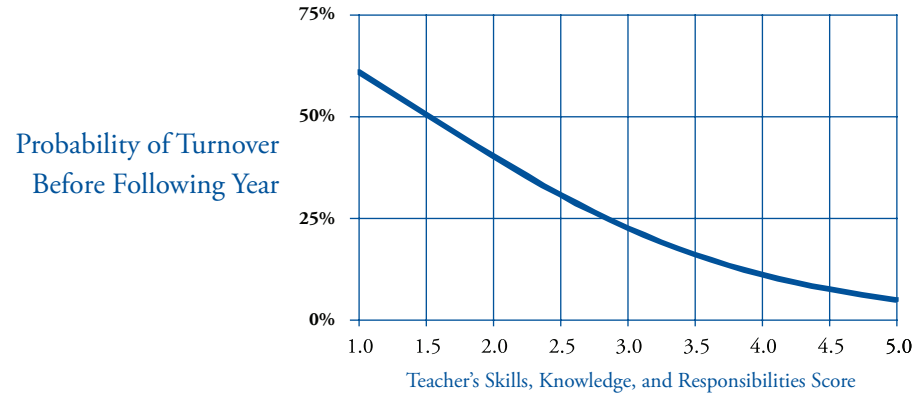
In addition to the impact of individual teacher growth, the quality of the teaching staff at TAP schools improves over time as a result of differences in the retention and turnover of teachers related to their instructional effectiveness. As illustrated in Figure 6, for each point higher that a teacher's SKR score is in one year, the teacher's odds of remaining in a TAP school the following year increases by 87%. Figure 7 shows the same relationship, inverted to emphasize that teachers with lower classroom evaluation scores are more likely to leave a TAP school.

Figure 6. Relationship between Teacher Evaluation Ratings and Retention



n = 7,377 teacher-level records by year from 2005 through 2009
Retention includes teachers who stayed in TAP schools, including teachers who became TAP master and mentor teachers. Retention does not include teachers who became administrators, moved to non-TAP schools, or left teaching.

Figure 7. Relationship between Teacher Evaluation Ratings and Turnover



n = 7,377 teacher-level records by year from 2005 through 2009
Turnover includes teachers who became administrators, moved to non-TAP schools or left teaching. Turnover does not include teachers who stayed in TAP schools, including teachers who became TAP master and mentor teachers.

This difference in teacher retention as it relates to quality is consistent with the theory that the TAP system motivates good teachers to stay, while giving less effective teachers *both* an opportunity to improve *and* an incentive not to stay where they are less likely to receive high ratings and bonuses. Since observational ratings are correlated with student value added, this will result in a more effective teaching staff and greater student growth over time.

National Institute for Excellence in Teaching

Implications of the Findings for Policy and Practice

As educators and policymakers work to improve the quality of education in American schools, a central focus of their efforts is the evaluation of teachers. Although teacher evaluation by itself is sometimes criticized as arbitrary, one-dimensional, undifferentiated, disconnected from the needs of students, and unaligned with professional development opportunities for improvement, this study shows that a well-designed, integrated system can be objective, rigorous, differentiated, multidimensional, linked to student learning and supportive of teacher improvement.

Underlying these abstractions are many concrete details of design and implementation, as described in the NIET Working Paper from this study. Creating the capacity for teacher evaluation and evaluation-guided instructional improvement in schools requires the right tools as well as the sustained engagement of teachers and leaders. The example of TAP implies that teacher evaluation should not be pursued as a one-time, one-size-fits-all policy prescription, but should be integrated within a comprehensive, site-based system with specific practical elements to support teachers and improve teaching and learning in the classroom.

For the NIET Working Paper on this study, please see:
http://www.tapsystem.org/publications/wp_eval.pdf

For additional information on TAP, please visit:
http://www.tapsystem.org

The National Institute for Excellence in Teaching (NIET) was established in 2005 as an independent 501(c)(3) public charity. With a staff experienced in teaching, school leadership, program evaluation, research, and business management, NIET operates TAP and works to ensure the system's effectiveness and sustainability. NIET researchers study the design, operations, and impact of TAP in the context of other research literature and public policy perspectives.

National Institute for Excellence in Teaching
1250 Fourth Street
Santa Monica, CA 90401
310-570-4860