More than Measurement

The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems

by Craig D. Jerald and Kristan Van Hook

January 2011

The Joyce Foundation

National Institute for Excellence in Teaching®

More than Measurement

The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems

by Craig D. Jerald and Kristan Van Hook

January 2011

Funding Provided By

The Joyce Foundation

Message from the NIET President

Effective teachers are central to assuring excellence and rigor in the educational experience of every young person in America. TAP: The System for Teacher and Student Advancement is a comprehensive, research-based reform designed to develop a corps of highly effective teachers and principals for America's schools. By generating an environment with powerful and sustained opportunities for career advancement, professional growth, teacher accountability, and competitive compensation, TAP is transforming the way schools, districts, and states support their most important asset – human capital.

TAP was developed by Lowell Milken and colleagues at the Milken Family Foundation and is now managed and operated by the National Institute for Excellence in Teaching. Over the course of the last ten years, we have developed and refined our approach to teacher evaluation as a key element of TAP's comprehensive system for improving teacher effectiveness. As potential reforms to teacher evaluation systems receive an increasing level of focus and attention, we recognize that we have a great deal to add to the conversation.

I would like to thank Craig Jerald for immersing himself in the details of the TAP evaluation system. He spent extensive time with numerous TAP evaluators, as well as those responsible for the implementation and training. Working with NIET staff and educators in TAP schools, districts, and states, Mr. Jerald brought a powerful outside perspective to identifying essential design and implementation decisions that must be considered in any evaluation system reform.

We look forward to the discussions that we hope this paper will provoke, and to working with others to ensure that teacher evaluation systems effectively measure teacher performance and support teachers in improving their instructional skills. We firmly believe that we must focus equally on both of these goals to transform teaching and create stimulating professional learning communities where educators and students alike will thrive.

Gary Stark
President and CEO
National Institute for Excellence in Teaching

Table of Contents

Executive Summary

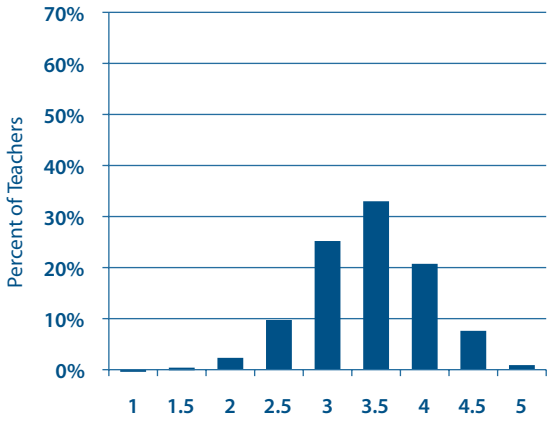
Teacher evaluation has emerged as a major focus for reform at the highest levels of education policymaking, and for good reason. Most evaluations are based on scant evidence of actual effectiveness, produce inflated ratings, and provide teachers with little useful feedback.

This paper offers policymakers and practitioners important "lessons learned" from TAP: The System for Teacher and Student Advancement. The TAP system is a comprehensive strategy to boost teacher effectiveness through opportunities for career advancement, professional growth, performance evaluation, and competitive compensation. TAP represents the longest sustained and most successful effort to radically transform teacher evaluation using multiple measures, including student achievement gains, in the United States today. Since its inception in 2000-2001, the TAP system has grown to serve over 10,000 teachers and 100,000 students across the country, and it will expand to reach twice that many teachers and students by the 2011-2012 school year.

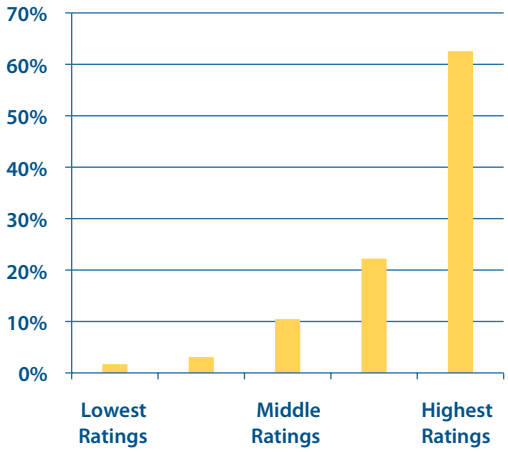
TAP uses a comprehensive approach to teacher evaluation which considers multiple measures of performance including student achievement gains in addition to teachers' instructional practices. Recent analyses have shown that TAP's evaluation system produces much more valid performance ratings than do traditional teacher evaluations. TAP's observational measure of classroom instruction represents a true distribution of teacher performance *even before* "value-added"[1] student learning gains are calculated.

Observational Ratings in TAP Schools versus Urban Districts with Traditional Evaluation Systems

Observational (SKR) Ratings in TAP Schools Before Value-Added Scores Are Calculated



Observational Ratings in Urban Districts with Non-Binary Scales Reported in *The Widget Effect*[2]



NOTE: TAP ratings include "Skills, Knowledge, and Responsibilities (SKR) " scores only, not value-added results. For *The Widget Effect* districts, scores on 3-point and 4-point scales were interpolated to a 5-point scale using a cumulative probability density function based on the reported data.

LESSON 1: Identify specific goals for teacher evaluation that can guide difficult system design decisions.

Any attempt to design a complex system (whether for teacher evaluation or school accountability or education finance) inevitably entails trade-offs among many possible design features, all of which might seem desirable in the abstract. Gaining clarity on specific goals for the system will help policymakers make thoughtful choices about difficult trade-offs when it comes time to actually design and implement a new teacher evaluation system.

TAP's system of teacher evaluation was intentionally designed to serve two equally important goals: accurately *measure* teachers' effectiveness for the purpose of making performance and personnel decisions, and provide teachers with intensive support to *improve* their performance over time. Neither goal can be achieved merely though lip service; both require targeted investments in specific mechanisms.

LESSON 2: Use multiple, complementary measures—including student achievement gains—to evaluate teachers.

TAP evaluates teachers based on multiple indicators of performance that take into account *both* teaching practices *and* teaching outcomes. Teachers are observed four to six times throughout the school year by multiple trained and certified evaluators who consider 19 areas of effective instructional practice for an overall "Skills, Knowledge, and Responsibilities (SKR)" score. Then, in grades and subjects where classroom-level student learning gains can be reliably calculated, each teacher also receives a value-added score that measures the teacher's impact on student learning growth. It is essential to look at student learning *gains* rather than attainment levels, in order to assess teacher effectiveness separate from other factors, as well as to ensure that teachers of all levels of students have an equal opportunity to demonstrate effectiveness.

For multiple measures to work in a teacher evaluation system, such measures should be *different yet complementary*. Across TAP schools, SKR scores and value-added scores are significantly correlated: When teachers demonstrate strong instructional skills as measured by the TAP observation methods and Rubric, their students show higher academic growth regardless of previous achievement and socioeconomic status. At the same time, the measures are never perfectly correlated because they are measuring different aspects of teacher performance, each of which is an important measure in its own right.

While it is important to consider student learning when evaluating teacher performance, it would be a mistake to consider value-added or other learning growth measures some kind of panacea for the problem of inflated evaluation ratings. If "multiple measures" are to work, then *all of the measures, including classroom observations or other qualitative measures, must be valid and rigorous in and of themselves*. TAP's experience proves this to be an achievable policy goal, but it requires a significant investment in time and resources.

LESSON 3: Invest sufficiently in "wrap-around" quality control mechanisms.

Advocates for improving teacher evaluation systems need to be aware that producing valid performance ratings requires more than simply adopting a new set of tools and procedures and providing some initial training. The TAP system works hard to ensure valid, non-inflated ratings through an extensive set of quality control mechanisms that "wrap around" the school year—taking place before, during, and after teachers are evaluated.

Before evaluating teachers, administrators and teacher leaders must successfully complete four days of TAP evaluation training and pass a performance-based certification assessment, and all evaluators must be recertified annually. During the school year, evaluators meet monthly to discuss issues related to evaluation and to identify potential problems with inter-rater reliability and score inflation. The National Institute for Excellence in Teaching (NIET) has developed a Comprehensive Online Data Entry (CODE) system that stores evaluation results and can produce a range of tables and charts for analyzing inter-rater reliability and validity. At the end of the year, schools compare SKR scores to value-added scores to review correlation between the measures and to analyze areas for improvement revealed by each measure.

LESSON 4: Train evaluators to conduct in-depth post-conferences that can help teachers improve their effectiveness.

In TAP schools, evaluators meet with teachers within 48 hours after an observation has taken place to analyze the lesson and provide teachers with detailed feedback. Such "post-conferences" cannot be treated as a bureaucratic formality to be checked off a list. In fact, they are one of the most critical features of an effective teacher evaluation system if the goal is not just to measure the quality of teaching, but also to improve it.

Post-conferences last from 40 minutes to one hour and provide teachers with two kinds of explicit feedback in addition to formal scores—one "area for reinforcement" and one "area for refinement," each of which is tied to a specific indicator on the TAP Rubric. Evaluators and teachers analyze how a particular strength of the lesson contributed to student learning and discuss how the teacher can continue to build on that area of strength in future lessons (reinforcement). Then they analyze an element of the lesson that could have been improved, thus better contributing to student learning, and discuss how the teacher can work to improve that area in future lessons (refinement).

NIET's training for evaluators teaches them how to ensure that such post-conferences are evidence-based and analytical. Principals and other school leaders must demonstrate that they can plan effective post-conferences in order to become certified evaluators.

LESSON 5: Look for ways to provide teachers with targeted follow-up support.

In TAP schools, evaluators have a variety of ways to follow up with teachers after observations and post-conferences, including "real-time coaching" through demonstration lessons, modeling, or team teaching during actual classtime in the teacher's own classroom.

Obviously, such intensive follow-up support is much easier to provide in schools that have adopted all elements of the full TAP system model—one in which a cadre of master and mentor teachers is granted explicit authority and responsibility for a range of school functions related to evaluation, professional development, and school improvement. However, other districts and schools could provide follow-up support by leveraging existing resources, including mentors or instructional coaches.

LESSON 6: Identify deliberate strategies for integrating evaluation and professional development.

Teacher evaluation and professional development need not be hermetically sealed off from one another; in fact, education leaders should take deliberate steps to ensure they are tightly integrated.

Leaders can promote integration by selecting an evaluation rubric that offers a coherent vision and language for talking about effective instruction. In addition, several increasingly popular professional development activities offer perfect opportunities to reinforce skills on the evaluation rubric: common instructional planning time, individualized coaching, and peer observations. In TAP schools, the CODE data system helps master and mentor teachers leverage such opportunities by producing reports that analyze SKR scores and areas for reinforcement and refinement across the school or within professional development "cluster groups."

LESSON 7: Include teacher leaders as well as administrators among evaluators.

TAP demonstrates that evaluation systems which include teacher leaders as evaluators can indeed produce valid, "non-inflated" ratings, even when such individuals work in the same schools as their evaluatees.

Moreover, this approach offers distinct benefits for improving teacher effectiveness: Since evaluators know the teachers and coach them on a regular basis, they have a more robust context for selecting areas for reinforcement and refinement after observations, and they have far more opportunities to provide teachers with intensive follow-up support in those areas following evaluation post-conferences.

LESSON 8: Use an evidence-based evaluation rubric that balances breadth and depth.

For evaluation systems to support individual growth as well as accurate measurement, evaluation instruments must be suitable for both purposes—comprehensive enough to capture essential elements of effective instruction but not so broad as to sacrifice depth for breadth or to become unwieldy in practice.

The 19 areas covered by TAP's instructional rubric provide sufficient breadth to ensure that evaluation ratings reflect the kind of effective instructional practices that predict positive learning outcomes. At the same time, the Rubric is not so broad as to overwhelm either evaluators, who must rate teachers in each of those areas, or evaluatees, who need to have sophisticated enough understanding of each area covered by the Rubric in order to think about how to improve their practice.

Moreover, NIET has found that both SKR ratings and value-added ratings work best when they are measured on a 1-to-5 scale. A 1-to-5 scale avoids "floor" and "ceiling" effects that can obscure important differences in teacher effectiveness, and it provides *nearly all* teachers in a district or school with encouragement and support to continue to improve. Thus, the evaluation system is relevant and useful to all teachers, not just "low-performing" ones.

LESSON 9: Attend to the "human side" of evaluation by offering teachers plenty of opportunities to understand how and why the new system works.

Breaking the cycle of inflated ratings requires more than just new evaluation frameworks and processes. Successful and sustained implementation of a sophisticated teacher evaluation system that produces valid ratings requires careful attention to school culture and to the "human side" of performance measurement.

In TAP schools, master and mentor teachers spend a significant amount of time helping teachers understand the evaluation rubric before teachers are ever evaluated, explaining and modeling what solid performance (level 3 on the 1-to-5 scale) "looks like and sounds like" in actual practice. Moreover, teachers come to understand that, unlike traditional evalu-

ation systems, the highest levels on the scale represent truly exceptional performance attained by a small percentage of the most expert and effective teachers in any given system.

Teachers also score their own lessons after each observation and bring that self-evaluation to share in the post-conference. Self-scores count for 10 percent of the annual SKR average score, which promotes a reflective attitude toward personal performance and gives teachers a structured way to calibrate their personal vision of effective practice.

LESSON 10: Provide sufficient technical assistance to implement the system.

NIET has found that districts and schools need substantial external support and technical assistance if they are to successfully implement more sophisticated teacher evaluation systems, including, at a minimum:

- intensive training, certification, and annual recertification based on expertly rated videotaped lessons;
- protocols and tools for monitoring and remediating inter-rater reliability and score inflation over the course of the school year; and
- systems to help school leaders use data from evaluations both for quality control and for planning professional development.

Introduction

Teacher evaluation has emerged as a major focus for reform at the highest levels of education policymaking. The Obama administration awarded states more points for plans to improve teacher evaluation in their Race to the Top applications than for nearly any other policy area, and it is requiring all states to provide information about local teacher evaluation systems in exchange for formula-based stimulus funding. In March the administration went a step further: Its *Blueprint for Reform* for reauthorizing the Elementary and Secondary Education Act would *require* states to revamp teacher evaluation in order to continue receiving significant amounts of formula funding.

That unprecedented policy push stems partly from a spate of reports exposing deep flaws in how districts currently evaluate their teachers. A report by Craig Jerald for the Center for American Progress published last summer offered a damning summary of the many problems exposed by recent studies: In most places, teacher evaluations are infrequent; are based on scant evidence; rely on crude instruments; include few reliable quality controls; fail to use adequately trained evaluators; provide almost no useful feedback to teachers; and yield vastly inflated performance ratings that are not taken seriously enough to inform basic personnel decisions.[3]

One study by The New Teacher Project found that, in five districts with "binary" ratings systems (usually "satisfactory" or "unsatisfactory"), more than 99 percent of teachers received satisfactory ratings; in five districts with more than two possible performance levels, 70 percent of tenured teachers received the very highest rating and an additional 24 percent received the second-highest. Despite low levels of *student* performance across many schools in those districts, nearly three quarters of teachers in those districts received no specific feedback about how to improve their instruction.[4]

Clearly, reformers and national policymakers are right to push for major improvements in teacher evaluation. How can schools, districts, and states hope to dramatically improve teacher effectiveness when they lack any reliable way even to measure it?

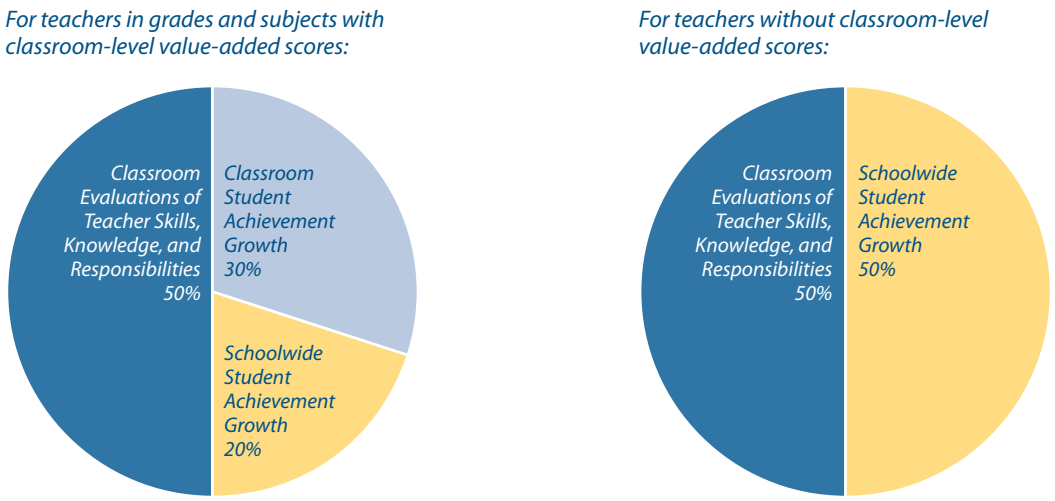
At the same time, states and districts currently have very little access to informed advice and practical guidance about exactly *how* to redesign teacher evaluation systems. While there are many ways to design and implement better approaches to teacher evaluation, there are also many ways to get it wrong. Indeed, research has shown that districts can adopt new evaluation systems that fit criteria suggested by reformers only to find that their new systems reproduce many of the same old problems—including vastly inflated performance ratings—as traditional evaluations.[5]

This paper offers policymakers and practitioners important "lessons learned" from TAP: The System for Teacher and Student Advancement. The TAP system is a comprehensive strategy to boost teacher effectiveness through opportunities for career advancement, professional growth, performance evaluation, and competitive compensation. TAP was developed by Lowell Milken and colleagues at the Milken Family Foundation and is now managed by the National Institute for Excellence in Teaching (NIET). TAP's rigorous evaluation system—including value-added measures of student growth—has been implemented in schools across the country for more than a decade.

There are good reasons for policymakers to pay close attention to lessons learned from TAP's evaluation system:

First, TAP uses a comprehensive approach to teacher evaluation which considers multiple measures of performance, including student achievement gains in addition to teachers' instructional practices. (See Figure 1.) A growing number of states and districts can calculate "value-added scores" (or simpler kinds of student growth measures), but such data are seldom, if ever, considered as part of formal teacher evaluations. Conversely, some districts have tried to reform teacher evaluations by using more sophisticated frameworks and conducting more frequent classroom observations, but those experiments have seldom, if ever, incorporated student achievement gains as a complementary measure.

Figure 1. Multiple Measures: TAP Considers Both Student Achievement Gains and Teaching Practices



Second, a recent analysis of data from across TAP schools offers compelling evidence that TAP's evaluation system produces much more valid performance ratings than traditional teacher evaluations.[6] (See Sidebar 2.)

Finally, unlike other proposals to reform teacher evaluation, TAP's evaluation system has been tried and tested with thousands of American teachers in real school settings over a significant period of time. In fact, TAP represents the longest sustained and most successful effort to radically transform teacher evaluation using multiple measures, including student achievement gains, in the United States today. Since its inception in the 2000-2001 school year, the TAP system has grown to serve over 10,000 teachers and 100,000 students across the country. Based on projected growth, NIET anticipates that TAP will expand to reach twice that many teachers and students by the 2011-2012 school year.

The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems

LESSON 1:

Policymakers hoping to strengthen teacher evaluation systems must begin by identifying specific goals for teacher evaluation, since those goals will guide important system design decisions.

It is almost impossible to design a sophisticated teacher evaluation system without making clear, specific, and honest decisions about exactly what goals the system will be expected to accomplish. Obviously, one could list many potential reasons for evaluating teacher performance. But any *specific* system will accomplish *specific* purposes, not a generic list of all possible purposes. For example, when it comes to a *particular* evaluation system, will the purpose be to:

- Hold teachers professionally accountable for their on-the-job performance?
- Produce valid and reliable data on teacher performance?
- Make personnel decisions (pay, tenure, etc.) based on measured teacher performance?
- Provide parents or other stakeholders with reports on individual teacher effectiveness?
- Offer an incentive for teachers to work to improve their performance?
- Provide various kinds of support to help teachers improve their performance?
- Some or all of the above?

This is far from an academic question. Any attempt to design a complex system (whether for teacher evaluation or school accountability or education finance) inevitably entails trade-offs among many possible design features, all of which might seem desirable in the abstract. Gaining clarity on specific goals for the system will help policymakers make thoughtful choices about difficult trade-offs when it comes time to actually design and implement the system.

TAP's approach to teacher evaluation focuses on two equally important objectives which can be considered the "dual goals" of the system: One goal is to produce sound summative data on teacher effectiveness that can be used to make performance and personnel decisions. For example, in TAP schools, teachers can earn annual bonuses for superior performance; therefore, teacher evaluation must function as a fair and reliable system for measuring professional performance. The second goal is to provide individualized and intensive support to teachers to help them improve their performance over time.

Those two goals for evaluation translate into two distinct levers for raising the overall level of teacher effectiveness in a school or district. For example, providing differential incentives based on performance (the first goal) can have a salutary impact on teacher turnover so that highly effective teachers become more likely to remain and less effective teachers become more likely to leave, which in turn elevates the effectiveness of the teacher workforce as a whole over time. Providing intensive feedback and assistance as part of the evaluation process (the second goal) gives every teacher the opportunity to improve on the job, regardless of his or her current level of measured performance, which also raises the average effectiveness of the workforce over time.

Seriously addressing multiple goals is a complex challenge and requires a greater investment of resources. However, TAP's designers deemed the investment worthwhile since the overarching aim for TAP was not just to increase instructional effectiveness in schools but to *dramatically* increase it. Recent research confirms what those designers took as an article of

faith ten years ago: Increases in teacher effectiveness can derive not only from *attracting and retaining* talented teachers, but also from *growing* the talent of every teacher every year.[7] To dramatically improve teacher effectiveness, TAP schools simply couldn't afford to leave either source of teacher effectiveness on the table.

In turn, gaining clarity on those dual goals enabled TAP's designers and early adopters to make difficult design decisions and navigate various trade-offs with greater confidence. Consider a common design decision mentioned in recent policy papers: Who should evaluate teachers? Should new evaluation systems rely only on administrators, as traditional systems do? Can teacher leaders (mentor teachers or instructional coaches) conduct evaluations in addition to administrators? If so, can teacher leaders who work in the same school as evaluatees conduct evaluations fairly and accurately, or should new evaluation systems rely only on evaluators who are "external" to the schools in which teachers work?

There is no single correct answer to those questions, since any particular design offers advantages as well as challenges. The only logical way to decide is to ask which approach best fits the specific goals for a *particular* evaluation system.

SIDEBAR 1: Brief Overview of How Teachers Are Evaluated in the TAP System

The TAP system evaluates teachers based on multiple indicators of performance that take into account both teaching practices and teaching outcomes.

Teaching Practices. In TAP schools, teachers are observed four to six times throughout the school year by multiple trained and certified evaluators, including principals (or other administrators), TAP master teachers, and TAP mentor teachers, who together form the leadership team. After each observation, the evaluator scores the teacher's lesson on a 1-to-5 scale based on an empirically-validated rubric called the *TAP Teaching Skills, Knowledge, and Responsibilities Performance Standards*. The Rubric allows evaluators to measure performance on 19 indicators of effective instructional practice. At the end of each year, teachers also receive scores in a fourth domain—professional responsibilities—that credits teachers for their efforts to improve teaching.

Teachers' scores on each of the 19 Rubric areas from all classroom observations are combined with the professional responsibility scores to calculate a final, year-end rating, again on a scale from 1 to 5, called the "Skills, Knowledge, and Responsibilities (SKR)" score. The SKR score is weighted according to various factors including the type of teacher being observed, the type of evaluator who conducted each observation, and the emphasis given to each domain and indicator in the rubric. The result is a standardized yet highly sophisticated and carefully calibrated evaluation rating for each teacher in each TAP school.

Teaching Outcomes. TAP schools also consider student learning outcomes when measuring teacher performance. In grades and subjects where such measures can be reliably calculated and reported, each teacher receives a value-added score, again on a scale from 1 to 5, which measures the teacher's impact on his or her students' learning by considering *gains* on state assessments. Teachers in TAP schools receive performance bonuses based on their SKR scores, schoolwide value-added scores, and (when available) individual value-added scores.

For both SKR and value-added scores, a rating of "3" is meant to reflect a truly proficient or "effective" level of performance (i.e., in the case of value-added scores, a full year of expected academic growth). A rating of "1" represents unsatisfactory performance on the teaching rubric, and it represents significantly lower than one year of average student growth as compared to classrooms of students with similar previous achievement.

SIDEBAR 1: Brief Overview of How Teachers Are Evaluated in the TAP System - *continued*

A "5" represents an exemplary level of performance on the teaching rubric and significantly higher than one year of growth for similar students.

Finally, TAP's approach to teacher evaluation is intended to help schools aggressively improve teacher effectiveness, not just reliably measure it. Announced observations are preceded by a pre-conference meeting between the evaluator and the teacher, and *all* observations are quickly followed by an in-depth post-conference meeting in which the evaluator provides specific, actionable feedback to the teacher. During post-conferences, evaluators highlight one area from the TAP Rubric for "reinforcement"—because the lesson revealed relative strengths in that area—and one area for "refinement" in which the teacher will work to improve. Following each observation, teachers receive intensive support from mentor and master teachers as they work to improve their performance in the area selected for refinement.

When designing the TAP system, Lowell Milken and his colleagues faced the difficult decision of whether to give master teachers and mentor teachers responsibility for teacher evaluation in addition to their responsibilities for instructional coaching and professional development. On the one hand, such an approach might undermine the kind of collegial relationships so critical for effective coaching and collaboration. Indeed, experts on instructional coaching had long warned that formal evaluation must be kept strictly walled off from instructional coaching.

On the other hand, the designers recognized that there might be important advantages to employing master and mentor teachers as evaluators along with the school principal. First, there were advantages related to cost and practicality: When they did the math, they realized that administrators alone would never be able to perform all of the observations and post-conferences necessary to ensure that every teacher was observed four to six times every year—at least not with any seriousness. Second, it would be cost-efficient to have master teachers and mentor teachers—who already would be earning salary bonuses for duties related to professional development—take some additional responsibility for conducting evaluations.

But the most convincing reasons were related to realizing the second of the dual goals—ensuring that the teacher evaluation system would, through expert feedback and advice, directly support improvements in teacher performance. External evaluators can provide feedback on whether and how a classroom lesson met a set of indicators or standards, but they lack the kind of detailed contextual knowledge of the school, students, and teacher to provide advice about how to change practice based on the feedback. And external evaluators would find it much more difficult to provide intensive follow-up support to teachers following a classroom observation and post-conference.

Moreover, Lowell Milken and his colleagues were determined to make the TAP system overcome the fragmentation in human capital policies typical of most public schools, where functions like evaluation and professional development work in isolation rather than reinforcing one another. They believed that giving school-based master teachers and mentor teachers major responsibilities for both functions would be a significant step toward achieving a more integrated approach to improving human capital in public schools.

TAP's unique approach to teacher evaluation is a result of many such technical design decisions, each of which called for the designers to determine how best to maximize attainment of both of the system's dual goals.

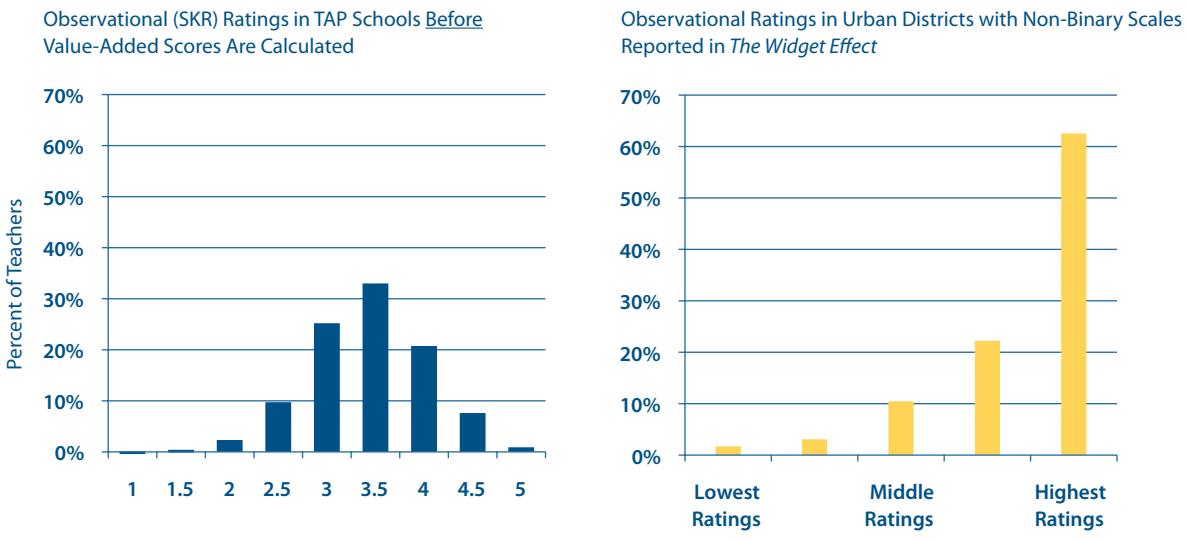
Sidebar 2: TAP's Evaluation System Produces Valid Performance Ratings

TAP's experience proves it is possible to design rigorous yet supportive teacher evaluation systems that produce more valid ratings than the embarrassingly inflated results documented in recent national reports. The findings below are from a recent analysis by Glenn Daley and Lydia Kim in NIET's research division. Their full report is available on NIET's website at http://www.tapsystem.org/publications/wp_eval.pdf.

1) SKR Scores Reflect Differentiated Performance. The average SKR score for teachers across TAP schools is 3.3 on a scale of 1 to 5. Moreover, the distribution of SKR results resembles a bell-shaped curve rather than a ramp, with most scores falling in the middle range of performance rather than at the top. Figure 2 below contrasts SKR scores from TAP schools with ratings from urban districts using traditional evaluation systems studied by The New Teacher Project for its 2009 report *The Widget Effect*.

Nationally, between one-quarter and one-third of TAP teachers score a 4, 4.5, or 5 on the SKR measure, and, among the subset of those teachers for whom classroom learning gains can be calculated, about one-third score a 4 or 5 on the value-added measure. Clearly, although TAP evaluates teachers based on student achievement gains as well as observational ratings, TAP does not rely on student test data to make teacher evaluations "rigorous." SKR results are rigorous in and of themselves, reflecting valid measurement of teacher performance even before the value-added results are calculated at the end of the year.

Figure 2. Observational Ratings in TAP Schools versus Urban Districts with Traditional Evaluation Systems



NOTE: TAP ratings include "Skills, Knowledge, and Responsibilities (SKR) " scores only, not value-added results. For *The Widget Effect* districts, scores on 3-point and 4-point scales were interpolated to a 5-point scale using a cumulative probability density function based on the reported data.

2) SKR Scores Are Significantly Correlated with Value-Added Scores. Among TAP teachers for whom value-added scores can be calculated, SKR scores are significantly correlated with value-added scores. In other words, on average, higher observed instructional quality during the year predicts higher student learning gains by the end of the year. Across TAP schools, for every one point that a teacher's SKR score improves, his or her value-added score improves by more than half a point. That relationship holds up across a variety of statistical models controlling for school characteristics, including schoolwide performance. (See Figure 3.)

LESSON 2: Student achievement gains can and should be one of multiple, complementary measures used to evaluate teacher performance.

TAP's approach to evaluation is grounded on the belief that validly measuring teacher performance requires consideration of multiple, complementary measures, and that one measure must take into account the *results* of teachers' efforts—student learning. Like K. Anders Ericsson, founder of the branch of cognitive science that studies expert performance, NIET believes that, no matter the field of endeavor, "real expertise produces concrete results. Brain surgeons, for example, not only must be skillful with their scalpels but also must have successful outcomes with their patients. A chess player must be able to win matches in tournaments." [8]

TAP schools and districts can create their own systems for measuring value-added student growth, or select an external provider of this analysis. Many select the value-added approach developed by Dr. William Sanders, who pioneered the methodology in Tennessee and is now the senior manager of value-added assessment and research for SAS Institute Inc. in Cary, North Carolina. As a variety of learning growth models have become more widely available, some schools and districts have selected other vendors or have developed their own methods and systems.

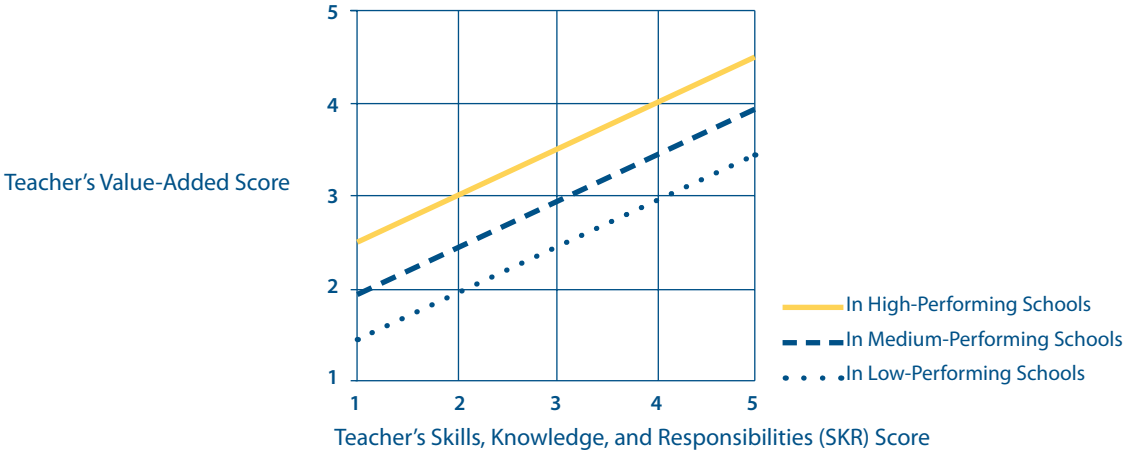
The value-added method is a fair and accurate way to measure student learning as one indicator of teacher performance. First, it is based on how much academic *growth* students make over the course of the year while they are assigned to a given classroom and school. Second, it estimates the direct impact that teachers and schools have on student learning—isolated from other contributing factors such as family characteristics and socioeconomic background.

For multiple measures to work in a teacher evaluation system, the measures must be *different yet complementary*. One way to test whether measures are complementary is to statistically analyze the extent to which they are correlated. For the TAP system, that means that teachers who receive higher SKR scores should, on average, have higher value-added scores. A recent analysis of data from across TAP schools revealed that is exactly the case: "When teachers demonstrate strong instructional skills as measured by the TAP observation methods and rubrics, their students show higher academic growth regardless of previous achievement and socioeconomic status."[9] The strong relationship between SKR and value-added scores holds up across a variety of statistical models controlling for school characteristics, including schoolwide performance. (See Figure 3.)

At the same time, different measures should be measuring different aspects of performance. While multiple measures should correlate statistically, they need not, and should not be expected to, correlate perfectly. Otherwise the measures would be redundant, and the system could be operated more simply and cheaply by eliminating redundant indicators. Figure 3 shows that in TAP schools, SKR scores and value-added scores function as different yet complementary measures of teacher effectiveness.

Figure 3. Strong Correlation between TAP's Measures of Teacher Performance

This chart shows the relationship between observational SKR scores (which measure teaching practices), and value-added scores (which measure teaching outcomes), across TAP schools. The chart only includes teachers for whom classroom-level value-added scores can be calculated. The correlation between measures of teacher performance remains significant even when schoolwide performance is taken into account.



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.

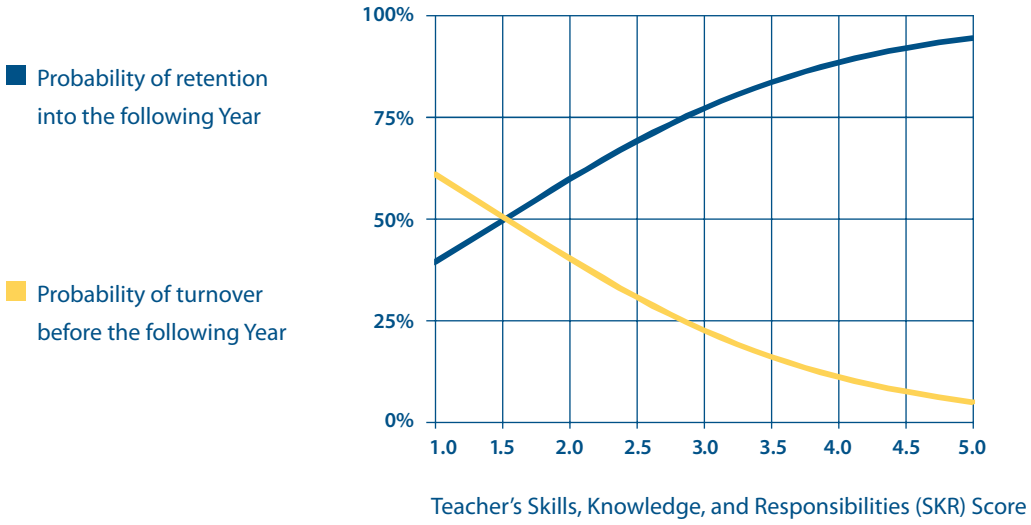
Multiple measures are important for achieving *both* goals of TAP's evaluation system—accurately measuring teachers' performance and helping teachers improve their performance over time.

First, different yet complementary measures should enable more valid and reliable measurement of teachers' current performance than would any one measure individually. Researchers have recently found evidence to support this contention in non-TAP settings as well: An analysis of data from Cincinnati Public Schools by Thomas J. Kane and colleagues yielded results which "provide initial support for the notion that multiple alternative measures of teacher effectiveness may be more predictive of future student achievement effects than any single measure."[10]

Such considerations are especially important whenever the results of evaluation are used to make significant personnel decisions. For example, it is important to target incentives to the right teachers not only for the sake of fairness but also because such policies can impact teacher retention and turnover which, in turn, affects who is teaching and the quality of instruction students receive.

A recent analysis by NIET provides preliminary evidence that both SKR scores and value-added scores are having a complementary and positive impact on teachers' career decisions: For each point higher that a teacher's SKR score is in one year, that teacher's odds of remaining a teacher in a TAP school the following year increase by 87 percent. For every full point higher that a teacher scored on the value-added scale the previous year, the odds of retention increase by 27 percent. At the same time, teachers who scored lower on those measures were more likely to leave.[11] (See Figure 4.)

Figure 4. Relationship between Teachers' SKR Scores and Teacher Retention and Turnover across TAP Schools



n = 7,377 teacher-year cases from 2005 through 2009
Retention includes teachers who stayed in TAP schools, including as master and mentor teachers.
Turnover includes teachers who left teaching, left TAP schools, or became administrators.

Second, different yet complementary measures provide multiple tools to help teachers improve. In TAP schools, classroom observations yield not only SKR scores based on the TAP Rubric but specific advice for how to improve on dimensions of the Rubric, including targeted areas for reinforcement and refinement after each observation along with intensive follow-up support in the classroom. In addition, teachers can use value-added results to analyze growth for different student subgroups such as high-, medium- and low-performing students to determine whether their instruction is having a differential impact that should be addressed.

A final caveat: While it is important to consider student learning when evaluating teacher performance, it would be a mistake to consider value-added or other learning measures some kind of panacea for the problem of inflated evaluation ratings. If "multiple measures" are to work, then *all of the measures, including classroom observations or other qualitative measures, must be valid and rigorous in and of themselves*. Otherwise teachers will receive mixed signals on their effectiveness and confidence in the system will begin to erode. (See Sidebar 3.) Results from TAP schools prove this is an achievable policy goal, although, as we discuss in Lesson 3, it requires considerable attention to up-front training and ongoing quality control.

SIDEBAR 3. Why Do Classroom Teaching Measures Need to Be Significantly Correlated with Student Growth Measures?

Sometimes discussions about improving teacher evaluation seem to assume that observational scores will always be higher than value-added scores. But that need not be the case if policymakers design evaluation systems thoughtfully and make strategic *up-front* investments in quality control. In fact, policymakers should anticipate the following negative consequences *if they do not take steps* to ensure that measures of teaching practice sufficiently correlate with measures of student learning gains:

- Teachers will receive conflicting signals about what matters and how they can improve their performance.
- Teachers will have little incentive to put in the effort necessary to improve on the measured instructional practices, since they will have no confidence it will pay off in improved student learning.
- Evaluators will struggle to help teachers make the connection between practices measured by the rubric in the classroom lessons they observe and evidence of student learning growth.
- Principals will have less incentive to put in the hard work necessary to maintain reliability and accuracy of observation-based scoring if these evaluations bear no connection to student achievement in their schools.
- As a result, evaluations will still feel like a bureaucratic exercise rather than an authentic school improvement activity, and busy principals will begin to "cut corners" during observations and post-conferences.
- Public confidence in the evaluation system will suffer, and support for using evaluation results to make personnel decisions will begin to erode.

LESSON 3: Ensuring reliability and validity of evaluation results requires significant, strategic investments in quality control.

Advocates for improving teacher evaluation systems need to be aware that producing valid performance ratings requires much more than simply adopting a new set of tools and procedures and providing some initial training on them. The TAP system has worked hard to ensure valid, non-inflated ratings by investing in extensive quality control mechanisms that "wrap around" the school year—taking place before, during, and after teachers are evaluated.

1. Before Observations Take Place. Before members of a school's leadership team can perform evaluations, they must successfully complete an eight-day training program (with four days devoted to evaluation and four days to other elements of TAP) that culminates in a performance-based certification assessment and is followed by annual recertification tests. Since school leadership teams bear responsibility for ensuring valid and reliable ratings, all members of the team must train together.

First, team members are provided with in-depth instruction on the *TAP Teaching Skills, Knowledge, and Responsibilities Performance Standards*, more commonly known as the TAP Rubric, breaking down each domain and carefully examining every performance indicator. Then they receive training on how to "script" a lesson (i.e., transcribing parts of the lesson and taking meaningful notes on the rest, including student behaviors), recognizing what kinds of evidence to capture in order to accurately score the lesson against each of 19 indicators in the Rubric.

Extensive opportunities for practice follow, during which teams observe and script videotaped lessons; discuss evidence from the lessons related to the Rubric; and arrive at a consensus on scores for each Rubric indicator and for the lesson overall. After teams reach a consensus on the scores they would give a particular lesson, trainers share the scores for the lesson assigned by "national raters" (highly experienced executive master teachers or members of the NIET national staff), along with national raters' evidence from the lesson to justify the scores.

According to Sue Way, a TAP executive master teacher with the Louisiana Department of Education, this is a critical point in the training process. "That's a big 'ah ha' moment for them, because usually they have given the lesson a much higher score than the national rater," she says. "At first, they generally see non-proficient instruction as proficient, and they still see the TAP Rubric as a checklist rather than a tool for analyzing the entire lesson."

This aspect of the training helps evaluators understand that the Rubric should not be used as a mere checklist but rather as an analytical tool. Instead of simply waiting for a teacher to exhibit some form of a behavior that matches an area of the Rubric so that indicator can be "checked off," evaluators learn to analyze whether a strategy described in the Rubric is used appropriately in the context of the lesson itself as a vehicle for helping students learn the content in question. For example, exemplary teachers do not just use student groups for the sake of doing so, but rather use grouping in particular ways depending on the structure and content of the lesson being taught and the needs of the students in the classroom. For administrators who have long experience with the checklist approach used in traditional evaluations, this can be a revelation.

The TAP System Training Portal provides evaluators with a library of nationally rated classroom lessons with detailed evidence and scores for each indicator on the Rubric. Evaluators can use the Portal to practice and improve their evaluation skills outside of the formal, "in person" observations they conduct during the year.

Importantly, the training sessions also teach evaluators how to plan for and conduct the post-conference meetings with teachers that must take place after each observation. Because the training on scoring emphasizes collection and use of evidence from the lesson—including teacher practices, student behaviors, and student work—to arrive at and to justify a score, trainees are very well prepared to understand the critical role that such evidence plays in the post-conference conversation.

At the end of the training each member of the leadership team must pass a performance assessment in which they show they can gather sufficient evidence by "scripting" a lesson, can analyze evidence from the lesson to arrive at an accurate score that is in line with national raters, and can apply that evidence to plan an effective post-conference. Team members must pass a recertification assessment every year.

2. During the Observation Cycle. During the school year, leadership teams take explicit responsibility for ensuring the quality of teacher evaluations. Teams devote at least one meeting per month to discussing issues related to evaluation and analyzing data to identify potential problems with *inter-rater reliability*, the extent to which evaluators are consistently applying the TAP Rubric when evaluating lessons. NIET has developed a TAP System Comprehensive Online Data Entry (CODE) system that stores evaluation results and can produce a range of tables and charts to examine inter-rater reli-

ability and guard against score inflation. Figures 5 and 6 provide examples of CODE analysis charts that school leadership teams use to monitor inter-rater reliability.

CODE analysis charts might reveal that evaluators are inconsistent in their ratings of a particular *Rubric Indicator* across the school. (See Figure 6.) For example, in the area of *Questioning*, a leadership team might find that evaluators vary on how they categorize low-level and high-level questions that teachers ask of students, which is causing a lack of inter-rater agreement on that indicator. The team would make time for an in-depth discussion of this topic, referencing real examples from lessons, that results in a shared operational understanding of what constitutes high-level versus low-level questioning.

Leadership teams can employ a number of strategies to monitor inter-rater reliability and guard against score inflation or to calibrate evaluations if CODE reports reveal problems. They can conduct teamed evaluations, either as a formal part of the evaluation process or on an informal basis as necessary. They can invite highly experienced evaluators from outside the school, such as executive master teachers or TAP national staff members, to assist in calibrating evaluation scores.

Finally, NIET has compiled an extensive DVD library of videotaped lessons available on the TAP System Training Portal that have been scored by national raters. School leadership teams are encouraged to make use of the videos during leadership team meetings to troubleshoot issues and ensure that team members are continuing to apply the TAP Rubric consistently and accurately after they have been certified.

3. After Observations Take Place. The classroom-level value-added scores that principals receive for a subset of teachers in the school also provide an important tool to monitor whether there are problems with score inflation. Leadership teams can analyze the relationship between final SKR scores and value-added scores on a schoolwide level, a cluster group level, and an individual teacher level.

State-level and NIET staff also monitor inter-rater reliability and correlations between SKR and value-added scores at the school, district, and state levels. Several years ago the TAP team at the South Carolina Department of Education noticed that the highest-performing TAP schools (based on schoolwide value-added gains) were achieving much higher rates of inter-rater reliability among evaluators than lower-performing schools, so they took steps to help school leadership teams reach stronger consensus on the vision for effective instructional practices described in the TAP Rubric.

Figure 5. Example of CODE Chart for Monitoring Inter-Rater Reliability: A Case of Inconsistent Scoring Across Evaluators

The following chart illustrates one of the reports the CODE system can produce to help school leadership teams analyze inter-rater reliability. In this example, the average evaluation ratings of classroom teachers observed by one master teacher are significantly higher than the average evaluation ratings of classroom teachers observed by a second master teacher. Noticing that pattern, the leadership team can probe more deeply to determine whether the variance reflects true differences in skills across teachers or instead represents a problem with inter-rater reliability which must be remedied immediately.

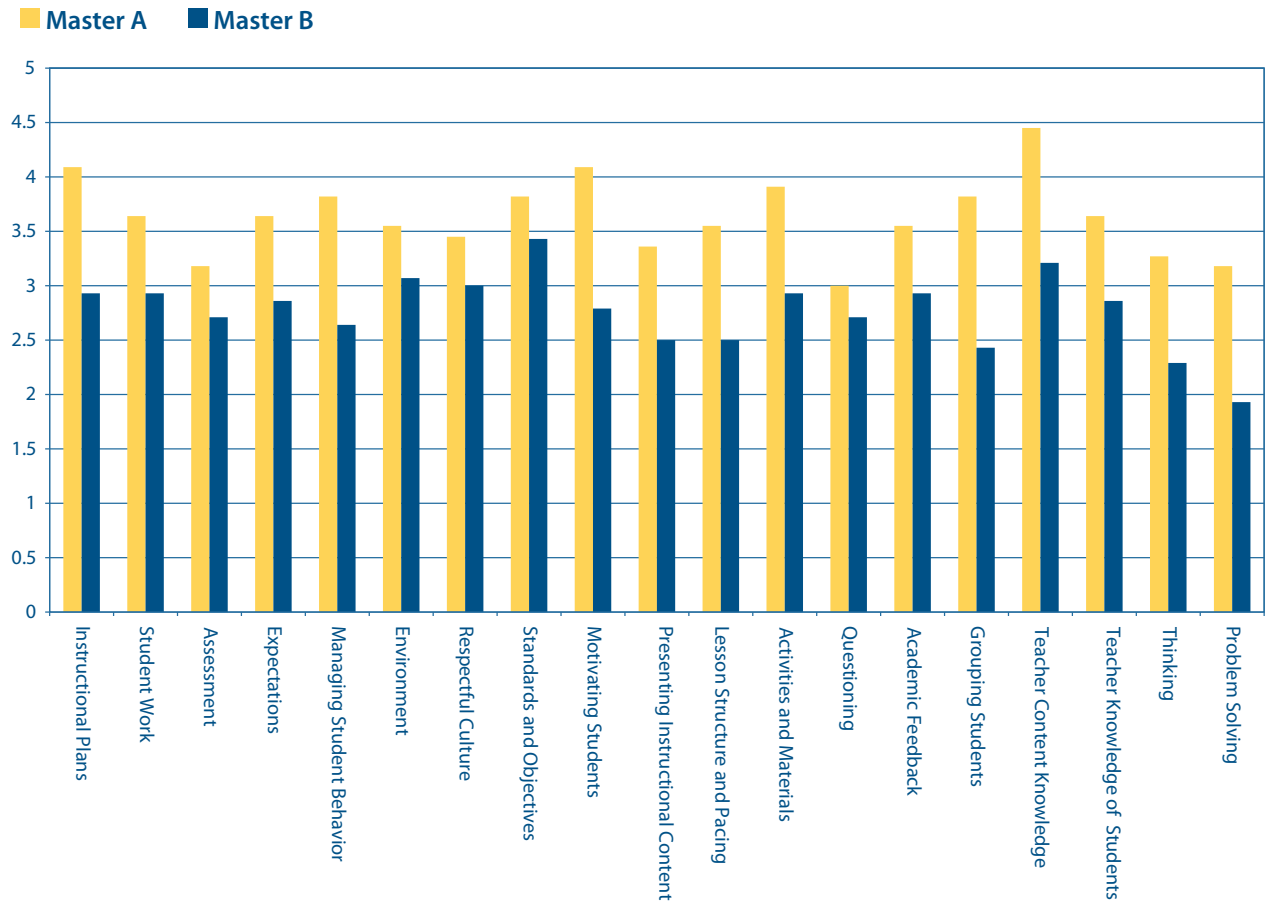


Figure 6. Example of CODE Chart for Monitoring Inter-Rater Reliability: A Case of Inconsistent Scoring of One Rubric Indicator

The following CODE report shows that mentor teachers have been relatively consistent in rating observed lessons across all TAP Rubric indicators except one—*Questioning*—which means that they might have different ideas of what that particular teaching skill looks like at different performance levels on the standardized 1-to-5 scale. In a case such as this, the leadership team can employ a range of strategies for calibrating the expectations of evaluators, including practicing with videotaped lessons that illustrate what student *Questioning* looks like at each performance level on the Rubric.

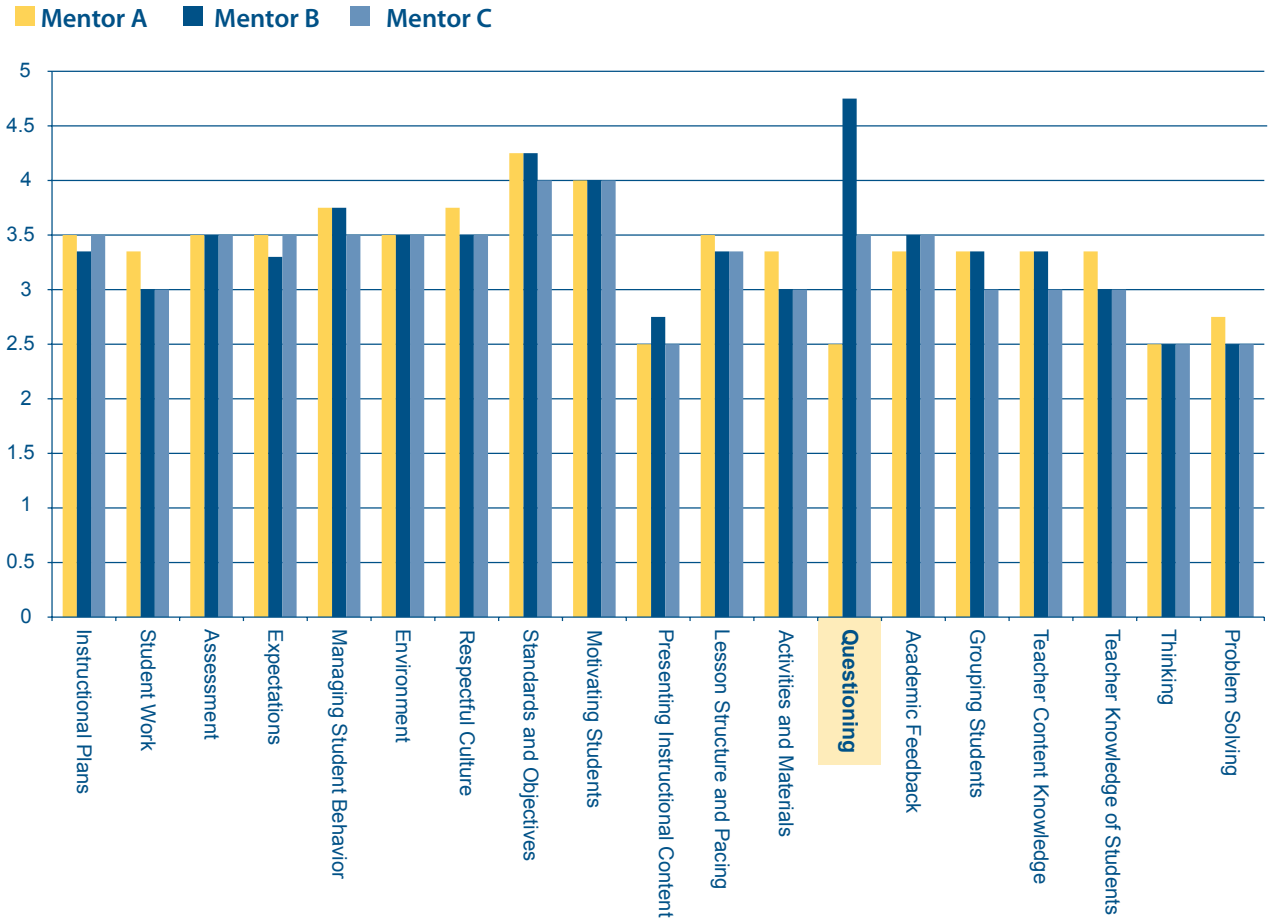


Figure 7. Meeting Four Major Measurement Challenges of Rating Teachers' Practices

Measurement Challenge	What Is It?	How Does the TAP System Address It?
1) Reliability	Reliability refers to consistency in measuring the aspects of practice described in a rubric such that scores reflect actual teaching rather than non-relevant personal or contextual factors. For example, in a system with high reliability of scoring, the same evaluator would give two teachers who taught the same lesson the same way the same scores, or would give a teacher consistent scores over time if his or her teaching practices did not improve.	» TAP uses a rubric that has been documented to support reliable scoring under optimal conditions and in the field. » TAP requires evaluators to undergo thorough training that addresses reliability concerns (e.g., how to base scores for each indicator on recorded evidence from the lesson). » TAP requires evaluators to pass a performance-based assessment following training in order to be formally certified. » TAP requires evaluators to pass an annual assessment for recertification. » School-based evaluation teams meet monthly to monitor quality of scoring. » TAP's CODE data management system provides standardized reports that help evaluators identify potential problems with scoring.
2) Inter-Rater Reliability	Inter-rater reliability is a particular aspect of reliability that requires extra attention in evaluation systems that rely on multiple evaluators. It concerns the consistency of scoring across evaluators such that all evaluators would give the same observed lesson the same scores.	» TAP requires all members of a school's evaluation team to undergo training and certification at the same time, which enables discussion and practice to calibrate expectations. » During training, members of the evaluation team observe and score the same videotaped lessons until a sufficient level of consensus has been clearly established. » School-based evaluation teams meet monthly to monitor quality of scoring, with a special emphasis on inter-rater reliability. » TAP's CODE data management system provides standardized reports for monitoring inter-rater reliability. (See Figures 5 and 6.) » NIET provides evaluation teams with assistance, strategies, and tools to immediately remediate any identified problems with inter-rater reliability, including an online video library of annotated, nationally rated lessons.

Figure 7. Meeting Four Major Measurement Challenges of Rating Teachers' Practices - *continued*

Measurement Challenge	What Is It?	How Does the TAP System Address It?
3) Accuracy[12]	Even if scoring is generally reliable and a high level of inter-rater reliability has been established, scores could still be systematically inflated or deflated such that they do not reflect true performance against the standardized scoring scale. In technical measurement terms, such scores would be "biased upward" or "biased downward." For example, members of an evaluation team might consistently be assigning scores of "4" for *Lesson Structure and Pacing* to teachers who actually should be earning scores of "3" on that indicator.	» The TAP Rubric has clear "descriptors" for different performance levels on each of the 19 indicators. » TAP's training enables evaluators to gain detailed understanding of the different performance levels for each of the 19 rubric indicators. » TAP's training teaches evaluators how to base scores for each indicator on concrete, recorded evidence from the observed lesson. » TAP's training provides evaluation team members with extensive practice scoring videotaped lessons and then comparing their scores with the scores assigned by "national raters," who also justify their scores based on evidence from the lesson. » TAP requires evaluators to pass a performance-based certification assessment during which they must demonstrate that they can script and score lessons within an acceptable margin of accuracy. » TAP requires evaluators to undergo an annual assessment for recertification, which guard against "expectations drift" over time. » TAP requires four to six observations per year, half of which must be unannounced, in order to capture a *representative* picture of each teacher's instructional practices and students' responses to those practices. » Evaluation teams and NIET staff members analyze CODE data to compare SKR scores and value-added scores, which can signal problems with score inflation or deflation.

Figure 7. Meeting Four Major Measurement Challenges of Rating Teachers' Practices - *continued*

Measurement Challenge	What Is It?	How Does the TAP System Address It?
4) Validity	Validity has to do with whether conclusions based on evaluation results are justified. (Reliability and accuracy are necessary but not sufficient to ensure that observational scores offer valid, rather than misleading, information about teachers' effectiveness.) Validity has multiple aspects. Following are two of the most important for policymakers to consider when designing and implementing teacher evaluation systems: A) "Construct Validity." Does the evaluation really measure what it is intended to measure? If evaluation systems are not designed carefully, they can measure some things that are irrelevant to teacher effectiveness while ignoring other things that are essential to it. For example, traditional evaluations have been criticized for considering teachers' grooming habits and style of dress while ignoring classroom practices that actually promote student learning. One way experts assess construct validity is to examine whether results on one measure "converge" with results on other measures that *should* be closely related. B) "Predictive Validity." How well do observational scores predict the desired *outcome* of effective classroom instruction—student learning? If the goal of evaluation is to measure effective teaching, then teachers with higher observational scores should have students who make higher learning gains, other factors being equal.	» During development of the Rubric, the Milken Family Foundation conducted studies to validate that teachers who scored higher on targeted instructional practices had students who achieved higher learning gains; the results of one such study were published in a peer-reviewed academic journal, *Economics of Education Review*.[13] » TAP's training for evaluators equips them to capture evidence of students' behavior during observed lessons in addition to what the teacher says and does, since both kinds of information are critical for measuring effective classroom instruction. » As described in Lesson 3, NIET's research team closely monitors the relationship between SKR scores and value-added scores across TAP schools as the evaluation system is implemented on a wide scale.[14]

LESSON 4: The post-conference cannot be treated as a bureaucratic formality; it is one of the most critical features of an effective teacher evaluation system if the goal is not just to measure the quality of teaching, but also to improve it.

Simply mandating that evaluators meet with teachers to provide feedback after observations take place is not a sufficient strategy for ensuring the kind of high-quality feedback that can improve instruction. Even if policymakers are vigilant about putting in place systems to be sure that post-conferences are being conducted, the conferences themselves will most likely be of low quality. In fact, TAP adopters have found that providing effective feedback is one of the most complex and difficult challenges of implementing a professional evaluation system. Evaluators need significant expertise, training, and ongoing support to ensure that post-conferences are conducted in ways that help teachers improve their performance.

In the TAP system, members of school leadership teams receive explicit training on how to conduct effective post-conferences as part of the formal process for becoming trained and certified evaluators. During the training, evaluators learn what effective feedback is and how to deliver it in ways that teachers can actually use when they return to their classrooms. At the end of the training, they must convincingly demonstrate that they can plan and carry out high-quality post-conferences in order to obtain the certification necessary to evaluate teachers in a TAP school.

Given the lack of attention to high-quality post-conferences in traditional evaluation systems, training must help evaluators break bad habits by understanding what the post-conference is and what it *is not*:

- The post-conference is not a meaningless formality that can be skipped or delayed, but rather a high-priority meeting that occurs soon after the observation.
- The post-conference is not a "quick check-in," but rather an in-depth, highly analytical conversation lasting at least 40 minutes.
- The post-conference is not a free-form "informal chat," but rather a focused discussion following a protocol that requires meaningful "hard feedback" on the observed lesson.
- The point of the post-conference is not simply to convey information ("tell the teacher how she did") but rather to collaboratively analyze how and why teacher actions have a positive or negative impact on student learning.
- The post-conference is not a bureaucratic "check-the-box" exercise that happens a few times a year, but rather a valuable coaching activity that contributes to teachers' ongoing growth and development.
- The post-conference is not an end in itself, but rather the basis for explicit follow-up actions on the part of the teacher and the evaluator(s).
- The post-conference should take place within 48 hours of the observation to be as meaningful and impactful as possible.

Teachers in TAP schools receive two kinds of explicit feedback in addition to learning how the evaluator scored the observed lesson—an "area for reinforcement" and an "area for refinement." Each is related to one of the 19 indicators on the TAP Rubric. The area for reinforcement is one in which the teacher demonstrated relative strength during the lesson and should continue to build. For example, the lesson might have demonstrated relative strength in providing students with *Academic Feedback*. The area for refinement is one in which the teacher could improve based on what the evaluator observed and documented during the lesson. For example, the lesson might have revealed significant room for the teacher to improve aspects of *Lesson Structure and Pacing*.

That protocol for structuring post-conferences ensures that evaluators provide teachers with valuable "hard feedback" rather than falling back on the kind of polite platitudes that severely undermine so much instructional coaching in U.S. schools.[15] At the same time, it provides sufficient flexibility to ensure that post-conferences do not become rote and that evaluators can address teachers' unique questions and concerns. However, selecting areas for reinforcement and refinement and discussing those areas with teachers requires much more sophistication than one might expect.

1. Feedback must be highly focused and specific to be actionable. Early in the development of the TAP system, TAP's designers and early adopters realized that it is very easy to overwhelm teachers with too many goals for improvement after each observation. In such cases, teachers are unable to reflect deeply on the feedback and to act on it when they return

to the classroom. We now know that cognitive scientists have found this to be generally true across all fields of endeavor: Stretching performance beyond one's current capabilities by focusing on targeted areas for improvement takes up a great deal of working memory and requires "enormous concentration."[16]

Therefore, even if the observed lesson demonstrated multiple areas of relative strength or many areas that could be improved, TAP evaluators focus only on one reinforcement objective and one refinement objective per post-conference, allowing for a much more in-depth analysis of each. Evaluators and teachers are able to examine evidence from the lesson regarding each area; discuss how particular teacher behaviors promote student learning; brainstorm ways to build on the area for reinforcement and improve the area for refinement in future lessons; and generate explicit actions the teacher will take to improve. (In fact, in order for the conversation to focus intensively on those two areas, evaluators typically wait until the end of the post-conference to share scores for each Rubric indicator and for the lesson overall.)

One reason it is possible to focus on a single reinforcement objective and a single refinement objective is that TAP requires four to six observations for every teacher every year. Thus, each teacher receives in-depth, specific feedback and support on eight to twelve instructional skills in a given year. A system that relies on fewer observations might be forced to address multiple areas each time, especially for teachers who initially fall far short of "satisfactory" performance. As we discussed under Lesson 1, policymakers need to be very aware of such tradeoffs when designing evaluation systems.

2. Areas for reinforcement and refinement must be carefully selected. Objectives for reinforcement and refinement are not just automatically determined by which of the 19 indicators the teacher scored highest and lowest on during the observation. Rather, to achieve maximum positive impact for teachers and students, the selection must be much more deliberate than that. To choose an area for refinement, TAP evaluators weigh all of the following considerations at once:

- Which areas of the TAP Rubric received relatively low scores during the observation?
- Which areas would have the greatest impact on student achievement?
- Since most of the 19 instructional indicators on the TAP Rubric are related to one another in practice, which indicator would help the teacher improve in *other* areas of the Rubric?
- Conversely, would a weakness in another area make it difficult for the teacher to work to improve in this area? For example, teachers often need to achieve a certain level of mastery in *Presenting Instructional Content* as a foundation for improving their skills in *Lesson Structure and Pacing*.
- Given the teacher's current level of expertise, which areas of the TAP Rubric present the greatest immediate opportunities for growth?
- Does the evaluator's script offer sufficient evidence from the lesson to support the choice, especially since the evaluator will be asking the teacher to spend a significant amount of time and energy to improve in that area?
- Does the evaluator have sufficient expertise or access to expertise in the area for refinement to answer the teacher's questions about it and to provide explicit examples of how the teacher can apply the feedback in future lessons?

Obviously, given that list of considerations, choosing an area for refinement is an intellectually challenging task requiring deep expertise and strong analytical (or "critical thinking") skills. As Sidebar 4 makes clear, choosing the area for reinforcement is no different.

SIDEBAR 4. The "Area for Reinforcement": Identifying and Analyzing Instructional Strengths in Evaluation Post-Conferences

In TAP system schools, the purpose of identifying instructional strengths documented during a classroom observation is not to provide a "pat on the back" that puts teachers at ease before discussing areas that can be improved. Rather, it is an important element of feedback in its own right: Teachers need to understand how their instructional strengths positively impact student learning and think about how they can build on those strengths as they plan future lessons. Certified TAP evaluators are provided with training on how to carefully select and discuss areas of reinforcement with the teachers they have observed.

✓ Some of the questions that evaluators consider when selecting an area of reinforcement from the TAP Rubric based on strengths observed in the lesson include:

- » Which indicator would have the greatest impact on student achievement?
- » Is the identified area of reinforcement an area in which the teacher is exhibiting noticeable growth?
- » Are there connections between an area of strength and other indicators on the TAP Rubric, such that being strong in one area strengthens overall teaching practice?

✓ Questions that evaluators might ask teachers about the area of reinforcement during the post-conference to elicit "metacognition" about their instruction:

- » Why did you decide to do this in your lesson?
- » How would you explain to a first-year teacher the importance of doing this?
- » How did this help students learn the material or master the skill during the lesson? Can we see it in the student work from the lesson?
- » How can you continue to use this in your future lessons?
- » How can you continue to strengthen this area of your teaching?

Source: Adapted from Ancar, A. & Bolen, N. (April 30, 2010). *The Reinforcement Area of the Post-Conference.* Presentation at the Louisiana "WE BELIEVE" Master Teacher Networking Meeting.

3. Training and protocols should ensure that the post-conference is analytical, not simply descriptive, and should encourage teachers themselves to become more independently reflective about their own performance. For teachers to put feedback to practical use, they must understand not just *what* the areas for reinforcement and refinement are, but *why* they are important and *how* they contribute to improved student learning. TAP's expert support staff often put it this way when training evaluators: "The post-conference must be conceptual, not just procedural."

TAP's evaluation system encourages those aims in several ways. First, evaluators learn how to script a lesson with sufficient clarity and detail that they will be able to draw on specific evidence from the lesson when discussing the area for reinforcement and the area for refinement with the teacher. Second, they learn how to use evidence of student work produced during the lesson to illustrate how the area for reinforcement or area for refinement had an impact on student learning. Third, they learn how to use "leading questions" to elicit reflective thinking from the teacher, often beginning with a very general question ("How did you think the lesson went?") and leading up to more specific questions ("How would you have structured the lesson differently to pace it better given what we've discussed today?").

Encouraging self-reflection takes considerable energy and skill on the part of evaluators, but it is critical for helping teachers improve their effectiveness. Cognitive scientists have found that expert performers have well-developed "metacognitive" skills (the ability to monitor, evaluate, and reason about one's own performance) and that such skills are a prerequisite for advancing from novice to expert in any field. According to K. Anders Ericsson, the leading researcher in the field, "After years of daily practice, the aspiring expert performers … start taking over the evaluative activity of the teacher and coach. They acquire and refine cognitive mechanisms that … allow them to monitor performance in representative situations to identify errors as well as improvable aspects." [17]

For example, Ericsson has observed the following of novice chess players who eventually become world-class masters: "By more careful and extended analysis, the aspiring player is generally able to discover the reasons for the chess master's move[s]. Most importantly, they can then reflect on their [own] thought process during their faulty move selection and examine how they need to change their planning methods when encountering other related situations and games." [18]

Some might consider that example a bit extreme: Does the TAP system expect every teacher to work toward such high levels of mastery? Obviously, not every single teacher will attain highly advanced expertise and "level 5" effectiveness. However, if an evaluation system provides every teacher with "stretch goals" regardless of his or her current level of performance, many more teachers will attain that level than would have been the case otherwise. And just as importantly: Every single teacher will have been given the *opportunity* and *encouragement* to strive for such high levels of professional effectiveness. The TAP system is founded on the belief that teachers—and their students—deserve no less.

Finally, quality control is important here, too. Just as school leadership teams are charged with ensuring that the system generates valid and reliable performance scores, they also take explicit responsibility for monitoring and maintaining the quality of post-conferences. Leadership teams adopt a common template for structuring post-conferences. Team members can model effective post-conferencing techniques for each other, and they can informally observe and provide feedback on colleagues' post-conferences. Principals must *formally* observe and provide feedback on master teachers' and mentor teachers' post-conferences as part of TAP's evaluation process for those positions.

Sidebar 5. What Does Effective "Feedback" Look Like? Portion of a Post-Conference Related to the "Area for Refinement"

This is a transcript of a portion of a post-conference focused on the area for refinement. The post-conference occurred the day after a master teacher observed a fourth-grade teacher's reading lesson. The transcript begins after the master teacher and the fourth-grade teacher (known as a "career teacher" in TAP terminology) have already discussed how the lesson went in general and have selected and discussed an area of strength for reinforcement. The next portion of the post-conference focuses on the area for refinement in which the teacher has room to improve her instruction.

Because this career teacher already is relatively high-performing, the master teacher guides her to a refinement skill that is discussed in the "exemplary" performance level under the Rubric's Academic Feedback *indicator—helping students give high-quality academic feedback to one another. (See Figure 10.) Thus, this transcript illustrates how TAP's evaluation system provides critical feedback and support for continued growth even to teachers near the upper end of the performance range.*

Master Teacher (MT): Think about your lesson now. You know as strong teachers, we typically focus on way too many areas [for refinement]. But if you thought about your lesson and you thought, 'I wish I had included something in the lesson that I left out,' or 'I wish I had done something a little bit differently' … If there was one thing you could have changed, what do you think that might have been?

Sidebar 5. What Does Effective "Feedback" Look Like? Portion of a Post-Conference Related to the "Area for Refinement" - *continued*

Career Teacher (CT): Well, my choice would have been different before the first part of this conference. [Teacher references the discussion of the area for reinforcement, which she thinks could have been even stronger in the lesson.] Now I would say that I would have given the students more time to interact with each other and reflect on their thinking.

MT: Okay. I know exactly what you're saying there. However, in that vein, I want to push that just a little bit more, because we did talk about the students actually modeling their thinking to one another. So, thinking about that, how important do you think it is for children to give high-quality academic feedback to one another?

CT: Oh, it's very important, like we talked about before. And it validates the partner as well, because then you're listening for another purpose. You're listening not only just to listen, but you're also listening to give each other feedback and to push each other [intellectually].

MT: You're absolutely right. In the area of *Academic Feedback*, I want you to understand that you did an excellent job of providing teacher level academic feedback to your students. For example, the word 'slaughter.' When you were trying to help them understand what a 'slaughterhouse' is. When they gave you an accurate definition, you didn't just say 'good job' and move on to the next person, but you allowed them to give their definitions, and then you explained to them why they were right. So you helped them make those connections. Also when [child's name] was worried about … [Laughter] the pork. I thought it was funny that they didn't at first make the connection that a pig was bacon …

CT: Well, the setting of this book is on a farm, and, you know, we live in inner-city Minneapolis, so not only have most of them not been to a farm, they also don't understand the process of how [animals become food] …

MT: So you provided that background for them by giving them quality feedback. So when she was worried about the pork being left out and possibly spoiling, through questioning her and giving her feedback, you helped her make the connection that most moms do take care of their families so they wouldn't have left meat out, helping her think through that and pushing her a little bit. But think about when the kids were acting out their characters. Talk to me a little bit about how the students could have given feedback to one another about that.

CT: I think it didn't go deep enough. And part of that was time. I was kind of debating, 'Should I save this activity for the next time we meet?' But knowing that it would be several days … Or, if I would have backed up a little bit, I could have given them some questions to think about in listening to their mates' response.

MT: And then with those questions, for them to provide feedback to …

CT: … to each other, yeah.

MT: It would then strengthen their thinking, because they're hearing that feedback. For students to be affirmed by one another is also just very important. They seem to be a very cohesive group of kids and so …

CT: They would have been very comfortable with that, I think.

Sidebar 5. What Does Effective "Feedback" Look Like? Portion of a Post-Conference Related to the "Area for Refinement" - *continued*

MT: I think they're pretty comfortable with just about anything. But so, just thinking about, when you're planning your lessons, providing opportunities for students to give that quality academic feedback to one another.

CT: So, can you think of a question I could have prefaced ahead of time to help them with that, to help the students' give academic feedback to each other?

MT: Well, let's do that together. Think about that segment of the lesson where they were acting out the characters. You were asking them to make inferences about how the character was thinking and feeling. Okay. So what could have been your focus question, if you want to call it that, that they could have been thinking about with that?

CT: [Long pause] Well, it could even have been, 'Are they giving feelings and thoughts, or are they just retelling what happened?'

MT: Exactly. Because I think you had to redirect [child's name], because what you were asking her was, 'Is that an inference? Are you projecting how that character feels? Or is that something that's here in the book?'

CT: So doing that as a way for them to evaluate each other's responses and provide feedback, and then be able to justify their own responses.

MT: Or it could have been something as simple as: 'As your partner is acting out the character, listen to the inferences they're making. And we know what good inferences are, and we know the things we use to make those good inferences. And I want you to give feedback to your partner on …' And it may have even been having a little rubric for them as to what are the qualities of good inferences, just something real short, not even anything big, about the qualities of a good inference. And then looking to give feedback to their partners. That would really strengthen their knowledge.

CT: Especially because that's something we've been working on as far as inferring in different ways … So that could be something to connect to: 'See, when I arrive at an inference, I need to use my knowledge plus what's in the text.'

MT: That's exactly right. You could have said, 'Okay, once they've made their inference, give feedback about what kind you think that was and what you think they used to arrive at that.' I think that would have really strengthened the lesson. Think about, and this may be a little more difficult, think about a lesson you're going to do today.

CT: Actually, that's what I was thinking about! [Laughter]

MT: Good! So how do you think you could get the kids to give quality feedback in a lesson you're going to do today?

CT: When I have them do their 'Think-Pair-Share,' I'm going to have them set up, not only are they going to share their thinking with their partners, but the partners will give them feedback on whether they answered the question.

MT: That will definitely strengthen the lesson. Now, thinking about what we talked about here today, tell me how you think this is going to impact how you plan lessons in the future and how you set your lessons up.

Sidebar 5. What Does Effective "Feedback" Look Like? Portion of a Post-Conference Related to the "Area for Refinement" - *continued*

CT: Well, I'm going to add an extra question in before I do a 'Think-Pair-Share.' Rather than just giving them a question to think about before we do the pairing, then I'll say, now when you share this thinking with your partner, they will give you feedback … So I'm going to put an extra step in between the 'thinking' and the 'pairing.' Because if we do it before the thinking, then they won't be doing the thinking about …

MT: Right … And I think you're such a strong model, that as you model the thinking for them, even saying for them, 'You know guys, you hear me model for you, and what I'd like you to do is do that for each other.'

CT: And probably before I do that I'll model giving some feedback on my end, so they have a common experience: Did she do this and where did it come from and is it a good inference? Because then we can debrief the whole process together before I ask them to do it with partners.

LESSON 5: Providing teachers with targeted support following post-conferences can help them improve more rapidly.

In TAP schools, evaluators have a variety of ways to follow up with teachers after observations and post-conferences, including "real time coaching" through demonstration lessons or team teaching during actual class-time in the teacher's own classroom.

The particular form that such support takes depends on the needs of the teacher in question. For example, if a master teacher identifies *Lesson Structure and Pacing* as an area for refinement during a post-conference, he or she might meet with the teacher to review an upcoming lesson plan and then visit the classroom to team-teach that lesson to students, explicitly modeling how to maintain appropriate pacing and make key transitions. In another instance, the master teacher might visit a teacher's classroom for only 15 minutes in order to demonstrate effective *Academic Feedback* during the appropriate chunk of an upcoming lesson.

Providing individualized support after observations can be important for a number of reasons: In some cases a teacher literally needs to see and hear what a particular Rubric skill or behavior "looks like" and "sounds like" in order to apply it. In other cases, a teacher might try to incorporate a new skill or behavior, but not succeed for some reason he or she cannot identify; in that case, the teacher needs help troubleshooting the problem. Or a teacher might simply need additional rounds of feedback as he or she works to improve on a particular dimension of instructional practice.

In addition, providing intensive support after evaluation post-conferences sends a very clear signal to teachers that the evaluation system is meant to support them, not just to "measure them." Such signals can help substantially increase teacher acceptance of and investment in a new evaluation system, especially when most teachers initially will receive lower ratings than in the past. (See Lesson 9 for more on this topic.)

Obviously, such intensive follow-up coaching is much easier to provide in schools that have adopted all elements of the TAP system—one in which a cadre of school-based master and mentor teachers is granted explicit authority and responsibility for a wide range of school functions related to evaluation, professional development, and school improvement.

However, other districts and schools could provide some follow-up support by leveraging existing resources. For example, many districts have invested heavily in instructional coaching over the past few years. Even if such coaches are not given a role to play in evaluation, they still could model skills and behaviors related to the evaluation rubric during their coaching sessions if they have access to information about areas for refinement identified during observations.

LESSON 6: Teacher evaluation and professional development need not be hermetically sealed off from one another; in fact, education leaders should take deliberate steps to ensure they are tightly integrated.

Integrating teacher evaluation with professional development can help ensure that evaluation meets the dual goals of better measurement for accountability and intensive support for improvement. However, policymakers need to understand that integrating evaluation and professional development does not come easily and does not result from simply talking about the importance of "alignment." It cannot be an afterthought and can only be accomplished through specific design decisions.

Sidebar 6. Brief Overview of TAP's Approach to Professional Development

Unlike the fragmented and disconnected approach to professional development still common in most schools, the TAP system provides teachers with a highly structured and focused form of professional development that is ongoing, job-embedded, collaborative, driven by analysis of student achievement data, and led by expert instructors. In TAP, school-based instructional experts called "master teachers" and "mentor teachers" have explicit responsibility for planning and leading a range of inter-related professional development activities. While the professional development structure is common across TAP schools, the content is entirely driven by careful analysis of student and teacher needs in any given school. Typical professional development activities include:

Cluster Groups. TAP restructures the school schedule to provide time during the regular school day for groups of teachers to collaborate on analyzing student data and learning new instructional strategies to improve student learning. Cluster groups of five to eight teachers led by a master or mentor teacher meet for 60 to 90 minutes a week. Strategies are selected by master teachers based on detailed analyses of student achievement data and are only introduced to teachers in the cluster group after the masters successfully "field test" and refine the strategies in actual classrooms so they can demonstrate student learning gains. After master teachers introduce a new strategy, teachers use the strategy in their own classrooms, then return to cluster meetings with pre- and post-data from formative assessments so that group can discuss how well the strategy worked and refine it further if necessary.

Individualized Coaching. Master and mentor teachers regularly visit teachers' classrooms to provide highly intensive and personalized coaching that can take a wide variety of forms, from teaching demonstration lessons to modeling specific instructional strategies or skills to team teaching. For example, master or mentor teachers often visit classrooms to coach teachers on a new instructional strategy after introducing it during a cluster group meeting. Coaching can take place outside the classroom, too: Mentor or master teachers can meet with teachers to brainstorm, troubleshoot, collaborate on lesson planning, review student work, provide feedback on teachers' plans and ideas, or—most often—to review and discuss how a lesson went.

An important prerequisite for integrating evaluation and professional development is selection of an evaluation instrument that provides a coherent vision for effective teaching that is detailed enough to provide teachers with a roadmap toward excellence in each instructional skill. Evaluation results can then be used by the teacher, and those providing professional development to that teacher, to move from one level on the rubric to a higher level in any particular skill.

Beyond selecting an appropriate instrument, leaders must ensure that teachers are given sufficient opportunities to develop deep understanding of the vision for effective practice embedded in the rubric or framework. During the first year of TAP implementation, master teachers and mentor teachers spend a considerable amount of time helping teachers become familiar with the TAP Rubric. As a result, the Rubric comes to form the basis for a shared language about the fundamentals of effective practice that is often lacking in American schools.

Once that prerequisite has been fulfilled, districts and schools have many additional ways to integrate evaluation and professional development, especially in places that are moving toward "job-embedded" approaches and relying less on workshops offered by outside providers. Of course, specific strategies will depend greatly on the particular form the professional development takes in any given district or school—and that still varies greatly across the nation and even within states.

However, TAP schools have found that several increasingly popular professional development activities offer perfect opportunities for integration, including the following:

1. Common Instructional Planning Time. According to the most recent federal data, 70 percent of teachers participate in regularly scheduled collaboration with other teachers on issues related to instruction, an activity that can account for two-thirds of professional development spending at the school level.[19] In TAP schools, this takes the form of weekly "cluster group" meetings led by master and mentor teachers. (See Sidebar 6 for an overview of professional development in TAP schools.)

During cluster meetings, master and mentor teachers take advantage of many opportunities to reinforce skills from the TAP Rubric. For example, if they are introducing a new instructional strategy teachers will implement (for example, a technique for improving reading comprehension), they will point out specific ways that particular rubric skills will help teachers use the new strategy well. As they model the new strategy for teachers (i.e., demonstrate the technique by role playing the part of a teacher using it the classroom), they will incorporate specific rubric behaviors, stepping "in and out of character" to point out how those behaviors will help teachers implement the strategy effectively in their own classrooms.

2. Instructional Coaching. Nationwide, 46 percent of American teachers reported that they received individualized coaching or mentoring or provided it to colleagues in 2003-04,[20] an activity that offers another opportunity for integrating evaluation and professional development.

In TAP schools, master and mentor teachers provide intensive coaching to teachers in their own classrooms on a regular basis; coaching can take the form of modeling particular instructional strategies, giving demonstration lessons, or team teaching. In some cases, they take advantage of coaching time to work with teachers on a particular rubric skill, most often the "area for refinement" identified by an evaluator following the teacher's most recent observation. In other cases coaches are helping teachers master a new instructional strategy introduced during the weekly cluster meetings, and they can reinforce and model particular rubric skills, including the area for refinement, as they do so.

3. Peer Observations. Nearly two-thirds of American teachers (63 percent) reported that they had opportunities to observe other teachers or be observed by other teachers in 2003-04.[21] However, experts suggest that the value of such activities

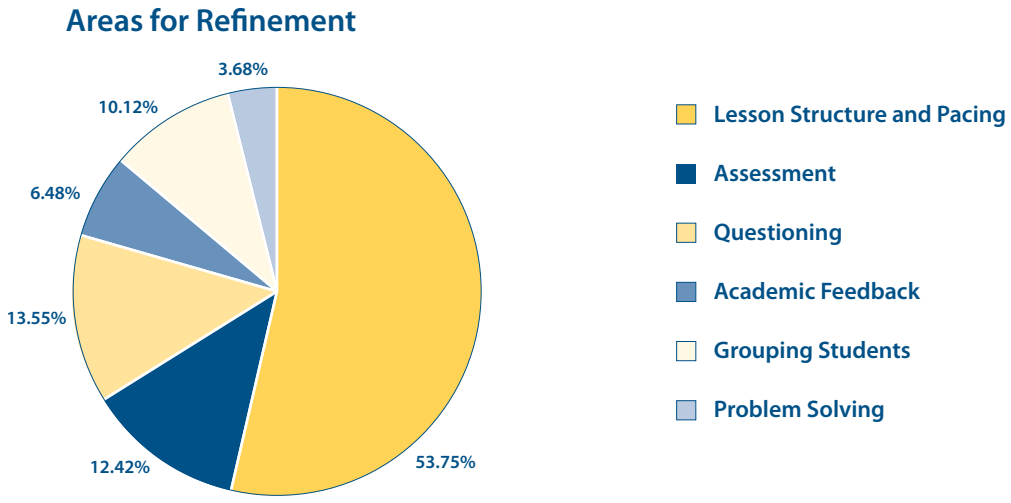
is often severely hampered by the lack of a shared understanding of what effective instruction should look like across classrooms[22]—a gap that can be filled with the right kind of evaluation rubric. (In fact, adopting a framework for thinking about good instruction that is different from the evaluation rubric would create confusion and unnecessary fragmentation of evaluation and professional development.)

Finally, the data management system that tracks evaluation results should be designed to produce standardized reports that support professional development planning as well as reports that help monitor reliability and accuracy of scores. In TAP's case, the CODE data system offers many reports that summarize and break down SKR scores for teachers in a school or district, display trends in those scores, and aggregate information about areas for reinforcement and refinement identified in evaluation post-conferences. (See Figure 8.) Members of the school leadership team can use such reports to better plan professional development activities across the school or within cluster groups, as well as to track individual teacher progress.

According to Monica Knauer, a master teacher at the Dwight D. Eisenhower Academy of Global Studies in New Orleans, "We use CODE to check on inter-rater reliability, but we can also use it to analyze overall strengths and weaknesses within our school on the Rubric. Early on, the data told us we needed to hone in on teachers' lessons being better aligned with standards and objectives, but after that we moved on to the student *Questioning* element of the Rubric because the CODE data told us that area was not really strong. So we embedded *Questioning* into our weekly cluster meetings, pointing it out and modeling it for teachers, even as they were learning a new instructional strategy."

Figure 8. Example of CODE Chart to Support Professional Development Planning

The following chart shows how often particular indicators on the TAP Rubric have been chosen as the area for refinement during post-conference. In this case, more than half of observations have led to *Lesson Structure and Pacing* being targeted as an area for improvement, suggesting that master and mentor teachers might want to pay particular attention to this skill in upcoming professional development activities such as cluster group meetings.



LESSON 7: Evaluation systems that include teacher leaders as evaluators can indeed produce non-inflated ratings, even when such individuals work in the same schools as evaluatees. This approach offers distinct benefits for improving teachers' effectiveness.

One of the most important decisions for designing an evaluation system is who will do the actual evaluating. As described in Lesson 1, this is a decision that can be fraught with trade-offs. Some experts who are primarily concerned with using evaluation to measure performance, i.e., produce data for accountability or other personnel decisions, suggest that evaluators should have no professional relationship with evaluatees. They worry that such relationships will cloud evaluators' objectivity and lead to score inflation. Moreover, some policymakers worry that asking teachers to objectively evaluate other teachers is simply asking too much.

However, as described in Sidebar 2, a recent analysis of SKR and value-added scores from across TAP schools suggests that TAP's reliance on school-based, expert master and mentor teachers has not led to the kind of inflated scores evident in traditional evaluation systems—in fact, quite the opposite. In addition, relying on such school-based teacher leaders offers significant advantages when it comes to leveraging evaluation to promote teachers' individual growth and improvement.

For example, as previously described, choosing the most effective areas for reinforcement and for refinement after an observation often calls for significant contextual knowledge: Which areas can best support what teachers are learning in professional development? Which areas best fit the specific goals for student achievement in the schoolwide improvement plan or the grade level or subject area? Which areas would best benefit the teacher and his or her students alike given the school's curriculum and upcoming units in the curriculum?

Moreover, using internal observers allows TAP schools to weave the evaluation rubric into the professional development activities at many different points, from cluster group meetings to individualized classroom coaching. While it is not inconceivable that external observers could follow up with evaluatees to provide coaching in the classroom, chances are they would find it far more difficult to do so. In TAP schools, coaching is a part of teachers' and administrators' daily lives; on a practical basis, it is very easy for evaluators to use such opportunities to provide targeted support following observations.

Ultimately, the TAP system has demonstrated that trained and certified teacher leaders are capable of taking professional responsibility for validly and reliably evaluating fellow teachers—even when they work in the same buildings as their evaluatees and are responsible for providing professional development to their evaluatees. Perhaps that should not come as such a surprise: In TAP schools, master and mentor teachers are proven professionals who have *already demonstrated* that they can be trusted with an even more sacred responsibility, successfully educating students to very high levels.

If a state or district plans to spend money on a teacher "career ladder" initiative, it should consider including evaluation among the responsibilities for at least some of those new roles. Career ladders are expensive: It makes tremendous economic sense to leverage that investment to support improvements in other areas, such as evaluation, especially at a time of shrinking education budgets.

One caution, however: If policymakers decide to rely on expert teacher leaders (or "career ladder" teachers) to help conduct evaluations, they should be sure to incorporate criteria related to that job function into the recruitment and selection process for such positions. Evaluation requires a particular skill set that complements, but is distinct from, effective classroom teaching and instructional coaching.

LESSON 8:

For evaluation systems to support individual growth as well as accurate measurement, evaluation instruments must be suitable for both purposes—comprehensive enough to capture essential elements of effective instruction, but not so broad as to sacrifice depth for breadth or to become unwieldy in practice.

Often it seems like the selection of a rubric or framework dominates conversations about improving teacher evaluation, as if upgrading evaluation were mostly a matter of replacing the crude checklists used in traditional systems with more detailed instruments such as the TAP Rubric. By now it should be clear to readers that simply swapping out evaluation instruments is not an adequate solution to the challenge of improving teacher evaluation.

However, the choice of instrument (whether it is called a "rubric" or "framework" or "teaching standards") is important in several ways:

1. The instrument must be objectively validated. The rubric or framework must incorporate a vision for effective instruction that has been *empirically linked* to improvements in student learning. In other words, there must be sound evidence that when evaluators apply the instrument accurately, higher scores on the instrument predict better outcomes for students. In TAP's case, the 2010 NIET analysis described previously in Figure 3 supports that claim.[23] In addition, during development of the Rubric, the Milken Family Foundation conducted studies to validate that teachers who scored higher on targeted instructional practices had students who achieved higher learning gains; the results of one such study were published in a peer-reviewed academic journal, *Economics of Education Review.*[24]

2. The instrument must be ***practical***. The framework or rubric must be able to serve as a manageable tool for collecting and scoring evidence of instructional practices as a basis for reliable scoring. Analyzing and evaluating classroom lessons is a challenging endeavor that requires a high level of training and skill. An instrument with too many domains and indicators to track and document will frustrate the efforts of even the most conscientious observer. Moreover, in order for multiple observers to achieve acceptable levels of inter-rater reliability for summative scores, they must do so for every discrete indicator on which teachers must be scored.

Therefore, instruments that seek to describe every conceivable aspect of instructional practice or to include every useful teaching technique ("everything-but-the-kitchen-sink rubrics") might not support reliable and accurate scoring during day-to-day application in the field even if they do so in controlled research settings. The TAP Rubric includes 19 areas in which teachers are scored during each observation, plus an area related to Professional Responsibilities on which teachers are scored at the end of each year. (See Figure 9 and Figure 10 for a list of domains and indicators on the TAP Rubric and for an in-depth example of one indicator, *Academic Feedback,* respectively.)

Figure 9. Domains and Indicators in the TAP Rubric
(*TAP Teaching Skills, Knowledge, and Responsibilities Performance Standards*)

INSTRUCTION	THE LEARNING ENVIRONMENT
1. Standards and Objectives 2. Motivating Students 3. Presenting Instructional Content 4. Lesson Structure and Pacing 5. Activities and Materials 6. Questioning 7. Academic Feedback 8. Grouping Students 9. Teacher Content Knowledge 10. Teacher Knowledge of Students 11. Thinking 12. Problem Solving	1. Expectations 2. Managing Student Behavior 3. Environment 4. Respectful Culture
DESIGNING AND PLANNING INSTRUCTION	**RESPONSIBILITIES (Assessed at the end of the school year)**
1. Instructional Plans 2. Student Work 3. Assessment	1. Staff Development* 2. Instructional Supervision* 3. Mentoring* 4. Community Involvement* 5. School Responsibilities* 6. Growing and Developing Professionally 7. Reflecting on Teaching

* Denotes an indicator that only applies to master teachers and mentor teachers

Figure 10. Example of One Indicator in the TAP Rubric

	Exemplary (5)	*Proficient (3)*	*Unsatisfactory (1)*
Academic Feedback	» Oral and written feedback is consistently academically focused, frequent, and high-quality. » Feedback is frequently given during guided practice and homework review. » The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. » Feedback from students is regularly used to monitor and adjust instruction. » Teacher engages students in giving specific and high-quality feedback to one another.	» Oral and written feedback is mostly academically focused, frequent, and mostly high-quality. » Feedback is sometimes given during guided practice and homework review. » The teacher circulates during instructional activities to support engagement and monitor student work. » Feedback from students is sometimes used to monitor and adjust instruction.	» The quality and timeliness of feedback is inconsistent. » Feedback is rarely given during guided practice and homework review. » The teacher circulates during instructional activities, but monitors mostly behavior. » Feedback from students is rarely used to monitor or adjust instruction.

3. *The instrument must be **meaningful***. The rubric or framework must offer a coherent vision of effective practice that makes sense to teachers and contributes to conversations about improving teaching and learning at the school and classroom levels. Here again, the challenge is to incorporate sufficient breadth without sacrificing the kind of depth that can help teachers acquire a vivid understanding of what performance looks like at various levels of expertise in each domain or area.

Indeed, teachers in TAP schools often talk about how the Rubric has encouraged a rich, meaningful, shared language about effective instruction that previously was lacking, enabling teachers to better collaborate and improve both instruction and student achievement. That could not happen if the Rubric were either not comprehensive enough to capture the essential elements of effective instruction or so impractically extensive that each indicator or domain could only be understood at a very shallow level. Achieving the right balance between depth and breadth is challenging but important.

4. *The instrument must capture the **full range of performance***. NIET has learned that a seemingly very technical decision has enormous implications in the field—how many performance levels to build into the evaluation instrument. Some might argue that fewer levels might make it easier to achieve higher rates of inter-rater reliability, suggesting that performance levels be limited to four or even three. However, TAP's experience suggests that ratings scales with five performance levels are better able to support multiple goals for teacher evaluation.

Importance for Measurement: A scale of 1 to 4 creates the problem of defining the expected level of performance, particularly in systems that use student academic growth as one of multiple measures. For example, value-added scores define an "expected" level of performance based on predicted growth: If there are only four possible ratings, i.e. no natural middle level, would level 2 or level 3 represent that expected level of growth? Should there be two "lower than satisfactory" levels and only one "higher than satisfactory" level, or vice versa?

A scale of 1 to 3 would solve that particular dilemma, but at the expense of not being able to differentiate between exceptionally effective teachers and "above satisfactory" teachers, or between exceptionally low-performing teachers and merely "below satisfactory" teachers. In effect, a scale of 1 to 3 introduces what measurement experts refer to as "floor and ceiling effects," which make it impossible to capture performance at the high and low end of the range and limit the use of ratings to make informed personnel decisions.

Importance for Individual Growth: TAP's 1-to-5 scale enables the system to provide valuable feedback to more teachers along a much wider range of performance levels, including experienced and relatively expert teachers who would naturally score at the highest level in systems with fewer categories. In the TAP system, scoring a 5 in any area or overall connotes not just satisfactory performance or even superior performance, but truly exceptional performance. Because 5's are very difficult for even the best teachers to achieve, the evaluation system provides *very nearly every TAP teacher* with "stretch goals" that encourage him or her to continue to improve. That in turn makes the evaluation system relevant, meaningful, and useful to a much larger proportion of teachers in any given school or district.

The value of providing goals for continuous improvement to all teachers can hardly be overestimated: Ensuring that an evaluation system does not just focus on "bad teachers," but challenges and supports all teachers to improve engenders not only more "buy-in" but, over the long term, more improvement. A recent cohort analysis conducted by NIET found that teachers' SKR scores grew at all performance levels; while lower-performing teachers improved the most, on average, teachers at every performance level improved.[25] Such a system benefits students as well as teachers: The more teachers who have room to grow and who receive follow-up support to grow, the more students will receive the benefit of improvements in instructional practice and effectiveness over time.

Consider the words of Lynn Kuykendall, a master teacher at Clinton Elementary School in South Carolina who was quoted in a report published by the Center for American Progress last year: "Prior to TAP, I can honestly say that in the 25 years in this school, there were some years where no one in an administrative or certainly in an instructional leadership role would ever walk in to evaluate me. When I would question that, they'd say, 'Why? You're doing a great job. You're nationally board certified.' But that doesn't mean I shouldn't be evaluated. How do you know? And there were personal goals I wanted to work on so that I could keep improving. But if I went a whole year without feedback from an evaluation, it would be hard to work on those goals."[26]

LESSON 9: Successful and sustained implementation of new teacher evaluation systems requires careful attention to school culture and to the "human side" of performance evaluation.

The 2009 study of district evaluation systems by The New Teacher Project (TNTP) found that, even in systems that produced clearly inflated ratings, teachers' expectations were even more inflated.[27] In six districts where teachers could be rated at multiple levels, half of the teachers (both tenured and untenured) who did *not* receive the very highest rating thought they deserved to do so. In a subset of districts where researchers asked teachers to rate their own instructional performance, more than 43 percent rated themselves a 9 or higher on a scale of 1 to 10.

"This creates a culture in which teachers are strongly resistant to receiving an evaluation rating that suggests their practice needs improvement," the TNTP researchers concluded. "Schools then find themselves in a vicious cycle; administrators generally do not accurately evaluate poor performance, leading to an expectation of high performance ratings, which, in turn, cause administrators to face stiff cultural resistance when they do issue even marginally negative evaluations."

Policymakers should make no mistake: Breaking that cycle requires much more than just new evaluation tools, procedures, and training. It requires deliberate and ongoing efforts to help teachers understand why professional standards have to be raised and how they and their students will benefit and be supported—as well as a degree of stubborn insistence on maintaining expectations among members of school leadership teams. Over the longer term, the goal must be not just to get teachers to accept the new system, but to help them see how much room for growth exists and to internalize the new higher standards so that they can develop the capacity to critically reflect on their own performance.

Over the past ten years, NIET staff members, expert technical assistance providers, and school-level practitioners have developed a variety of tools to communicate and uphold high evaluation standards, to facilitate teacher understanding, and to support culture change.

One formal mechanism that helps teachers internalize higher standards is the self-evaluation built into the TAP system: After each observation, teachers score their own lesson based on the TAP Rubric and bring that self-evaluation to share in the post-conference. To ensure that the self-evaluation is a meaningful exercise, the TAP system counts such scores as 10 percent of the annual SKR average score. In addition to promoting a reflective attitude toward personal performance, this gives teachers a structured way to compare their own expectations for instructional performance with ratings given by trained and certified evaluators. (The CODE data system can produce bar charts comparing teachers' self-scores with evaluators' scores on the 19 Rubric indicators.)

When TAP's evaluation system is first introduced, teachers tend to score their own lessons much higher than do the evaluators, but that changes over time as teachers become familiar with the Rubric and calibrate their own expectations with it. "What I've encountered is that, over time, my teachers become harder on themselves and I'm often slightly higher than their self-scores," says Alma Velez, principal of Anson Jones Elementary School in Bryan, Texas. "If the score is within one

point, I bring it up with them just to ensure that they understand, but if it is more than one point, I look at the indicator and we study it together. I show them the evidence that I saw, and they share their evidence for that score. So it provides really meaningful conversations about performance and classroom instruction."

Even so, during the first year of implementation, teachers are often reluctant to accept lower evaluation ratings than they are used to receiving. School leadership team members find themselves having many conversations with teachers about the new expectations, trying out different metaphors and using different examples to make the point.

"When we started out, teachers immediately thought, 'How do I get a 5 on everything?'" says Eric Matheson, principal at Chapman Elementary School in South Carolina. "But the state TAP folks said, 'That's like walking on water.' It was hard for [teachers] to understand given the way evaluations had worked in the past. I had a lot of conversations where I said, 'That would mean there's no room for improvement. Can you honestly say that is true every day for every student on every objective, and that it would not be possible to improve in any area?'"

Sue Way, an executive master teacher with the Louisiana Department of Education, recalls that when she first implemented the TAP system as a school principal, her biggest worry was that teachers would not accept the new evaluation system. "After we spent the first 12 or 14 weeks studying the Rubric, I told teachers to pick any lesson they've taught and do the self-evaluation as an informal exercise. I just wanted to see where we were. And they were rating themselves all 4's and 5's! So we had to have a talk. I had them compare their self-evaluation ratings with our school's student proficiency results and asked, 'If we're *all* so wonderful in teaching, why aren't we off the charts in learning?' After that, and with more practice, their scores became better calibrated. Different leadership teams handle it in different ways, though we give them lots of advice from the state level."

Laura Roussel, a former master teacher at Lowery Intermediate School in Ascension Parish, Louisiana, cautions that introducing a new evaluation system requires spending a great deal of time helping teachers fully understand what performance looks like at each level, particularly the proficient level. "We really spent a lot of time in our cluster meetings that first year describing and modeling and promoting level 3 as 'rock solid' instruction," she recalls, "really furthering their understanding of what that looks like and sounds like."

Evaluators also learn useful strategies for delivering critical feedback during post-conferences. Giving someone critical feedback is difficult under the best of circumstances, and doing so for the first time in schools where teachers have never received such feedback can be doubly difficult. Therefore, TAP evaluators take pains to "focus on the performance, not the performer." Such an approach does not just make honest feedback easier to accept; it also helps teachers develop the kind of "growth mindset" that cognitive researchers have found to be another prerequisite for dramatically improving individual performance.[28]

Finally, because of the extensive classroom coaching that master and mentor teachers provide, teachers have plenty of opportunities to see expert instructional performance for themselves—teaching that would score high on TAP Rubric indicators and is taking place "in real time" with the very same students.

LESSON 10: Districts and schools need substantial external support and technical assistance to successfully implement more sophisticated teacher evaluation systems.

Perhaps the greatest danger in reforming teacher evaluations is that policymakers will underestimate or—for financial reasons—simply overlook how much intensive external support it takes to help schools implement such systems with fidelity. Because of significant strategic investments made by the Milken Family Foundation and NIET, TAP schools and

districts have access to an unusually rich set of tools, technical assistance, and support to ensure they can get evaluation right. As a recent paper by Jessica L. Lewis and Matthew G. Springer noted, the technical assistance provided by NIET has evolved from a purely face-to-face model, to one in which training content is electronically delivered, to one that enables TAP participants to share information with one another.[29]

Schools and districts can rely on the following kinds of support when implementing the TAP evaluation system:

TAP System Training Portal. In fall 2010, NIET launched a new Web-based training portal available to all TAP schools. The Portal includes a wide range of helpful tools and materials related to teacher evaluation, such as guidebooks and manuals, evaluation protocol templates, and—most significantly—the full video library of nationally-rated lessons described previously. Included in the Portal are a series of resources including:

> ***Video Library of Rated Lessons.*** Over time, NIET has developed a library of videotaped classroom lessons that can be used for evaluator training and, once training is over, by school leadership teams during the year. The videos illustrate different levels of performance on areas of the TAP Rubric. Each lesson has been rated by "national raters," and each video is accompanied by descriptions of the evidence the national raters used to justify their ratings.
>
> ***TAP CORE Training and Certification Sessions.*** NIET has developed a standardized eight-day series of training workshops and certification assessments, the TAP CORE Trainings. Four days are focused on evaluation and are required for all TAP evaluators. (See the Appendix for a more detailed description of these sessions.)
>
> ***Comprehensive Online Data Entry (CODE) System.*** CODE allows TAP leadership teams to input teacher evaluation ratings into a Web-based system for collecting, storing, and analyzing evaluation results. As previously described, the CODE system can produce a very wide range of tables and charts that school leadership teams can use to monitor inter-rater reliability and score inflation and that master teachers can use to plan schoolwide and cluster-group professional development activities.
>
> ***Strategies Library.*** A growing library of field-tested instructional strategies, submitted by TAP master and mentor teachers across the country and reviewed by NIET, provides leadership team members with instructional strategies organized by content and grade level to support them in planning cluster group professional development.
>
> ***Training Modules on Specific Instructional Skills.*** Career teachers have access to videotaped excerpts from high performing teachers' classroom lessons that illustrate specific instructional skills related to the Rubric. These video resources are accompanied by support materials to enable career teachers to work on improving their skills.

National Conferences and Summer Training Institutes. The annual National TAP Conference and state and regional TAP Summer Institutes offer multiple training sessions related to teacher evaluation. For example, the 2010 national conference offered several opportunities to attend a session called "Using CODE to Monitor Inter-Rater Reliability," during which expert evaluators explained various CODE charts and provided participants with guided practice in analyzing the CODE charts to identify patterns that might signal problems with inter-rater reliability.

Dedicated National and State-Level Support Staff. NIET staff members provide significant assistance to TAP schools. And in three states with large concentrations of TAP schools—Louisiana, South Carolina, and Texas—the state department of education houses an office that provides support for TAP leadership teams. State support staff members include "executive master teachers" who spend a significant portion of their time visiting TAP schools and working with principals and master teachers to troubleshoot problems, offer advice, and formally evaluate the leadership team's performance, including its implementation of the TAP evaluation system.

Conclusion

The TAP system's successful reform of teacher evaluation reveals that policymakers need not trade relevance for rigor, or vice versa, in teacher evaluation. It is possible to design robust evaluation systems that equally produce valid measurement of teacher effectiveness and simultaneously help individual teachers become more effective over time.

However, crafting such sophisticated evaluation systems requires a great deal of thought about design trade-offs, and implementing them successfully requires a significant investment in time and resources. "The journey to truly superior performance is neither for the faint of heart nor for the impatient," Ericsson advises professionals who hope to develop high levels of expertise in their fields. "The development of genuine expertise requires struggle, sacrifice, and honest, often painful, self-assessment. There are no shortcuts."[30] We offer the same caution to policymakers who hope to radically reform teacher evaluation systems in the United States.

Appendix: TAP's Training and Certification for Evaluators

Training and certification for school leaders to conduct teacher evaluations takes four days total, two days for Evaluation A and two days for Evaluation B.

Evaluation A

1. Understanding the TAP Rubric

Participants are guided through an in-depth examination of the TAP Instructional Rubric and how it is used in the evaluation process. The Rubric includes 19 indicators of effective teaching that are evaluated during classroom observations, each with detailed performance descriptors and an explanation of what represents "exemplary", "proficient," and "unsatisfactory" teaching on a scale of 1 to 5. The training includes group discussions about specific indicators on the Rubric and what proficient teaching looks like for that indicator.

2. Applying the TAP Rubric

After understanding the Rubric, participants watch videos of actual classroom lessons; "script" or record evidence from the lesson on two or three key Rubric indicators; discuss this evidence with colleagues; and agree on a score for the indicators reviewed. The training provides evaluators the opportunity to work with colleagues to build their common understanding of proficient teaching and to compare their scores against those given the same lessons by TAP's "national raters." Participants then have the opportunity to review additional indicators from the Rubric, continue to practice scoring by viewing lesson videos, collecting evidence through scripting and observation, categorizing the evidence, and scoring.

3. Planning and Conducting Pre- and Post-Conferences

Evaluator training also includes explanations, outlines, and videos that help participants understand how to conduct high-quality pre-conferences and post-conferences with teachers. Participants view videotaped pre- and post-conferences between evaluators and teachers (along with watching the relevant lessons) and engage in rich discussions with fellow participants about what they observed.

After the Evaluation A training, participants are encouraged to return to their schools and practice applying their new understanding of the TAP Rubric and the evaluation process with colleagues and teachers. New evaluators often practice evaluating classroom lessons in their own school building in pairs with another member of their leadership team. During this period, the scores generated by evaluators are not counted toward teachers' formal, summative evaluation results. Optimally the evaluators-in-training will have multiple opportunities to practice the evaluation process in their own schools before attending Evaluation B, the second and final training and certification session.

Evaluation B

1. Reinforcement and Additional Practice

After completion of the Evaluation A training and practicing what they learned in their own schools, participants return for the second portion of the evaluation training known as Evaluation B. In the Evaluation B training participants have several additional opportunities to view videotaped lessons and to practice scripting and scoring lessons through the lens of the TAP Rubric. The participants also review the pre- and post-conference process and practice developing written conference plans based upon their observation of the videotaped lessons. After the participants have had several opportunities to practice using the Rubric and applying their learning, they take the TAP Evaluator Certification test.

2. Certification to be an Evaluator

At the end of the training, participants must take a certification test that involves watching an hour-long videotaped classroom lesson, collecting evidence to justify the scores they assign on each indicator of the Rubric, assigning scores for each of the 19 indicators, and writing a post-conference plan that demonstrates their ability to analyze the lesson and discuss objectives for reinforcement and refinement with the teacher. If a participant is able to score the lesson within one point of the national raters' score and complete a satisfactory post-conference plan, he or she receives a one-year certification to formally evaluate teachers in TAP schools. However, evaluators must take and pass the certification test annually in order to be recertified.

References

1. 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.
2. Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness.* New York: The New Teacher Project.
3. Jerald, C. (2009, July). *Aligned by Design: How Teacher Compensation Reform Can Support and Reinforce Other Educational Reforms*. Washington, DC: Center for American Progress.
4. Weisberg, Sexton, Mulhern, & Keeling.
5. See for example Halverson, R., Kelley, C. & Kimball, S. (2004). Implementing Teacher Evaluation Systems: How Principals Make Sense of Complex Artifacts to Shape Local Instructional Practice. In Wayne K. Hoy and Cecil G. Miskel, eds., *Educational Administration, Policy and Reform: Research and Measurement*. Greenwich, CT: Information Age Publishing.
6. Daley, G. & Kim, L. (2010, August). *A Teacher Evaluation System That Works*. Santa Monica, CA: National Institute for Excellence in Teaching.
7. See for example Silva, E. (2008, April). *The Benwood Plan: A Lesson in Comprehensive Teacher Reform*. Washington, DC: Education Sector.
8. Ericsson, K.A, Prietula, M.J., & Cokely, E.T. (2007, June 1). The Making of an Expert: New Research Shows that Outstanding Performance Is the Product of Years of Deliberate Practice and Coaching, Not of Any Innate Talent or Skill. *Harvard Business Review*, July-August 2007.
9. Daley & Kim.
10. Kane, T., Taylor, E., Tyler, H., & Wooten, A. (2010, March). *Identifying Effective Classroom Practices Using Student Achievement Data*. Cambridge, MA: National Bureau of Economic Research.
11. Daley & Kim.
12. Measurement experts often use the term "bias" to describe this concept, as in "the scores are biased upward." However, since that word has a narrower and very specific connotation for many non-technical readers, we have used the term "accuracy" to avoid confusion.
13. Schacter, J. & Thum, Y.M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review, 23*, 411–430.
14. Daley & Kim.
15. Jerald summarizes recent studies on instructional coaching that point to an over-reliance on "soft feedback" as one of the biggest problems in many coaching initiatives.
16. Ericsson, Prietula, & Cokely.
17. Ericsson, K.A. (2008, November). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine 15*(11), 988-994.
18. Ericsson.
19. Jerald.
20. Jerald.
21. Jerald.
22. See for example City, E.A., Elmore, R, Tietel, L., & Fiarman, S. (2009). *Instructional Rounds in Education: A Network Approach to Improving Teaching and Learning*. Cambridge, MA: Harvard Education Press.
23. Daley & Kim.
24. Schacter & Thum.
25. Daley & Kim.
26. Jerald.
27. Weisberg, Sexton, Mulhern, & Keeling.
28. See for example Dweck, C.S. (2006). *Mindset: The New Psychology of Success*. New York: Ballantine Books.
29. Lewis, J.L. & Springer, M.G. (2009, December). *Effective Technical Assistance Principles: Lessons from Three Performance Pay Programs*. Washington, DC: Center for American Progress.
30. Ericsson, Prietula, & Cokely.

About the Authors

Craig D. Jerald
Craig Jerald is president of Break the Curve Consulting, which provides expertise to leaders and policymakers on issues related to education policy, communications, research, and practice. Prior to founding Break the Curve, he served as a principal partner at the Education Trust, and a senior editor at Education Week from 1996 to 2000. He also has worked at the U.S. Department of Education, and he began his career as a Teach for America recruit and middle school teacher in California's Long Beach Unified School District.

Kristan Van Hook
Kristan Van Hook is senior vice president for public policy and development at the National Institute for Excellence in Teaching (NIET). In this capacity, she develops and implements strategies to build support for NIET's education initiatives, with a focus on increasing teacher and school leader effectiveness through TAP: The System for Teacher and Student Advancement. She brings 20 years of experience in government and policy to her work with NIET.

Acknowledgements

We would like to thank the Joyce Foundation for providing support for this paper and for NIET's ongoing work to analyze the TAP system's lessons learned. This report reflects contributions from many individuals involved with TAP. The authors would like to thank the Louisiana TAP staff for providing valuable insights and for opening up many doors to illustrate the importance of technical assistance in supporting a rigorous evaluation system. Vicky Condalary spent several days answering detailed questions and taking one of the authors on a tour of TAP schools. The authors also would like to thank NIET staff for support with key areas of the report, including Lisa Shapiro, Sarah Shoff, Glenn Daley, Geneva Galloway and Jason Culbertson, among others.

Funding Provided By

The Joyce Foundation

70 W. Madison Street, Ste. 2750
Chicago, Illinois 60602
Phone: (312) 782-2464
Fax: (312) 782-4160
www.joycefdn.org

National Institute for
Excellence in Teaching®

1250 Fourth Street
Santa Monica, California 90401
Phone: (310) 570-4860
Fax: (310) 570-4863
www.tapsystem.org