Ensuring Accurate Feedback *from* Observations

• PERSPECTIVES ON PRACTICE •

BILL & MELINDA GATES *foundation*

ABOUT THE AUTHOR

Craig Jerald is president of Break the Curve Consulting, which provides technical assistance and strategic advice to organizations working to improve education for all students. Prior to consulting, Craig was a principal partner at the Education Trust and a senior editor at *Education Week,* where he founded the organization's first full-time research team. Craig began his career as a middle school teacher of language arts, history, and mathematics in California's Long Beach Unified School District.

Bill & Melinda Gates Foundation

Guided by the belief that every life has equal value, the Bill & Melinda Gates Foundation works to help all people lead healthy, productive lives. In developing countries, it focuses on improving people's health and giving them the chance to lift themselves out of hunger and extreme poverty. In the United States, it seeks to ensure that all people—especially those with the fewest resources—have access to the opportunities they need to succeed in school and life. Based in Seattle, Washington, the foundation is led by CEO Jeff Raikes and Co-chair William H. Gates Sr., under the direction of Bill and Melinda Gates and Warren Buffett.

For more information on the U.S. Program, which works primarily to improve high school and postsecondary education, please visit www.gatesfoundation.org.

March 2012

Contents

Executive Summary

Three Areas for Action

1. Build Observers' Capacity
2. Create Conducive Conditions
3. Monitor & Ensure Quality

As states and districts design new teacher evaluation systems, they are keeping classroom observations as a core component. But the observations are being redesigned to provide better feedback that can help all teachers improve. To accomplish that goal, school systems are replacing the crude "observation checklists" of the past with more sophisticated instruments that allow teachers and observers to identify strengths and opportunities for growth on multiple dimensions of teaching practice.

New research suggests that accurate feedback based on such observation instruments can be a powerful resource for improving teaching and learning. One recent experimental study found that giving secondary school teachers frequent observational feedback based on the Classroom Assessment Scoring System (or CLASS) boosted their students' achievement by the equivalent of moving from the 50th to the 59th percentile on Virginia's state tests (on a 100-point scale).

However, feedback that *inaccurately* classifies observed practices as strong or weak can entail significant opportunity costs for both teachers and their students. Therefore, school systems adopting new evaluation systems face a common problem: How to ensure observers' feedback and coaching avoid major errors in classification and instead are based on a *reasonably* accurate judgment of a lesson.

This brief offers examples and lessons from leading states, districts, charter management organizations (CMOs), and other education organizations working to provide teachers with accurate feedback from observations. We identified three broad areas for action: Build observers' capacity to conduct accurate observations, create conducive conditions for observing accurately in the field, and monitor observations periodically to ensure quality.

1. Build Observers' Capacity

Make sure that observers are well equipped to conduct accurate observations by providing them with the knowledge, skills, and tools to do the job well.

▶ Provide observers with **intensive** training.

Most school systems are contracting with external consultants to design and deliver the first round of training while building internal capacity to provide it in the future. Many are discovering that observers can struggle to transfer skills to the field if training relies exclusively on video-recorded lessons, so they are building in more opportunities for "live practice" in local classrooms.

Conduct a **certification assessment** and require periodic re-certification to confirm skills have not deteriorated.

Some school systems grade observers on a pass-fail basis depending on whether they can score within a certain range of a "gold standard." Others include a middle category, "conditionally certified," to identify staff members who can observe with additional supervision and who need extra support to reach full certification by the following year.

Offer useful **tools,** including an observation instrument that supports reliable measurement of teacher and student behaviors.

Observation instruments need to be comprehensive enough to capture a robust vision of effective teaching without becoming so extensive that they become unmanageable for observers. Based on lessons learned during pilots or first-year implementation, school systems are clarifying language, adding more precise descriptors, and streamlining instruments by collapsing redundant dimensions. School systems also are providing additional tools to support accurate observations, from LiveScribe audio pens to example-laden evidence guides.

Help observers **reinforce** their skills after training.

School systems are finding that observers can experience "drift" following training and certification, so they are using a wide range of strategies to help observers maintain and improve their skills. Strategies include:

- **Deep-dive training** for groups of observers focused on specific dimensions of the observation instrument;
- **One-on-one coaching** provided by school system leaders or expert consultants;
- **Paired observations** of live or video-recorded lessons; and
- **Group calibration** sessions based on live or video-recorded lessons, sometimes using videoconferencing, to allow large groups to view, score, and discuss a live lesson together.

2. Create Conducive Conditions

Even the most highly trained and rigorously certified staff members can fail to conduct accurate observations if they encounter significant obstacles in the field.

Ensure that observers have **manageable caseloads** given their other time commitments.

Asking principals to conduct too many observations might force them to cut corners in ways that undermine accuracy. Some school systems are reducing the number of observations for experienced or effective teachers, the required time for some observations, or the number of dimensions to be scored in some observations. Others are certifying additional administrators or groups of teacher-leaders who can share the burden for conducting formal observations.

Promote a **positive culture** for accurate observations and feedback.

School systems are providing teachers with meaningful opportunities to understand and improve on the effective practices embedded in the observation instrument and ensuring that principals know how to align evaluation with professional development in their

schools. Strategic communication is necessary and useful, but deep culture change happens when school systems provide tangible opportunities for teachers to learn from, and grow from, classroom observations and other measures of effective teaching.

3. Monitor and Ensure Quality

School systems are taking the extra step to audit the accuracy and reliability of observations to identify potential problems.

- **Analyze data from observations to flag patterns that suggest problems with accuracy.**

  Data systems can be expensive, but school systems that invest in them for warehousing and analyzing observation results cite the ability to keep frequent tabs on inter-rater reliability and possible problems with accuracy.

- **Audit evidence collected by observers to confirm that it aligns with the scores they assigned.**

  School systems that audit evidence from observations say that it also offers an opportunity to provide observers with feedback on how to improve their evidence-collection and evidence-analysis skills.

- **Conduct a reliability audit based on additional observations in a sample of classrooms.**

  The MET project has outlined an auditing procedure that a school system could use to monitor reliability by conducting additional observations in a sample of classrooms.[1]

One thing is clear: Ensuring accurate feedback from observations presents a complex challenge for school systems. Leading states, districts, and organizations have learned that simply providing initial training for observers is not a sufficient solution. Therefore, school systems must design a robust solution involving multiple strategies tailored to their own unique circumstances. But those that make such an investment are likely to reap significant dividends. Students benefit greatly when teachers are provided accurate, actionable feedback that helps them improve classroom practices.

Possible Strategies for Ensuring Accurate Observations

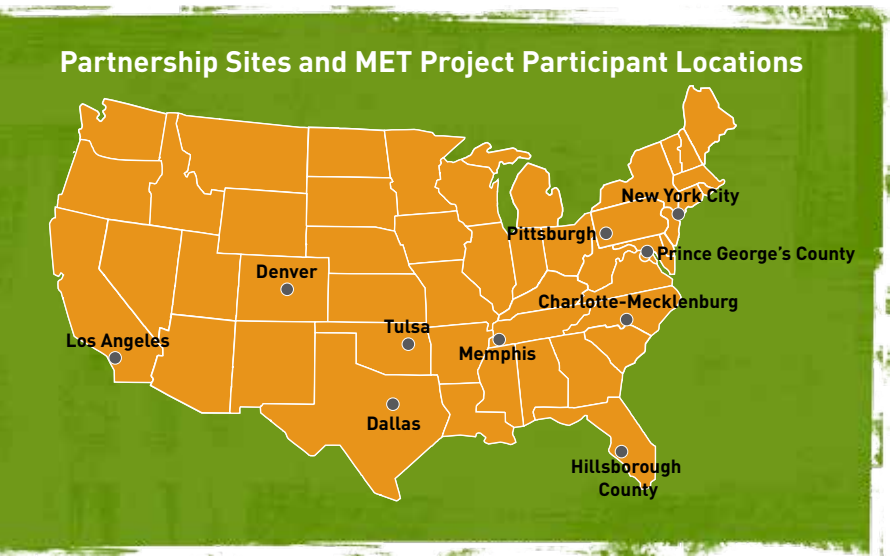
The figure to the right shows how school systems can select and deploy strategies that promote accuracy at key points before, during, and after the school year. In the sections that follow, we discuss the strategies in greater depth and illustrate them with specific examples and advice offered by leading school systems and organizations.

	Before School Year	During School Year	After School Year
1. Build Observers' Capacity			
Training	✔		
Certification	✔		
Tools	✔	✔	
Reinforcement		✔	✔
Recertification			✔
2. Create Conducive Conditions			
Culture change	✔	✔	
Caseload		✔	
3. Monitor & Ensure Quality			
Data analysis and evidence audits		✔	✔
Additional observers		✔	

Introduction

In 2009, The New Teacher Project (TNTP) released a report called *The Widget Effect* that exposed teacher evaluations as infrequent and overly subjective, yielding inflated performance ratings that are not taken seriously enough to inform basic personnel decisions. Just as alarmingly, even though the evaluations relied mostly on direct observations of teachers' practices, they provided almost no useful feedback to help teachers improve. Across the 12 districts TNTP studied, evaluations identified areas for improvement for only one of four teachers.[2]

While *The Widget Effect* garnered immense attention in the media, its findings came as no surprise to America's classroom educators. In a national survey conducted by Education Sector, 73 percent of teachers dismissed evaluations either as "just a formality" or as "well-intentioned but not particularly helpful to [my] teaching practice."[3]



Fueled by such exposés, as well as by philanthropic investments and the federal Race to the Top Program, states and school districts around the country are moving rapidly to overhaul teacher evaluation policies. They aim to produce much better information about teaching effectiveness, not only to enable more informed personnel decisions but also to provide greater support for improving teacher practice.

In 2009, the Bill & Melinda Gates Foundation launched two significant investments to support effective teaching. **The Measures of Effective Teaching (MET) project** is a research partnership involving nearly 3,000 teachers from across the country to investigate and build better ways to identify and develop effective teaching through the use of multiple measures. At the same time, the foundation made grants to a group of **Partnership Sites to Empower Effective Teaching,** school systems in which district leadership, board leadership, and union leadership collaborated to develop plans to implement new multiple measure systems of teacher evaluation; to strengthen supports for teachers; to recognize and reward effective teaching; and to ensure that the most underserved students have access to highly effective teaching.

Partnership sites are finding new ways to link teacher evaluation with teacher support, but doing so requires development of new kinds of systems, processes, and tools to empower administrators and teachers in their daily work. That represents both a significant opportunity and a major challenge for the field. This paper examines some of the practices partnership sites and others are using to strengthen classroom observations as one strategy for improving teaching practice and student learning.

Context

As partnership sites and other school systems design new evaluation systems, they are keeping classroom observations as a mainstay component. But they are designing new tools and protocols to ensure observations can provide much more objective and precise information about teachers' practices. Instead of the crude checklists principals relied on in the past, leading school systems are equipping administrators and other evaluators with sophisticated observation instruments, often referred to as "frameworks" or "rubrics." Some school systems are creating customized observation instruments while others are adopting or adapting commercially available ones.

The new instruments enable observers to identify teaching practices along multiple dimensions and to classify practices along a continuum of performance levels, with the highest level of performance painting a picture of what excellent practice should look like. For example, based on evidence collected during an observation, an administrator or peer evaluator might score "instructional dialogue" a two and "regard for student perspectives" a four. Some instruments rate dimensions of practice on a three-point scale while others use a seven-point scale. Figure 1 shows an excerpt from the **Denver Public Schools** *Framework for Effective Teaching Evidence Guide* related to the dimension "provides rigorous tasks and ensures student success through supports."

Figure 1. Excerpt from Denver Public Schools *Framework for Effective Teaching Evidence Guide (Pilot Year 2011–12)*

Domain: INSTRUCTION **Expectation:** STANDARDS-BASED GOALS

Indicator: I-3: Provides rigorous tasks and ensures student success through supports

Observable Evidence	Not Meeting (1-2)	Approaching (3-4)	Effective (5-6)	Distinguished (7)
Teacher Behaviors	• Tasks may seem like busy work as evidenced by students not needing to think through their work. Teacher does not incorporate rigorous tasks. • If teacher provides rigorous task(s), strategies are not used to support students with rigorous tasks (see examples in "Effective"), as seen by most (~75%) students struggling with tasks. • Tasks may be rigorous, but are overly scaffolded, so most (~75%) students are not required to think through work.	• Teacher incorporates tasks that may not be rigorous (i.e., do not require students to think at high levels). • Teacher may use strategies to support students with rigorous tasks (see examples in "Effective"), but some (25–50%) students may still struggle with tasks. • Tasks may be rigorous, but are so scaffolded, some (25-50%) students are not required to think through work.	• Teacher incorporates rigorous tasks that require students to use higher order thinking skills. • Supports from teacher to complete rigorous tasks may include: – Purposefully creating student groups to execute tasks. – Using gradual release: Model ("I do"), guide students through shared practice ("We do"), and provide independent practice ("You do"). – Using inquiry model: Allow students to explore initially, then regroup them to discuss experience or findings. – Using think-alouds to model approaches to tasks. • Sufficient but not too much support is in place for almost all (>75%) students' success while still requiring them to think through work.	In addition to "effective": • Rigorous tasks are aligned to student need so that regardless of support needed, all students are engaged in tasks that require higher order thinking skills. • Teacher supports all students with appropriate academic tools that promote their success with rigorous tasks.

(cont.)

(cont.)

Observable Evidence	Not Meeting (1-2)	Approaching (3-4)	Effective (5-6)	Distinguished (7)
Teacher Behaviors	Effective examples would include rigorous tasks where: • Types of thinking required are higher order (i.e., analyzing, evaluating, creating/synthesizing). • Tasks demonstrate usefulness and value of discipline (i.e., they illustrate application and relevance of discipline beyond classroom). • Degree of scaffolding or cue is appropriate, so students are required to think through work, but not struggle to a level of frustration. • Students are transferring higher-level thinking from speaking and thinking aloud to writing. • There is more than one way to approach tasks. • Instruction and tasks build and integrate learners' listening, reading, and writing skills as their oral language develops. • Activities are increasingly difficult (additional skills and/or effort required) during lesson or sequence of lessons. • Students have to understand complex texts, data sets, events, etc., using prior learning and inquiry skills. • Students have to draw inferences to generalize from new data and/or facts. • Superficial features of new challenges do not look familiar, and students need to demonstrate ability to apply skills or understanding in different contexts.			
Student Behaviors	• Some students work on tasks that may or may not be aligned to objective(s). • Most (~75%) students struggle to remain engaged in tasks because they lack support. • Most (~75%) students are observed not thinking through the work because tasks lack rigor or are overly scaffolded.	• Students are observed engaged in tasks, but may not be using high-level thinking skills as they complete tasks aligned to learning objective(s). • Students receive some support for rigorous tasks, but some (25–50%) students still struggle with tasks. • Some (25-50%) students are observed not thinking through the work because tasks are not rigorous enough or are too scaffolded.	• Students are observed using high-level thinking skills as they complete tasks aligned to learning objective(s). • Almost all (>75%) students receive support for rigorous tasks but are still required to think through the work.	• All students, regardless of support needed, are engaged in tasks that require higher order thinking skills. • Students are observed using appropriate academic tools that support their success with rigorous tasks.

Source: Denver Public Schools.

Better Feedback to Improve Teaching

New evaluation systems provide information from multiple sources that can be incorporated into several forms of useful feedback. One kind of feedback can rely on *quantitative information* about a teacher's overall skill set based on one or more measures of effective teaching that are captured over an extended timeframe. Data from student surveys can tell a 7th grade science teacher that her overall body of practice is more successful in "captivating students" than in "challenging students." Scores from multiple classroom observations might signal to a 4th grade teacher that his lessons tend to be quite strong on the dimension "managing student behavior" but could be greatly strengthened when it comes to "using questioning and discussion techniques."

Another kind of feedback takes the form of a *qualitative coaching conversation* that an observer holds with a teacher during a "post-conference" following an observation. The scoring of the lesson informs the post-conference, but the coaching conversation delves much deeper than those numerical data. Indeed, some school systems advise observers to wait until the end of the post-conference to share the formal scores with teachers.

The MET project has found that a single observation is not enough to provide a complete and reliable picture of the relative strengths and weaknesses in a teacher's *overall* body of practice. Lessons can differ greatly due to many factors, including the material being taught,

and teachers employ practices related to various dimensions of an observation instrument differently from lesson to lesson. Therefore, providing feedback on a teacher's overall skill set, as opposed to specific instances of practice during a lesson, might require observation and scoring of multiple lessons over the course of a school year.

Greater reliability is particularly important when making major decisions based on observations. For example, **Hillsborough County Public Schools** is offering teachers professional development courses on specific dimensions of its observation instrument, the *Framework for Teaching*. Because such courses demand a significant time investment, a teacher might want to consider scores from across multiple observations before selecting a course on "using questioning and discussion techniques" instead of one on "managing student behavior." Similarly, principals might want to consider scores from multiple observations before deciding to devote large chunks of a school's professional development time to a particular dimension on the observation instrument.

Providing Objective, Immediate Feedback

But that should not be interpreted to mean that accuracy does not matter for observers charged with providing useful feedback after *each* lesson they observe. Just as a college baseball coach needs to provide accurate feedback on a pitcher's performance following a practice or a game, an administrator or peer evaluator needs to provide accurate feedback to a teacher during a post-conference. A major league scout might need to analyze a range of statistics and observe more than one game to get a reliable read on a pitcher's overall skill set, but a coach's job is to provide immediate feedback to the pitcher during *every* game and practice.

To that end, many school systems are training observers to ask probing questions that prompt teachers to reflect deeply about the instructional choices they made during a lesson, often by strategically analyzing how specific practices affected students' behavior and learning. For example, evidence from a lesson might show that a teacher's "regard for student perspectives" had an observably positive impact on her students' engagement, but that lower "quality of feedback" limited some students' learning of the material. Such conversations typically result in identification of relatively strong practices that teachers plan to *extend* into future lessons and one or two weaker practices they plan to *improve* in future lessons.

Cognitive psychologists who study expert performance have found that such feedback is the key resource for engaging in the kind of "deliberate practice" necessary to reach high levels of performance in any field. While athletes and musicians often receive regular doses of high-quality feedback, most professionals do not. "The greatest obstacle for deliberate practice during work is the lack of immediate objective feedback," says K. Anders Ericsson, a leading authority on expertise and expert performance.[4]

Recent educational research suggests that objective feedback can be a powerful resource for improving teaching and learning in schools. A study by Eric Taylor and John Tyler found that when mid-career teachers participated in Cincinnati's Teacher Evaluation System, their students scored significantly better on state tests in following years. "One likely mechanism

Value of Providing Both Quantitative and Qualitative Feedback

One kind of feedback can rely on **quantitative information** about a teacher's overall skill set based on one or more measures of effective teaching that are captured over an extended timeframe. Another kind of feedback takes the form of a **qualitative coaching conversation** that an observer holds with a teacher during a "post-conference" following an observation. The scoring of the lesson informs the post-conference, but the coaching conversation delves much deeper than those numerical data.

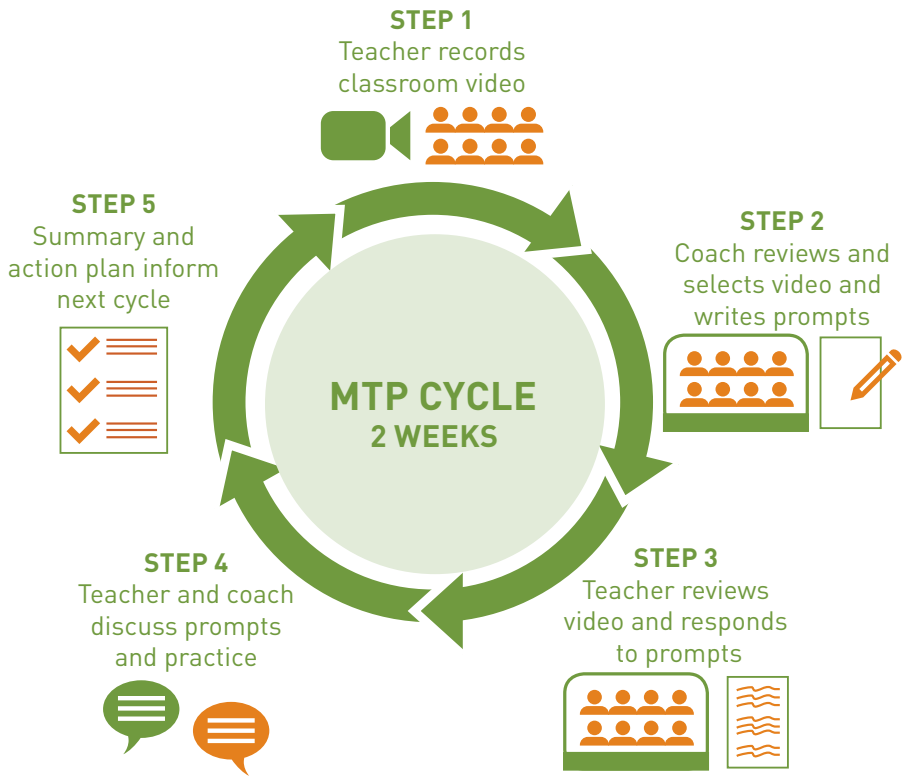
Both types of feedback are necessary to create both reliable and accurate observation systems. Teachers need to receive multiple observations to get a reliable picture of their overall strengths and weaknesses. But they also need to receive meaningful, high-quality feedback after each observation so that they have information to act on.

for such productivity growth is that the feedback provided in the evaluation spurs employee investments in human capital development," the researchers concluded.[5]

An experimental study of **MyTeachingPartner,** an online instructional coaching program that uses the Classroom Assessment Scoring System (CLASS), offers even stronger evidence. CLASS was developed by Robert C. Pianta and colleagues at the University of Virginia and is one of five observation instruments that the MET project recently tested for reliability and validity. In the experimental study of MyTeachingPartner, Pianta and his colleagues found that CLASS-based coaching had a substantial impact on secondary students' test scores the year after their teachers participated in the program. The impact was equivalent to moving from the 50th to the 59th percentile on Virginia's statewide assessments (on a 100-point scale).[6]

Although MyTeachingPartner does not contribute to teachers' formal evaluations, the program uses an observation-and-feedback cycle not dissimilar to the one at the heart of many new evaluation systems. Participating teachers share two video-recorded lessons per month with trained coaches called "consultants," who then select video segments that illustrate positive practices or areas for growth on dimensions of CLASS. The consultants pose questions to the teachers that prompt them to analyze the relationship between their behaviors and students' reactions, and they help teachers create a plan to enhance future instruction based on CLASS dimensions (see Figure 2).

Figure 2. The MyTeachingPartner Coaching Cycle: Analyzing Classroom Behavior through Videos and Feedback



Source: Pianta, R.C. (2011). *Teaching Children Well: New Evidence-Based Approaches to Teacher Professional Development and Training.* Washington: Center for American Progress. Used with permission.

But other studies suggest that *inaccurate* feedback, especially "soft feedback" that fails to pinpoint opportunities for growth, does little to improve teaching and learning. An experimental study of a rural, National Science Foundation-funded peer program found that too much coaching "consisted of explanations of what occurred during the classroom observations rather than meaningful analyses of how classroom instructions could be improved." The program had no impact on students' achievement.[7] An in-depth, three-year study of instructional coaching in a sample of Urban Systemic Initiative sites reached the following conclusion: "The crux of the one-on-one work appears to lie in the structure of teacher-leaders' feedback to classroom teachers, and here the overreliance on soft feedback can be crippling."[8]

Failure to provide accurate feedback following a lesson might entail significant opportunity costs both for the teacher and her students even when it identified areas for growth. For example, if weak practices are misclassified as strong and subsequently extended into future lessons, student engagement and learning could suffer. If strong practices are misclassified as weak, the teacher might work hard to improve in a particular area even though investing to improve in a different area would have yielded much bigger dividends. Therefore, school systems adopting new evaluation systems face a common challenge: How to ensure that feedback and coaching avoid major errors in classification and instead are based on a *reasonably* accurate judgment of the lesson.

Learning from the Field

This brief offers examples of strategies that some states, districts, charter management organizations (CMO)s, and other education organizations are using to ensure that post-conference feedback can help teachers build on strengths and identify areas for growth based on accurate observations. While the foundation could not collect information from every community and program working to improve teacher evaluations, the ones we examined represent a significant sample of leading edge reform efforts.

Information about the practices in partnership sites was gathered through a brief survey, follow-up discussions, and additional interviews. Appendix 1 on page 30 provides an overview of the basic policies and procedures that the partnership sites supported by the foundation have adopted for conducting formal observations.

The foundation supplemented its learning from partnership sites by looking at several additional school systems and organizations that offer unique perspectives on solving this problem of practice. They included the **American Federation of Teachers,** the **District of Columbia Public Schools,** the **National Institute for Excellence in Teaching,** the **Tennessee Department of Education,** and the **University of Virginia.** Appendix 2 on page 31 provides background information on the work that each is doing in this area.

School systems that assign staff members to conduct accurate classroom observations must take steps to ensure that those individuals can, and do, consistently meet that responsibility, as would any organization that asks an employee to tackle a challenging new assignment. A school system can take three broad kinds of actions to ensure that observers provide teachers with sound feedback based on accurate observations. The next three sections look at those areas in depth.

Three Areas for Action

1. Build Observers' Capacity
2. Create Conducive Conditions
3. Monitor & Ensure Quality

1. Build Observers' Capacity

School systems can make sure that observers are well equipped to conduct accurate observations by providing them with the knowledge, skills, and tools to do the job well. Most leading school systems told us they are using two or more of the strategies described below. They recognize the need to provide observers with intensive training but understand that training is not enough to ensure accurate observations.

Initial Training

Training for observers varies across the school systems and organizations that shared strategies with us. In every case, the training took place over at least several days and in some cases lasted as long as 80 hours. Most school systems are contracting with external consultants to design and deliver the training while building internal capacity for their own staff members to provide it in the future. For example, the **Tennessee Department of Education** contracted with the **National Institute for Excellence in Teaching** to train approximately 5,000 observers last summer but was able to train an additional 2,000 itself during the fall.

While specific training activities vary from place to place, the curriculum tends to cover a core set of topics and to follow a similar flow to build observers' understanding of the instrument and ability to use it. According to Angela Minnici, a former **American Federation of Teachers (AFT)** official who worked with 10 districts on common training, that is no coincidence. "No matter what instrument you use, there are common elements: how you collect objective evidence, how you align it to whatever instrument you are using, how you interpret it, and how you talk to a teacher about the results. Regardless of the observation instrument, you want to see that same set of skills in place."

Typically, trainers begin by introducing staff members to the instrument so they can become familiar with the dimensions of practice it includes and the vision for effective instruction at its core. Next, the trainers delve into each dimension of the rubric to understand the elements of practice it encompasses. For example, in the *TAP Rubric*, what is "questioning" and what is "academic feedback"? How are they different from one another and related to one another?

Trainers then guide staff members in exploring the *range* of performance the instrument describes for elements of practice under each dimension. This often requires an understanding of "performance descriptors" most instruments use as anchors to guide judgments. For example, at the "distinguished" level of "provides rigorous tasks and ensures student success through supports," **Denver Public Schools'** instrument says that teachers give all students "appropriate academic tools that promote their success with rigorous tasks" (see Figure 1). Many school systems and organizations show short video clips of lessons to illustrate what

a particular practice looks like at the high and low ends of the scale, allowing observers to begin to "calibrate" their own expectations for effective practice against the instrument's.

Trainers then show longer video segments to illustrate how to observe and score multiple dimensions of the instrument. At some point, they teach staff members how to collect objective evidence from the lesson, often involving some kind of note-taking or "scripting," since that is the cornerstone for making accurate judgments about practices. Observers learn about potential sources of bias and how to avoid them when collecting and scoring evidence. And they learn how to align evidence from the lesson with the descriptors in the instrument to classify practices and arrive at accurate scores.

Finally, trainers provide staff members with extensive opportunities to practice collecting evidence, aligning it to the instrument, and scoring, either through video clips or live observations or both. Staff members share their scores, justify them by citing evidence they collected, and work to reach a consensus judgment that brings their expectations into closer calibration over the course of the training. Many school systems use video segments that have been rated by expert scorers, either external consultants or internal staff members who serve on "norming committees."

School systems and organizations shared several common lessons they had learned about training observers. First, like any good professional development program, observer training is best when it allows for collective participation of staff members, provides frequent opportunities for interaction and discussion, and offers extensive opportunities to apply and practice new skills. For example, based on what they learned from the first round of training, leaders of the **Alliance College-Ready Public Schools,** a Los Angeles CMO, revised the curriculum to offer more opportunities for practice and discussion. "The dialogue resulted in much greater understanding of the rubric and the evidence required to support ratings," they told us.

For that reason, some partnership sites believe that smaller groups work best for observation training. One reason **Alliance College-Ready Public Schools** was able to build in more opportunities for dialogue was that it could train observers in smaller groups after the first large cohort had been trained. **Memphis City Schools** told us that, "as with students, adult learning seems to work best for us when we can have small groups with lots of opportunities to apply the rubric to tangible teaching scenarios. However, this is challenging at times in a district of our scale, so we have reached out strategically to various departments to give them tools that can facilitate this intensive, smaller-group approach."

Finally, a number of partnership sites have learned that observers often struggle to transfer their new skills to the field if they have only practiced using videos. Based on its pilot training and observations conducted last year, **Denver Public Schools** learned that "work in the field is where powerful learning occurs for our observers." Therefore, it has incorporated "live observation training" into the activities it is providing for principals, assistant principals, and peer observers over the course of this school year.

Strategies for Initial Training

- **Alliance College-Ready Public Schools** revised its curriculum to offer more opportunities for practice and discussion.
- **Memphis City Schools** found that training worked best with small groups and many opportunities to apply the training to real teaching scenarios. Often challenging for such a large district, they reached out to various departments to provide tools to facilitate intensive, smaller-group approaches.
- **Denver Public Schools** has incorporated "live observation training" into the activities it provides for principals, assistant principals, and peer observers over the course of this school year.

Certification Assessments

Every school system has some method to assess and certify observers' skills or is working to develop one. In every case, the assessment requires the staff member to observe and score classroom lessons using the observation instrument. However, specific methods and standards varied based on the different goals that school systems have for the assessment.

For example, while most school systems certify observers on a strictly pass-fail basis, others have adopted certification protocols that place observers into three or more performance categories. **Partnerships to Uplift Communities (PUC) Schools** designates school administrators as "not yet certified," "conditionally certified," "certified," or "certified with distinction." Similarly, **Prince George's County Public Schools** sends a letter to administrators telling them they have scored at one of three levels: "fully certified," "conditionally certified," or "certification not yet achieved." Both use the "conditional" category to identify administrators who may conduct formal observations only with additional supervision and who need extra support over the course of the year to become fully certified.

School systems typically require observers to score one or more lessons to some standard of accuracy to be certified. **Memphis City Schools** convened a 17-member Certification Committee comprising teachers, principals, and district administrators who watched video-recorded lessons to agree on "gold-standard" scores for each dimension of the observation instrument.To be certified, observers had to watch the same videos and score at least three of 11 dimensions exactly the same as the committee while not deviating by more than one point (on a five-point scale) on any of the dimensions.

Strategies for Certification Assessments

- **Partnerships to Uplift Communities Schools** designates school administrators as "not yet certified," "conditionally certified," "certified," or "certified with distinction." They use the "conditional" category to identify administrators who need supervision to conduct observations and who need extra support to become fully certified.
- **Prince George's County Public Schools** sends a letter to administrators telling them they have scored at one of three levels: "fully certified," "conditionally certified," or "certification not yet achieved." They use the "conditional" category in the same way as PUC Schools.
- **Memphis City Schools** created a committee made up of teachers, principals, and district administrators to define "gold-standard" scores for each dimension of the observation instrument. To be certified, observers had to watch the same videos that the committee had and score at least three of 11 dimensions the same as it did—not deviating by more than one point (on a five-point scale) on any of the dimensions.

A few partnership sites consider multiple measures to certify observers. **Prince George's County** contracts with the Danielson Group to assess the following three areas: whether the evidence the observer recorded is objective and free of bias; whether the observer aligned the evidence with appropriate scoring criteria in the instrument; and whether the ratings met standards for accuracy. Several of the CMOs participating in **The College-Ready Promise** take a similar approach.

In contrast to the trend toward highly standardized, video-based certification, **Hillsborough County's** "final exam" required observers to successfully complete two live observation cycles at a school site over a two-day period. Teachers volunteered to be observed, and expert training consultants shadowed the observers to certify that they completed both cycles—including preconferences, observations, and post-conferences—with fidelity and accuracy. The district is currently investigating whether it might be possible to conduct some part of the certification using an online, video-based system. However, because **Hillsborough's** observation instrument requires observers to question students in order to assess one dimension in the instrument, they anticipate that some live component will always be necessary.

The variation among school systems certification goals and policies begs an important question: What strategy works best? However, the work is too new to provide definitive answers,

and many school systems are still learning from their early efforts. For example, Jonathan Stewart, Implementation Lead for **PUC Schools,** says the "conditionally certified" category has provided useful flexibility that has been well received by administrators and teachers. "The obvious advantage is that it allows a site leader to engage in the observation process before they are fully certified, but with support," Stewart explains. "At the same time, that support from the certified partner will help the leader prepare for the next certification assessment." However, so far PUC leaders have found the "certification with distinction" category to offer few practical benefits, at least during this early stage of implementation.

Tools to Aid Observation

Leading school systems and programs have found that certain characteristics of observation instruments themselves can impede or promote observers' capacity to make accurate judgments. Accuracy can suffer if an instrument includes *too many* dimensions for an observer to consider during a single classroom observation, or if does not describe its vision of effective practice with *enough* detail and specificity. "On the one hand, you can't oversimplify, but on the other you can't overwhelm observers and teachers," Pianta says his team learned while they refined CLASS over a 10-year period. To help make the 70-page *CLASS Manual* more "digestible," they developed a set of complementary materials, including observation guides and *CLASS Dimension Guides*.

School systems are refining observation instruments to achieve a better balance between comprehensiveness and manageability based on what they learn from pilot testing and implementation. In **Pittsburgh Public Schools,** a sizeable team of teachers and administrators has led development of the new evaluation system. After **Pittsburgh** piloted the *RISE Rubric* during 2010–11, the team revised the language in the instrument to be cleaner and clearer and to eliminate confusing terminology. The team also incorporated new language in the form of "critical attributes" that describe what instructional practices look like in action at each level of proficiency. District leaders say that teachers and administrators have responded positively to the revisions.

Strategies to Aid Observation

- **Pittsburgh Public Schools** has revised its initial instrument language to be cleaner, clearer, and eliminate confusing terminology. Its team also incorporated new language that describes what instructional practices look like in action at each level of proficiency.
- **DC Public Schools** significantly streamlined its instrument after the first year of implementation, eliminating redundancies to reduce the number of dimensions from 13 to nine.
- **Partnerships to Uplift Communities Schools** purchased LiveScribe pens for every observer. The pens capture audio from the lesson that is time-synced with the written evidence the observer records during the lesson.

DC Public Schools significantly streamlined its instrument after the first year of implementation, eliminating redundancies to reduce the number of dimensions from 13 to nine. For example, district leaders found that scores on several different dimensions related to classroom management tracked very closely together, so they collapsed them into one.

At the same time, **DC Public Schools** built more flexibility into some indicators to enable observers to accurately capture evidence of effective practice in different kinds of lessons. Instead of requiring teachers to check for understanding a certain number of times in every lesson, the revised instrument looks at whether teachers check for understanding at appropriate and important points. Instead of asking whether the objective for the lesson has been posted in the room, observers assess whether students understand what they are learning and why. The revised instrument requires observers to exercise greater judgment, which can be a challenge for consistency of scoring, but it allows more meaningful measurement and feedback.

School systems also are providing additional tools to complement the observation instrument. For example, **PUC Schools** purchased LiveScribe pens for every observer. The pens capture audio from the lesson that is time-synced with the written evidence the observer records during the lesson. Leaders say the decision was prompted by teachers' concerns about observers' ability to capture all of the student dialogue necessary to score some dimensions of the observation instrument.

Last fall **PUC** also developed its nine-page *Evidence Guide* that describes a range of evidence observers might collect for each dimension of the observation instrument. The guide helps observers in several ways:

- **"Before observation,** to deepen observers' understanding of an indicator and guide them during collection if their evidence for a particular indicator is often sparse;
- **"While aligning evidence,** to help observers identify passages of their collected evidence that would be appropriate for a particular indicator;
- **"When seeking to improve rating accuracy,** to deepen understanding of an indicator and the full scope of evidence that pertains to it; or
- **"In conversation with teachers,** to deepen their understanding of how they might show proficiency on an indicator or to help answer why an observer included [or] omitted certain evidence when sorting for an indicator."[9]

Denver Public Schools has produced a similar evidence guide, along with a *Framework for Effective Teaching Handbook* that includes tangible examples of effective and ineffective practices. For example, for the dimension "provides rigorous tasks and ensures student success through supports," the handbook offers an example of effective practice from an elementary literacy lesson and an example of ineffective practice from a secondary geometry lesson.

Strategies for Reinforcement

- **Partnerships to Uplift Communities Schools** developed Individualized Calibration Assessment Plans describing observers' strengths, weaknesses, and "next steps" so implementation coaches could work with them to strengthen skills.
- **Prince George's County Public Schools** is offering 45-minute sessions on Thursdays that provide one-on-one coaching for observers who obtained "conditional" certification status and need to become fully certified. They are also using Polycom video-conferencing technology to allow large groups to view live lessons together and then discuss the evidence they collected and scores they assigned.
- **Pittsburgh Public Schools** convenes leadership team trainings that focus on a high-priority dimension of the observation instrument, and each "deep dive" is then replicated at the school level.

Opportunities for Reinforcement

Perhaps the most common lesson learned by sites that have piloted or implemented new evaluation systems over the past year is that they need to provide additional opportunities for observers to improve or maintain their skills following initial training. School systems often refer to such support as continuing "calibration."

Partnership sites gave two main reasons for investing in strategies to reinforce observers' skills throughout the school year. First, as described above, some districts and CMOs are using the certification process to identify observers who need extra support or to identify skill areas on which individual observers can improve. Last year, based on performance on the certification assessment, **PUC** developed Individualized Calibration Assessment Plans describing observers' strengths, weaknesses, and "next steps" so implementation coaches could work with them to strengthen skills. This year **Prince George's County** is offering 45-minute "Think It Through Thursday" sessions that provide one-on-one coaching for observers who obtained "conditional" certification status and need to become fully certified.

Second, sites are recognizing that many observers can experience "rater drift" over the course of the school year no matter how well they performed on the initial certification assessment. School districts and CMOs are using a wide variety of strategies to help observers maintain and improve their skills, including the following:

- **Deep-dive training** for groups of observers focused on specific dimensions of the observation instrument;
- **One-on-one coaching** provided by school system leaders or expert consultants;
- **Paired observations** of live or video-recorded lessons; and
- **Group calibration** sessions based on live or video-recorded lessons.

Some school systems have incorporated continuing calibration activities into the regular districtwide staff meetings that administrators attend each month, which can help save on costs while signaling that accurate observations are a central goal for the district. Others are providing additional opportunities for reinforcement. **Pittsburgh Public Schools** convenes leadership team trainings that focus on a high-priority dimension of the observation instrument, and each "deep dive" is then replicated at the school level. **Prince George's County** is using Polycom video-conferencing technology to enable large groups of up to 40 observers to view live lessons together and then discuss the evidence they collected and scores they assigned.

Although many school systems rely on expert consultants to deliver reinforcement and calibration opportunities, most are working to build internal capacity so their own staff members can take over such responsibilities in the future.

SUMMARY: Build Observers' Capacity

THREAT ASSESSMENT: Does the observation instrument support accurate observations? Is it too exhaustive, too vague and unclear, or needlessly redundant?

POSSIBLE STRATEGIES

- Determine if the instrument supports accurate observations by piloting it before implementation or by collecting feedback from observers during or following initial implementation.
- Revise the instrument to ensure it is not too extensive and does not include too many dimensions to be manageable.
- Analyze whether any dimensions are redundant, and if so, collapse or eliminate them.
- Revise the instrument to clarify language describing each dimension of practice.
- Revise the instrument to incorporate clearer and more specific descriptors of practice at various levels of performance and/or examples of practice at various levels of performance.
- Develop supporting materials that clarify the instrument's vision of effective practice or offer tangible examples of practices at different levels.

THREAT ASSESSMENT: Do observers have adequate understanding of the instrument, including a sound operational understanding of how practices described in each dimension might look "in real life"?

POSSIBLE STRATEGIES

- Use short video clips of lessons to illustrate what practices look like for various dimensions at various levels of performance.
- Provide examples of effective and ineffective practices under various dimensions of the instrument.

THREAT ASSESSMENT: Are observers equipped with sufficient skills and tools to capture appropriate evidence from an observation to arrive at accurate judgments?

POSSIBLE STRATEGIES

- Provide observers with training on how to collect evidence aligned with the instrument.
- Provide observers with an evidence guide that offers additional tips and examples.
- Provide observers with forms and other tools to aid evidence collection.

THREAT ASSESSMENT: Will observers have difficulty transferring the skills they acquired during initial training to conduct accurate live observations in the field?

POSSIBLE STRATEGIES

- Incorporate opportunities for live practice into training or provide a window for live practice at the beginning of the school year.

(cont.)

SUMMARY: Build Observers' Capacity *(cont.)*

THREAT ASSESSMENT: What if the training fails to provide some observers with sufficient knowledge and skills to conduct accurate observations?

POSSIBLE STRATEGIES

- Require observers to pass a performance assessment to certify that they can conduct accurate observations.
- Identify observers who might need extra oversight and support before they can observe and provide feedback to the required standard of accuracy.

THREAT ASSESSMENT: What if observers' skills erode during the school year following successful training and certification and their calibration begins to drift?

POSSIBLE STRATEGIES

- Provide observers with opportunities for reinforcement and "continuing calibration" throughout the year.
- Provide deep-dive training on specific dimensions of the instrument.
- Provide expert coaching.
- Provide paired observations.
- Provide group calibration events.
- Require observers to re-certify periodically by passing a version of the original performance assessment.

EARLY LESSONS

- School systems can check for redundancy by analyzing how closely observation scores correlate across dimensions.
- Feedback from observers and teachers is critical for identifying problems with vague and unclear language.
- Quantitative indicators that allow observers to check off or count teaching practices allow more consistent scoring, but they might also undermine meaningful measurement and feedback.
- If observation instruments require talking to students, some element of the certification assessment might have to take place in live classrooms.
- Districts can share costs by collaborating on common training and certification assessments, especially when they have similar observation instruments.
- Observers will drift despite rigorous training and certification, so school systems should plan to offer some kind of reinforcement throughout the year.
- In large districts, technology might provide efficient ways to conduct group calibration events.
- Incorporating reinforcement activities into regularly scheduled staff meetings can save on costs and signal that accurate observations are a high priority.
- If contracting with consultants, plan strategically to build internal staff capacity to deliver training, certification, and reinforcement in the future.

District Collaboration to Ensure Accurate Feedback from Observations

Ten school districts in **New York State** and **Rhode Island** have joined with the **American Federation of Teachers (AFT)** and their state-level AFT affiliates to form the **Educator Evaluation for Excellence in Teaching and Learning Consortium**. In 2010 the consortium obtained a $5 million i3 grant from the U.S. Department of Education to develop shared materials and procedures to train observers to conduct accurate classroom observations as one component of new evaluation systems the districts are adopting. During early 2011 the consortium conducted training for about 100 observers, including 60 who piloted the new observations in 26 schools across the 10 districts.

One key to the collaboration lies in similar definitions for effective teaching practices across the 10 districts. To participate in the consortium, each district agreed to adapt Charlotte Danielson's *Framework for Teaching*. As a result, the consortium was able to collaborate with expert consultants who co-developed shared materials and activities and then bring together observers from across the 10 districts for common training sessions. The majority of training focused on common elements, and observers broke into district-specific groups only for a few activities focused on particular elements of their own observation instruments.

However, Angela Minnici, former associate director for educational issues at the **AFT**, says a significant amount of collaboration still would have been possible even if some districts had adapted other instruments. "No matter what instrument you use, there are common elements: how you collect objective evidence, how you align it to whatever instrument you are using, how you interpret it, and how you talk to a teacher about the results. Regardless of the observation instrument, you want to see that same set of skills in place." In addition, she says, there are some common elements of teaching practice that most observation instruments identify as effective, including high levels of cognitive engagement.

Minnici says the biggest benefit is the obvious one: cost. During the start-up phase for new evaluation systems, districts often contract with consultants to help design and conduct training for observers. Sharing those costs can deliver considerable savings. Now the consortium is exploring other possible cost-efficiency strategies. For example, could districts in the same state share trained observers to maximize available personnel? Regulations regarding final evaluation judgments will need to be considered, but the consortium thinks the question is worth exploring.

The districts are discovering another important benefit: Working together on observation materials and training has created a natural platform for ongoing sharing and collaboration. For example, a district that develops a professional development unit will share the strategy with other districts in the consortium.

Minnici offers the following advice to districts contemplating similar collaborations:

- **Look for ways to build capacity from the start,** including identifying and developing district staff members who can play a larger role as the project progresses. "We identified a cadre of observers who showed promise and interest in deepening their skill sets," Minnici explains, "and we have continued to work with them to the point where they are co-facilitating trainings."
- **Identify what districts will need to implement accurate observations beyond initial training.** "You can't just develop materials and training as a 'one and done' exercise," explains Minnici. Part of the solution involves building capacity and expertise within each district prior to full implementation. Another involves identifying common challenges that will arise during implementation. For example, the consortium is exploring questions such as how often trained observers need reinforcement to stay "calibrated."
- **Look for ways to leverage technology.** In-person meetings are probably necessary for many initial activities, but convening staff members from across different states and districts can be very expensive. The consortium is considering how to leverage a blend of travel and technology to provide ongoing calibration activities for trained observers.
- **Recognize that districts can vary greatly in existing resources and human capacity** necessary to design, pilot, and implement sophisticated new evaluation systems. Different districts will need different kinds of technical assistance, and some will need much more than others. "You need to make sure that every district gets what it needs to be successful," says Minnici.

2. Create Conducive Conditions

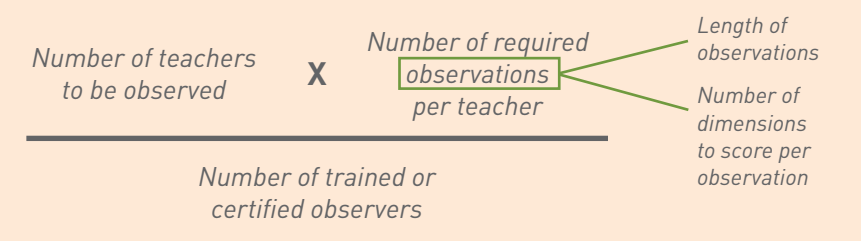
1. Build Observers' Capacity
2. **Create Conducive Conditions**
   - Ensure manageable caseloads
   - Promote a positive culture
3. Monitor & Ensure Quality

Even the most highly trained and rigorously certified staff members can fail to conduct accurate observations if they encounter significant obstacles in the field.

Create Manageable Caseloads

By requiring principals and other administrators to conduct multiple observations and post-conferences with teachers each year, new evaluation systems are asking them to spend more time on the instructional leadership aspect of their jobs. Many administrators view that as a positive development. At the same time, research suggests that administrators who are asked to conduct more observations than is reasonable given their other responsibilities often cut corners in ways that undermine the accuracy and usefulness of the feedback they give to teachers.[10] In essence, the problem can be expressed as a ratio that experts call the "span of review," shown in Figure 3.[11] The top of the ratio represents workload while the bottom represents capacity.

Figure 3. "Span of Review" Ratio for Analyzing Observation Caseloads



Source: Adapted from Shields, R.A. & Miles, K.H. (2008). *Strategic Designs: Lessons from Leading Edge Small Urban High Schools.* (Watertown, MA: Education Resource Strategies).

School systems face tough tradeoffs to keep the span of review manageable for administrators, and many struggle to find the right balance between their goals for new evaluations and concrete limits on human capacity. As the ratio suggests, one way to keep caseloads manageable is to reduce the workload by decreasing the number of teachers to be observed or the number of observations required per year, or both. **Hillsborough County Public Schools** adopted a "proportionate" observation policy that requires teachers to be observed from three to 11 times per year depending on their prior year's evaluation results.

Other school systems are adjusting the length of some observations or number of dimensions that must be scored. For its 2011–12 pilot year, **Denver** decided that during two of four observation windows, observers would score and provide feedback to teachers only on two "areas of focus" in the observation instrument, plus any areas relevant to English language acquisition.

Increasing the Number of Observers

Some school systems are addressing the bottom number in the ratio, boosting capacity by training more administrators to conduct observations or by training some groups of teachers to do so, or both. For example, to provide all teachers with the full complement of observations and feedback it had originally envisioned, **Partnerships to Uplift Communities (PUC) Schools** made a decision to certify two administrators in each school rather than only one as had been initially planned.

Pittsburgh Public Schools created a Peer Review Supported Growth Project in which about one-third of experienced teachers participate as an alternative to the main observation system in any given year. The project allows experienced teachers to spend a year focused on one dimension of the observation instrument that will significantly improve teaching and learning. According to the district, the program provides an opportunity for teachers to encourage each other to improve as well as creating more capacity for principals to spend time observing and meeting with beginning teachers and those who need extra support.

Hillsborough County trained and certified a cadre of teachers called "peer evaluators" to conduct formal observations of experienced teachers and "mentor evaluators" to observe beginning teachers. "If we were relying only on principals and assistant principals, we could never conduct as many observations as we're requiring," says David Steele, the district's chief information and technology officer. "The peer and mentor evaluators significantly increase our observation capacity." He points out three additional benefits: More observations can be conducted by observers who match the teacher's subject area or grade level; teachers can receive feedback from observers who have recent classroom experience; and the peer and mentor roles provide career opportunities and leadership pathways for effective teachers.

School systems that include teacher-leaders in the pool of trained and certified observers incur additional costs because those teachers must be released from regular duties. For example, **Hillsborough County's** peer evaluator initiative costs about $1,125 annually per evaluated teacher, which includes $5,000 stipends for peer evaluators, salaries and benefits for their classroom replacements, and $500 per peer evaluator for "continuing calibration" training. Mentor evaluators cost significantly more, about $4,320 per teacher evaluated, in part because beginning teachers are observed with greater frequency and in part because mentors devote a significant proportion of their time to coaching and mentoring in addition to conducting observations. Even so, the total cost of the program still amounts to less than 2 percent of the district's overall personnel budget for teachers.

The **District of Columbia Public Schools** has taken a somewhat different approach. Unlike **Hillsborough's** peer and mentor evaluators, who serve on a rotating basis and return to the classroom after two or three years, the DC master educator role is a permanent position. The district recruited some master educators from DC schools and others from outside the school system. Master educators earn starting salaries of $90,000 per year and spend about 75 to 80 percent of their time in activities related to observing and post-conferencing. The district estimates that the master educator program costs about $1,500 annually per evaluated teacher. Again, however, the total cost of the master educator program is relatively small, amounting to less than 2 percent of DC's personnel budget for teachers.

Strategies for Managing Caseloads

- **Pittsburgh Public Schools** created a Peer Review Supported Growth Project, which allows experienced teachers to spend a year honing their expertise on one area of the observation instrument.
- **Hillsborough County** adopted a "proportionate" observation policy that requires teachers to be observed from three to 11 times per year depending on their prior year's evaluation results.
- The **District of Columbia Public Schools** created permanent DC master educators, comprising experienced DC teachers and educators outside the system.

Of course, school systems that already plan to invest in "career ladders" for teachers could offset such costs by inviting some or all of those career-ladder teachers to conduct observations and provide feedback. Moreover, the study of **Cincinnati Public Schools'** evaluation system found that student achievement gains from the evaluations more than offset the "opportunity costs" of releasing effective teachers to conduct observations and post-conferences.

Change the Culture around Evaluations

In its *Widget Effect* study, TNTP found that years of useless feedback provided by principals had created "a culture in which teachers are strongly resistant to receiving an evaluation rating that suggests their practice needs improvement."The researchers described a vicious cycle in which "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" to accurate evaluations and feedback.[12]

Last year the **National Institute for Excellence in Teaching (NIET)** listed the need to directly and deliberately address that culture as one of the top lessons it had learned after a decade of **TAP: The System for Teacher and Student Advancement.** "Breaking that cycle requires much more than just new evaluation tools, procedures, and training," **NIET** had concluded.[13] Teachers and administrators in TAP system schools have found several concrete strategies to be helpful:

- Teachers use the observation instrument to score their own observed lessons, which allows them to calibrate with trained and certified observers and compare self-scores with observers' scores over time;
- During the first year of implementation, a number of regular weekly collaborative team meetings are devoted to helping teachers study the observation instrument and discuss its practices for effective instruction; and
- Master and mentor teachers model effective practices based on dimensions in the observation instrument during weekly instructional coaching sessions, allowing teachers to see the practices "live and in action" with their own students.

Some school systems are providing teachers with video-recorded modeling of effective practices on the observation instrument. For example, **Denver Public Schools** has gone into classrooms to video-record examples of effective teaching practices in each dimension of its framework. School systems also have found that providing teachers with meaningful opportunities to be involved in the design of new observations can lay a strong foundation for culture change. Many are involving teachers as full partners in selecting, adapting, or creating the new observation instruments.

According to Derrick Chau of **Alliance College-Ready Public Schools,** culture change also requires principals to understand how new evaluation systems are meant to support teacher development, not just better measurement. "If administrators themselves don't understand the philosophy of improvement behind standards-based evaluations, teachers will see it as a 'gotcha,'" Chau warns. He and other CMO leaders are working harder this year to help principals understand that rationale and to provide them with concrete strategies to link evaluation with professional development in their schools. Another suggestion they recently offered principals: Invite teachers to request observations and post-conferences beyond those formally required by the new evaluation system.

Strategies for Creating a Positive Culture for Accurate Observations

- **TAP System schools** have identified several strategies: ensure teachers **use** the observation instrument; make sure teachers become familiar with the instrument through regular, weekly meetings; and have master and mentor teachers model effective practices.
- **Alliance College-Ready Public Schools** are providing principals with strategies to link evaluations with professional development.
- **Denver Public Schools** are video-recording examples of effective teaching practices for each dimension of its framework.

Partnership sites also are working to align evaluation and professional development at the district level. In **Pittsburgh Public Schools,** teacher-leaders from two schools that opened in 2011–12 as Teaching Institutes have invited administrators and fellow teachers into their classrooms to provide support in accurately observing effective teaching across different content areas. Last year **Hillsborough County's** Office of Staff Development offered teachers online courses about its observation instrument, the *Framework for Teaching,* and this year it is aligning its online and in-person course offerings with specific dimensions of the framework. Teachers can earn in-service credits by completing the courses.

The "Four Cs" of Culture Change

Based on the examples above, one way to think about culture change might be through the lens of the "four Cs":

- **Communication:** Helping teachers and administrators understand the observation system and how it can support improvement;
- **Collaboration:** Inviting teachers to help develop or select the observation instrument and to establish gold standard scores for observers' certification assessments;
- **Calibration:** Providing teachers with opportunities to reach a deeper understanding of the observation instrument so they can begin to calibrate their own vision for effective practice against it; and
- **Coaching and Professional Development:** Providing teachers with meaningful opportunities to improve on the practices measured by the observation instrument.

Perhaps the biggest lesson learned so far is that changing the culture to support accurate observation and feedback requires more than just strategically communicating with teachers. Such efforts are necessary and useful, but deep culture change happens when school systems provide tangible opportunities for teachers to learn from, and grow from, classroom observations and other measures of effective teaching.

SUMMARY: Create Conducive Conditions

THREAT ASSESSMENT: Will observers have sufficient time to complete the required number of observations, or will they be forced to cut corners?

POSSIBLE STRATEGIES

- Reduce the workload by varying the number of observations required for different types of teachers (experienced vs. inexperienced, more effective vs. less effective).
- Reduce the workload by requiring less time or fewer dimensions scored for some observations.
- Boost capacity by training and certifying more administrators to conduct observations.
- Boost capacity by training and certifying some groups of teacher-leaders to conduct observations.

THREAT ASSESSMENT: What if years of inflated and meaningless evaluation ratings have created a culture resistant to accurate observations and feedback?

POSSIBLE STRATEGIES

- Provide opportunities for teachers to understand and discuss the new observation instrument and the vision for effective practice embedded in it.
- Provide opportunities for teachers to calibrate their current understanding of effective practice against the observation instrument.
- Give teachers opportunities to see effective practices in the instrument modeled in real classrooms, either through videos or live observations.
- Ensure that principals understand how new evaluation systems can help all teachers improve, and provide concrete strategies for them to align evaluation and professional development.
- Provide meaningful opportunities for teachers to improve on the practices measured by the observation instrument (in addition to feedback during post-conferences).

EARLY LESSONS

- There is a limit to how many observations principals can conduct accurately given their other responsibilities.
- Training and certifying teacher-leaders to share the observation burden can be expensive, but:
  - It can allow greater subject-area and grade-level matching and give teachers opportunities to receive feedback from someone with more recent experience in the classroom;
  - Research suggests it can result in significant improvements in teaching and learning; and
  - Districts already implementing career-ladder reforms can offset costs by asking some or all of those teachers to conduct observations.
- School systems should ensure that principals understand the developmental philosophy behind the new observations and that they have concrete strategies for aligning evaluation with professional development in their schools.
- Communicating strategically with teachers about the new evaluation system is important, but real culture change happens when teachers are given meaningful opportunities to understand and improve on practices in the observation instrument.

3. Monitor and Ensure Accuracy

Even after training and certifying observers, school systems are discovering they need to take extra steps to monitor accuracy as staff members conduct observations over the course of the school year. School systems that want to keep frequent tabs on accuracy probably will need to invest in a data system for warehousing and analyzing observation results. Other methods include auditing evidence-score alignment from a sample of observations or asking observers to conduct extra observations in a sample of classrooms.

Data Analysis

Hillsborough County Public Schools invested in a Lawson Talent Management System to warehouse and analyze results from its new evaluation system along with other kinds of personnel data. District leaders now run weekly reports to analyze patterns that might signal problems with accuracy or reliability. When an observer's scores show an unusual pattern, officials investigate to determine whether some kind of intervention and additional support will be necessary to help the observer "re-calibrate" scoring. While such data systems can be expensive, **Hillsborough's** leaders believe the investment has enabled them to monitor accuracy with much greater frequency so they can intervene immediately when re-calibration becomes necessary.

Tennessee invested in a similar electronic data system to support its new evaluations. The system allows the state department of education to monitor observation results to identify districts or schools where the range and distribution of scores is significantly different than expected. Because **Tennessee** adopted the *TAP Rubric* as its statewide observation instrument, the department was able to establish baseline benchmarks by analyzing the distribution of scores for thousands of teachers in **TAP System schools.** The department also can compare observation scores with Tennessee Value-Added Assessment System scores. Districts will be able to access the data system to monitor scores as well.

When a district or school exhibits an unusual distribution of observation scores, the department reaches out to local education officials to set up a conversation about what might be causing the pattern. If there is a problem with accuracy, the department can provide a consultant who can help with re-calibration by conducting additional training or paired observations. The department has contracted with nine consultants, one for each region of the state, to help communicate with

Strategies for Data Analysis

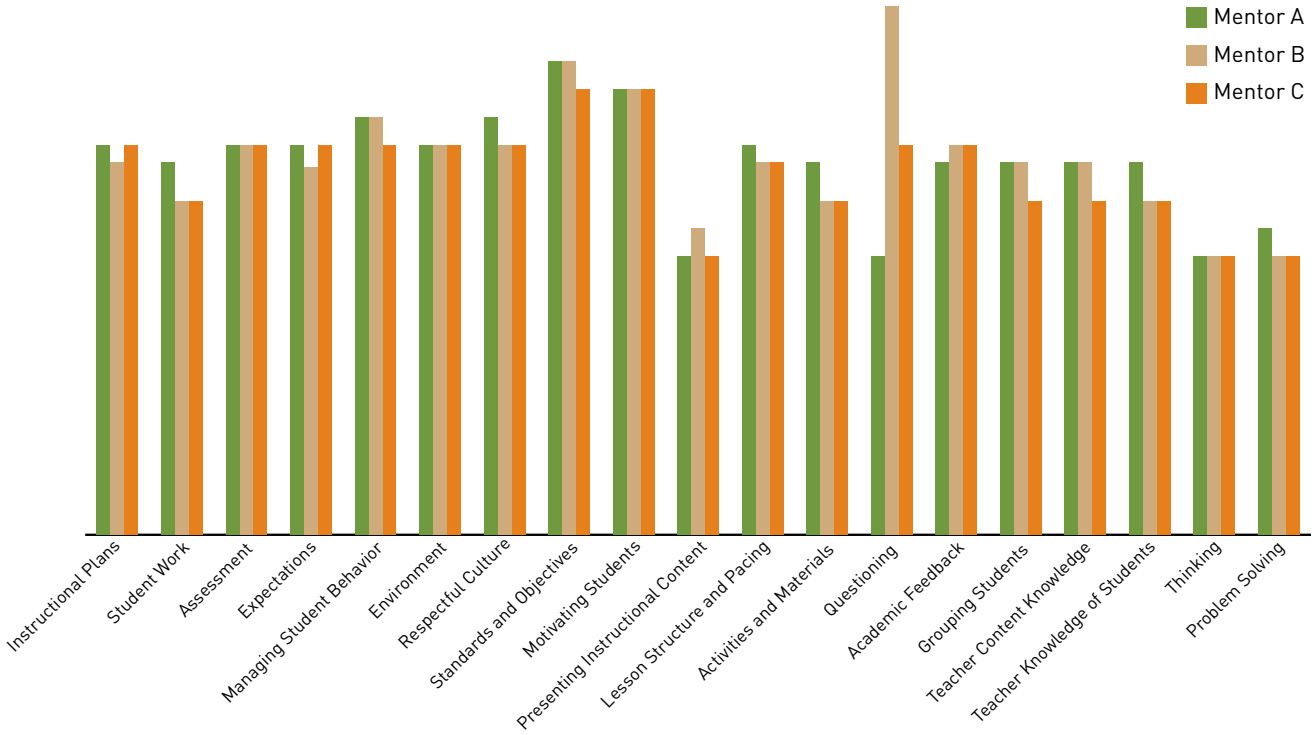
- **Hillsborough County Public Schools** invested in a Lawson Talent Management System to warehouse and analyze results from its new evaluation system along with other kinds of personnel data. District leaders now run weekly reports to analyze patterns that might signal problems with accuracy or reliability.
- **Tennessee** invested in a similar electronic data system that allows the state department of education to monitor observation results to identify districts or schools where the range and distribution of scores are significantly different than expected
- The **TAP System schools** are using a data system that can provide dozens of tables and charts to help monitor inter-rater reliability of observation scores. School-based evaluation teams made up of principals, master teachers, and mentor teachers are required to analyze data reports each month to flag, investigate, and if necessary, remediate problems with rater drift or reliability.

districts and provide technical support for implementation. The state board of education may decide that districts unable to resolve problems with accuracy should be subject to additional oversight or ineligible to propose alternative evaluation models.

Emily Barton, **Tennessee's** assistant commissioner for curriculum and instruction, explains that the benchmarks are not meant to be quotas. "We know there might be legitimate reasons for anomalous patterns, which is why we first hold conversations with districts," she says. "But we also want to guard against upward drift and unintended 're-norming' so that, for example, a score of three still means what it is supposed to mean based on the rubric."

National Institute for Excellence in Teaching (NIET) developed a data system called CODE that can provide dozens of tables and charts to help monitor inter-rater reliability of observation scores in and among **TAP System schools.** School-based evaluation teams comprising principals, master teachers, and mentor teachers are required to analyze CODE reports each month to flag, investigate, and if necessary, remediate problems with rater drift or reliability. Figure 4 provides an example of one kind of chart the CODE system can generate. **NIET** provides schools with an online library of video-recorded lessons that have been rated by gold-standard scorers, and evaluation teams use those videos along with live observations to re-calibrate scoring.

Figure 4. Example of a TAP System Chart for Analyzing Inter-Rater Reliability



Source: Jerald, C.D. & Van Hook, K. (2011). *More than Measurement: The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems.* (Santa Monica, CA: National Institute for Excellence in Teaching.)

Evidence Audits

Some school systems conduct "evidence audits" to confirm that evidence recorded by observers during a lesson matches the scores they assigned based on descriptors in the observation instrument. In addition to offering an accuracy check, the audits can also provide a powerful opportunity to strengthen observers' understanding and skills. For example, Jonathan Stewart and his team at **Partnerships to Uplift Communities** use the comments function in Microsoft Word to provide advice to observers inside the evidence records they audit. "Avoid summary statements, especially ones that include judgment," Stewart recently wrote in one such comment. "Include specific instances of the teacher telling students what she wants them to do." In another he advised that, "Evidence of the teacher choosing a story students were already familiar with would have shown stronger alignment with this indicator."

Reliability Audits

The MET project has outlined an auditing procedure that a school system could use to monitor reliability by conducting additional observations in a sample of classrooms.[14] School system officials can randomly select a representative sample of teachers (100 would suffice even for large districts) and then dispatch certified observers to conduct an additional "auditing observation" of those teachers on a day when they are not being formally observed. The auditing observers need not be expert consultants as long as they have been trained and certified to conduct observations and have no prior relationship with teachers they are assigned to observe in the sample.

SUMMARY: Monitor and Ensure Quality

THREAT ASSESSMENT: **Despite training, certification, and reinforcement, staff members might fail to consistently conduct accurate observations and post-conferences throughout the school year.**

POSSIBLE STRATEGIES

- Invest in a data warehouse that allows frequent analysis of observation scores to identify anomalous patterns.
- Audit evidence from a sample of observations to check alignment with descriptors in the observation instrument.
- Assign a group of observers to conduct additional observations in a representative sample of classrooms.
- Develop efficient strategies to immediately intervene and support observers when problems with accuracy are identified.

EARLY LESSONS

- Data systems can be expensive, but they enable more frequent monitoring of accuracy.
- Less expensive alternatives include auditing samples of evidence or conducting additional observations in a random sample of classrooms.
- School systems that require rigorous training, certification, re-certification, and frequent reinforcement might decide to monitor observation results on an annual or quarterly basis rather than every week or every month, while those that do not reinforce observation skills or require re-certification might need to be more vigilant about monitoring scores and auditing evidence to identify potential problems.

Designing a Coherent Solution

Ensuring accurate feedback from observations presents a complex challenge for school systems. Little formal research has been conducted on the issue as a practical problem of practice, and there is not yet convincing evidence to suggest that one combination of strategies works better than any others. Indeed, different strategies might work better for some school systems than others depending on their local contexts and specific observation policies.

Therefore, each school system should design a coherent solution that fits its particular needs, selecting a suitable set of strategies based on a logical theory of action about how those strategies can, taken together, ensure accurate observations. Moreover, the choice about whether to adopt any particular strategy can depend on the *other* strategies the school system selects. For example, depending on how heavily it invests in building and maintaining observers' capacity, a site might not need to monitor accuracy as frequently. A school system that requires rigorous training, certification, re-certification, and frequent reinforcement might decide to analyze observation results on a quarterly basis rather every week or every month. But a school system that decides not to require re-certification or to invest in frequent reinforcement might want to be more vigilant about monitoring scores and auditing evidence to identify potential problems.

Finally, states that adopt statewide evaluation systems might need to consider the problem differently than do local school districts. **Tennessee** recognized accuracy as a major challenge from the start. "The largest challenge I see is trying to ensure consistency in the range of distribution for the observation scores," Commissioner of Education Kevin Huffman testified before a U.S. congressional committee in July. However, state leaders also realized that they could not provide all of the reinforcement activities and ongoing "calibration events" from a state level that a district might at a local level. Therefore, they identified their two most feasible leverage points to be training and certification combined with data analysis and quality assurance.

One thing is clear: Leading states, districts, and organizations have learned that initial training alone is not sufficient to ensure accurate feedback from observations. But the investment in additional strategies is likely to pay valuable dividends. Recent research has shown that students benefit greatly when teachers are provided accurate, actionable feedback that helps them improve classroom practices.

Appendices

Appendix 1. Overview of Basic Policies for Conducting Formal Observations of Experienced Teachers among Selected Partnership Sites (as of Fall 2011)

	Teachers Included	Number and Duration of Observations (Minimum)	Observers	Tools	Additional Information
Denver Public Schools	All teachers (same process for experienced and new teachers)	4 observations per year for 45 minutes each	Principals/Assistant Principals (2); Peer Observers (2) ["Peer Observer" is a district-level position, enabling content-area and grade-level matching]	DPS *Framework for Effective Teaching*; note-taking/scripting templates; rating summary sheet; reflective feedback conversation template	All observations unannounced; observers score all dimensions during 1st & 4th observations and 2 "focus areas" (plus areas relevant to English Language Acquisition) during 2nd & 3rd
Hillsborough County Public Schools	Teachers not assigned a Mentor (generally those with 3+ years experience)	3-11 observations per year (depending on prior-year evaluation) for either a full period or 20-25 minutes each	Principals, Assistant Principals, Peer Evaluators, Subject Area Supervisors	Danielson's *Framework for Teaching*	Combination of scheduled (full-period) and unannounced (20–25-minute) observations
Memphis City Schools	Teachers who hold a Professional License	4 observations per year for at least 15 minutes each	District-Level and School-Level Administrators (Principals, Assistant Principals, Instructional Facilitators)	Memphis City Schools *Teacher Effectiveness Measure (TEM) Teaching and Learning Framework Rubric*	Based on feedback from teachers, will differentiate observation rubrics for teachers of certain content areas and student populations
Prince George's County Public Schools	Tenured, on-cycle teachers with no performance issues	2 observations per year for at least 30 minutes each	Principal and Assistant Principal	Danielson's *Framework for Teaching* observation evidence form; PGCPS *Standards for Excellence* form; "Look Fors"	Observations scheduled, preceded by pre-conferences; also requires 2 scheduled formative observations
Pittsburgh Public Schools	Tenured teachers not participating in the Supported Growth Peer Review Cycle or Intensive Support processes (approx. 2/3 of all experienced teachers)	2 "formal" observations per year for at least 30 minutes each, plus 2 "informal" observations per year	School Administrator (however, beginning in 2012-13, teachers in new career ladder positions also will contribute)	Pittsburgh Public Schools *Research-based Inclusive System of Evaluation (RISE) Rubric*	Some observations scheduled (with pre-conference) and some unannounced; 12 of *Rubric's* 24 dimensions are Power Components used for summative evaluation
TCRP-Alliance College-Ready Public Schools	All teachers (same process for experienced and beginning teachers)	2 observations per year for approx. 50 minutes each	School Administrator (Principal, Assistant Principal, Director of Instruction)	*The College-Ready Promise (TCRP) Framework for Effective Teaching;* online observation evidence collection form	One observation per semester
TCRP-Aspire Public Schools	Teachers with 3+ years of experience (post-induction)	2 observations per year for at least 30 minutes each	Any certified Principal, Dean, Assistant Principal, Area Superintendent, or Central Staff Member	*The College-Ready Promise (TCRP) Framework for Effective Teaching;* online observation data collection tool called Formative Learning	For *TCRP Framework* to be manageable, teachers focus on 2-3 dimensions per year but also receive evaluation on all dimensions
TCRP-Green Dot Public Schools	All teachers (same process for experienced and beginning teachers)	2 observations per year for at least 45 minutes each	Principal or Assistant Principal	*The College-Ready Promise (TCRP) Framework for Effective Teaching*	One formal observation per semester, each of which is preceded by 2 "informal" observations
TCRP-Partnerships to Uplift Communities (PUC) Schools	All teachers until they are identified as highly effective, after which the process is adjusted (same process for experienced and beginning teachers)	2 "formal" observations per year for a full period each, plus 2-4 informal observations totaling at least 20 minutes each semester	Assigned PUC Site Leader	*The College-Ready Promise (TCRP) Framework for Effective Teaching;* online observation data collection tool called Formative Learning; *Observation Guide*	Formal observations together account for 30% of total evaluation and informal observations for 10%
Tulsa Public Schools	Tenured teachers	2 observations per year for at least 20-30 minutes each	Principal or Assistant Principal	Tulsa Public Schools *Teacher and Leader Effectiveness (TLE) Rubric; TLE Observation Form*	Two observation windows: Beginning of year to Nov. 14 & 10 days from first observation to Jan. 14

NOTES: (1) Some partnership sites have alternative procedures for observing certain subgroups of experienced teachers (in one case affecting about one-third of experienced teachers), which are not captured in this table. (2) In many cases, sites have different policies for observing beginning teachers. (3) Some sites are considering changes to the policies listed in the table.

Appendix 2. Background on Additional School Systems and Organizations that Shared Information about Ensuring Accurate Observations

American Federation of Teachers

In partnership with two state-level affiliates, the American Federation of Teachers (AFT) is providing guidance and technical assistance to 10 school districts in New York State and Rhode Island that have formed the Educator Evaluation for Excellence in Teaching and Learning (E3TL) Consortium. The Consortium has developed common materials and procedures to train observers to conduct accurate classroom observations as one component of the new teacher evaluation systems the districts are adopting. Each district's observation instrument was originally based on Charlotte Danielson's *Framework for Teaching*, and expert consultants are helping the Consortium develop shared materials and training based on the *Framework*. During early 2011 the Consortium conducted training for about 100 observers, including 60 who piloted the new approach in 26 schools across the 10 districts.

District of Columbia Public Schools

During 2009–10, the District of Columbia Public Schools (DCPS) implemented a new evaluation system called IMPACT that evaluates teachers based on multiple measures including classroom observations. Under IMPACT, teachers are observed five times per year, three times by a school-based administrator and twice by a district-based master educator. Beginning this year, teachers who earned a rating of "highly effective" for two years in a row and scored 3.5 or higher on two observations this fall can choose to waive the remaining three observations. Administrators and master educators use an observation instrument called the *Teaching and Learning Framework*, which was developed by DCPS in 2008–09 and significantly revised in 2010. After each observation, the administrator or master educator holds a post-conference with the teacher to share a written report of the observation, including scores on each dimension of the *Framework*, and to discuss strategies for improvement.

National Institute for Excellence in Teaching

The National Institute for Excellence in Teaching (NIET) manages the TAP System, a comprehensive strategy for improving teaching effectiveness through aligned evaluation, professional development, compensation, and career advancement. In schools using the TAP System, teachers are evaluated based on multiple measures including four to six classroom observations per year. The observations are conducted by school-based administrators, master teachers, and mentor teachers who have been trained and certified by NIET to use an observation instrument called the *TAP Rubric*. During post-conferences, observers prompt teachers to analyze the lesson and discuss *Rubric*-based practices for "reinforcement" and for "refinement." During the summer of 2011 NIET trained and certified approximately 5,000 Tennessee administrators and educators to conduct observations and post-conferences based on the *TAP Rubric* as one component of the state's new teacher evaluation system.

Tennessee Department of Education

This year all school districts began implementing a new statewide evaluation system called the Tennessee Educator Acceleration Model (TEAM), which evaluates teachers on multiple measures of effectiveness including classroom observations. Under TEAM, teachers who hold a professional license are observed four times per year, and those who hold an apprentice license are observed six times per year. During 2011, more than 7,000 Tennessee administrators and educators completed training and certification to conduct formal observations and post-conferences using the *TAP Rubric*, which the state adopted as its formal observation instrument. The state board of education grants permission to some districts to use alternative evaluation models.

University of Virginia Center for Advanced Study of Teaching and Learning; Teachstone

The University of Virginia Center for Advanced Study of Teaching and Learning was the lead developer of the Classroom Assessment Scoring System (CLASS), a standardized metric for observing teacher practices and student behaviors. CLASS measures the quality of teacher-child interactions on a 7-point scale to rate the quality of emotional support and instructional support that students receive in classrooms. Over 10,000 individuals in the United States and several other countries have completed training on CLASS. As of October 2011, Teachstone had formally certified over 6,900 individuals to observe and rate classrooms using the CLASS, and it had certified over 500 individuals to provide formal CLASS training themselves.

Endnotes

1 See pages 39-40 of Kane, T. & Staiger, D. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* (Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

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 Duffett, A., Farkas, S., Rotherham, A.J., & Silva, E. (2008). *Waiting to Be Won Over: Teachers Speak on the Profession, Unions, and Reform.* Washington, DC: Education Sector.

4 Ericsson, K.A. (2009). "Enhancing the Development of Professional Performance: Implications from the Study of Deliberate Practice." In K. Anders Ericsson, ed., *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments.* Cambridge, U.K.: Cambridge University Press.

5 Taylor, E.S. & Tyler, J.H. (2011). "The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data on Mid-Career Teachers." Cambridge, MA: National Bureau of Economic Research.

6 Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011). "An Interaction-Based Approach to Enhancing Secondary School Instruction and Student Achievement," *Science* 333(6045): 1034-1037.

7 Murray, S., Ma, X., & Mazure, J. (2009). "Effects of Peer Coaching on Teachers' Collaborative Interactions and Students' Mathematics Achievement," *Journal of Educational Research* 102(3): 203-212.

8 Lord, B., Cress, K., & Miller, B. (2008). "Teacher Leadership in Support of Large-Scale Mathematics and Science Education Reform." In Melinda M. Mangin & Sara Ray Stoelinga, eds., *Effective Teacher Leadership: Using Research to Inform and Reform.* New York: Teachers College Press.

9 PUC Schools, *Evidence Guide.*

10 Milanowski, A.T. & Heneman, H.G. (2001). "Assessment of Teacher Reactions to a Standards-Based Teacher Evaluation System: A Pilot Study," *Journal of Personnel Evaluation in Education* 15(3): 193-212. Kimball, S.M. (2002). "Analysis of Feedback, Enabling Conditions, and Fairness Perceptions of Teachers in Three School Districts with New Standards-Based Evaluation Systems," *Journal of Personnel Evaluation in Education* 16(4): 241-268. 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 & Cecil G. Miskel, eds., *Educational Administration, Policy and Reform: Research and Measurement.* Greenwich, CT: Information Age Publishing.

11 Shields, R.A. & Miles, K.H. (2008). *Strategic Designs: Lessons from Leading Edge Small Urban High Schools.* Watertown, MA: Education Resource Strategies.

12 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.

13 Jerald, C.D. & Van Hook, K. (2011). *More than Measurement: The TAP System's Lessons Learned for Designing Better Teacher Evaluation Systems.* Santa Monica, CA: National Institute for Excellence in Teaching.

14 See pages 39–40 of Kane, T. & Staiger, D. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* (Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

BILL & MELINDA GATES *foundation*

www.gatesfoundation.org