MEASUREMENT MAKEOVER

Florida district revamps teacher evaluation to focus on student achievement

By Valerie von Frank

Hillsborough County (Fla.) Public Schools had a concern. A review in 2008 of teacher evaluations in the district found that more than 99% of the 12,000 teachers were rated satisfactory or outstanding, and nearly half of high school teachers received perfect scores.

While Hillsborough is a high-performing district and has several high schools on *Newsweek's* 2010 list of the nation's best, many in the district agreed the evaluations must be misleading — but there was no way to know

exactly how the ratings matched up to student learning. And teachers and administrators alike didn't know exactly how to describe what the evaluations were supposed to do: What does exceptional teaching look like?

Teachers were laboring under an unequal system of evaluation from school to school. Principals were asked to observe nontenured teachers once a year and tick off boxes on a 44-item checklist. Tenured teachers were observed once every three years.

"Truthfully," said David Steele, Hillsborough's chief information and technology officer, "many of our principals were not even reaching that goal. We were not in classrooms observing frequently enough."

The evaluations seemed to have little real connection

to teachers' daily work. Teachers also were frustrated that the effort yielded few specifics about what they could do to improve. So, in a system dedicated to continuous improvement, the administration and union jointly began to work to answer the critical questions to define teaching practices that lead to student learning.

Hillsborough is not alone. The question of what constitutes effective teaching is at the core of efforts around the nation to raise student achievement by focusing on teacher quality. In 2010 and 2011 legislative sessions, states including Colorado, Florida, Illinois, Indiana, and Tennessee passed legislation to mandate new systems of teacher evaluation based at least partly on student achievement, to lower barriers to dismissing underperforming teachers, and to change state policies that base layoffs on seniority.

They may have been goaded to some extent by a 2009 study of 12 districts over four states that concluded, "A teacher's effectiveness — the most important factor for schools in improving student achievement — is not measured, recorded, or used to inform decision-making in any meaningful way" (Weisberg, Sexton, Mulhern, & Keeling, 2009, p. 1). The report goes on to state, "In general, our schools are indifferent to instructional effectiveness — except when it comes time to remove a teacher" (p. 2).

Many also saw as an incentive grants from the Bill & Melinda Gates Foundation that put funding behind change efforts. Hillsborough has been involved in two grant projects, yielding $100 million over seven years, about 1% of the large district's annual operating budget, according to Steele.

"The Gates grant gives us money to do things we couldn't afford to do in the past," Steele said.

MEASURES OF EFFECTIVE TEACHING

Hillsborough began its work as part of the Gates Foundation's Measures of Effective Teaching Project, a $45-million effort with 3,000 teachers in seven districts over two years to develop objective and reliable measures of effective teaching. Researchers collected data from student feedback through surveys, student work, supplemental student assessments, assessments of teachers' ability to recognize and diagnose student problems, and teacher surveys on working conditions. In addition, a pivotal element of the research is more than 13,000 videotaped classroom lessons captured by 360-degree cameras, and teachers' subsequent reflections on their videotaped lessons.

Initial findings from that project across the districts indicated that "in every grade and subject, a teacher's past track record of value-added is among the strongest predictors of their students' achievement gains in other classes and academic years" (Measures of Effective Teach-

ing Project, 2010, p. 9). "The teachers who lead students to achievement gains in one year or in one class tend to do so in other years and other classes," a Measures of Effective Teaching Project report states. More conclusions based on two years of data from the project are expected to be released in January 2012.

Other recent studies underscore the complementary nature of student learning data and teacher observations in evaluations. A report of the National Board for Professional Teaching Standards (2011) states, "There will always be challenges in determining how much each teacher contributes to student learning. Education is a complex process … . For this reason, thoughtful evaluations of teacher performance must combine direct evidence of student learning such as 'value-added' data and examinations of teaching practice." A study analyzing student data from New York City between 2003 and 2008 found a correlation between teachers who did well on value-added measures and those who scored highly in observations, concluding that observations pick up on teacher skills not captured in student test scores — and that evaluation systems should incorporate both subjective measures by trained professionals and objective data (Rockoff & Speroni, 2011).

"It's studies like this that, as we learn more about effective teaching, will help us pinpoint the most effective skills," Steele said. Steele, who also is project manager for the district's teacher effectiveness initiative, said educators already know quite a bit about the essential elements of good teaching.

"What we're searching for right now are ways to measure teacher effectiveness, but right now, we're saying effective teaching is the person who scores the highest on the measures we have," Steele said. "One of the keys is having multiple measures. There is no one way to measure teacher effectiveness. It is a combination of different skills."

A NEW DEFINITION OF EFFECTIVENESS

Hillsborough began extensive work in spring 2009 to consider ways to improve teaching quality, continuing the Measures of Effective Teaching Project objectives as one of a half-dozen districts in the nation continuing work with

Hillsborough County Public Schools
Tampa, Fla.

Number of schools: **225**
Enrollment: **192,000**
Staff: **12,000**
Racial/ethnic mix:

White:	**40.1%**
Black:	**21.5%**
Hispanic:	**29.6%**
Asian/Pacific Islander:	**3.4%**
Native American:	**.3%**
Other:	**5.1%**

Free/reduced lunch: **57%**
Contact: **David Steele**, chief information and technology officer
Email: **David.Steele@sdhc.k12.fl.us**

the Gates Foundation in an intensive partnership project. The goal, Steele said, is to understand clearly which skills correlate to higher student performance and to work with teachers to develop those skills.

The district formed a teacher evaluation committee. Members spent summer 2009 researching evaluations and settled on adapting Charlotte Danielson's framework for effective teaching as a foundation for observations. Steele noted that while some simply use the framework Danielson published, Hillsborough worked with Danielson to modify it where appropriate for the district's context. Hiring Danielson as a consultant allowed changes to be made to the observation form through conversation and feedback, with Danielson able to explain to committee members the rationale behind the points included and how each worked with others.

"We wanted something first and foremost that was rubric-based, so a teacher would have a clear understanding of what he or she needed to do when being observed," Steele said. "That clarity was something we were looking for."

By spring 2010, a teacher evaluation committee had drafted a new teacher evaluation system with multiple measures. To measure classroom practice, two different observers, both a trained peer or mentor and supervising administrator, use the Danielson-based framework over multiple observations.

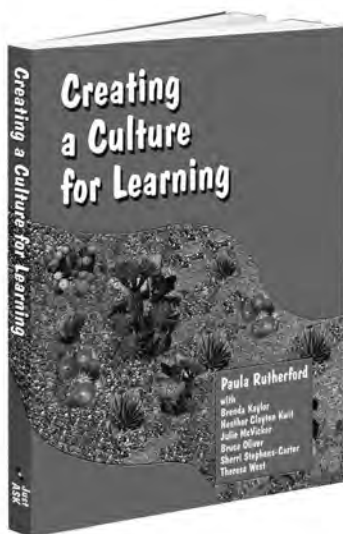The observation form no longer includes 44 items to check, but concentrates on five or six subcategories each in four main areas. Each of the domains within the observation is weighted: planning and preparation, 20%; classroom environment, 20%; instruction, 40%; and professional responsibilities, 20%.

"We want to get a better understanding of exactly which skills correlate most closely to higher student performance," Steele said. "We wrestled with that as a committee — how to weight the categories. Over time, we could very well adjust based on what we learn about the value of the components."

With the earlier observation form, according to Steele, teachers had no real direction for what to do to improve in a specific area for which the observer might have checked off a lower rating. With the new model, the classroom observation sheet is exactly the same as the teacher's end-of-year evaluation, so each time a teacher is observed, he or she knows which areas the observer thought needed to be developed and which looked good.

"It's a much more informative process and leads to a clear understanding on the teacher's part of the strategies he or she needs to use to teach more effectively," Steele said.

The other prong of the evaluation system is student learning data. Working with the University of Wisconsin's Value-Added Research Center, the district developed a method for a value-added assessment of teaching. Teacher evaluations now have student learning growth accounting for 40% of the evaluation.

Peer and mentor observations account for 30% of the evaluation, the administrators' observation for 30%.

"We wanted the value-added (student learning data) to be the biggest single piece" of the evaluation, Steele said, "but we didn't want any one piece to outweigh the other two."

PEER AND MENTOR OBSERVATIONS

The peer and mentor observations are a key component of the evaluations, Steele said, and the most powerful professional learning for teachers and the observers. The district and union had approved a peer assistance program in the 1990s with no evaluative component, but budget constraints sank it before it began.

In 2010-11, the district hired nearly 200 experienced educators for full-time roles as mentors and peer evaluators at a cost of about $12 million, or less than 1% of the district's annual operating budget. Six times as many applied for the positions and were selected by a committee of principals, teachers, members of the union, and district administrators.

Mentors focus their work on supporting new teachers, while peer observers work with veteran teachers. Each earns an additional $5,000 stipend and returns to the classroom after two

POWERFUL
TOOLS

FOR CLASSROOM
OBSERVATION AND EVALUATION

Classroom Walkthrough | **Teachscape *Walk***

Teacher Evaluation Tools | **Teachscape *Reflect* Live**

Video Capture and Sharing | **Teachscape *Reflect* Video**

Online Professional Learning | **Teachscape Professional Learning Suite**

Observer Training and Test | **Framework for Teaching Proficiency System**

⠿ teachscape

or three years. For peer teachers to understand firsthand what daily practice is like, Steele said, they must recently have been in the classroom. Hiring new observers every few years ensures that currency, he said.

"We hope and expect they will be more effective teachers by having had the experience of helping others," Steele said. "We also see that as a benefit. We make them cycle back because a peer adds an immediacy of teaching that a principal won't have."

Peers and mentors conduct formal observations. Mentors work regularly with about 20 new teachers, visiting first-year teachers once a week and second-year teachers every two weeks. For first- and second-year teacher observations, two mentors switch the teachers with whom they work so the mentoring relationship remains pure and the mentor is not seen in an evaluative role.

Peer observers at first had a caseload of about 150 teachers to observe regularly, which the district reduced to 110 teachers in the second year, recognizing that the number was too high.

Each evaluator is trained to conduct three parts of a cycle that helps teachers gain information and reflect on their practices: preobservation conference, observation, and post-observation conference. That reflective piece is essential for powerful learning, Steele said.

Before the preobservation conference, teachers complete a set of questions that they review with the peers and mentors, who use a preconference guide document to help stimulate thinking. Peer observers and mentors might ask:

- What is/are your lesson objective(s)?
- How is the lesson objective aligned with state curriculum standards?
- What data did you use to design this lesson? How did the data influence the planning of this lesson?
- How will you know if your lesson objective was achieved?

After the observation, the evaluators load their ratings into a data management system accessible by the teacher, principal, and evaluators. Teachers then can decide on follow-up.

LESSONS LEARNED

In a report for the Gates Foundation, the district listed six lessons learned:

- Build a foundation of board/district/union shared leadership;
- Create a sense of urgency rooted in student learning;
- Establish a high-capacity executive team of respected district leaders;
- Include teachers and principals in every phase;
- Community clearly and constantly through multiple channels; and
- Incorporate what is learned from implementation quickly.

Source: Hillsborough County Public Schools.

CONDITIONS FOR OBSERVATION

The National Comprehensive Center for Teacher Quality recommends that certain conditions be in place before classroom observations are used in evaluations (Goe & Croft, 2009):

- Use a high-quality observation instrument based on standards of effective teaching practice that include levels of performance.
- Allow teachers time and opportunity to familiarize themselves with the observation instrument so that they will understand what is expected.
- Train observers to use the instrument so that all observers are using it in the same way. The goal is to ensure that a teacher gets the same score no matter which rater conducts the observation. Furthermore, avoid potential rater bias (or the appearance of bias) by using trained raters.
- Calibrate observers. Calibration involves checking the scores of observers to ensure that they are not getting more stringent or lax in scoring over time, a condition called "rater drift."
- When the stakes are high, conduct multiple observations, preferably with different observers.
- For elementary teachers and other teachers of more than one subject, observing when they are teaching different subjects will help identify subject-specific strengths and weaknesses.
- Share rating with the teachers, preferably as part of an individual development plan.

Depending on the teacher's needs, the peers may offer a conference, model a lesson, or provide additional informal observations that focus on classroom skills.

PRINCIPAL OBSERVATIONS

The principal's role changed dramatically under the new system, Steele said. "We do believe very strongly that we want the principal to be the instructional leader of the school," he said. "And if you're going to have a high-stakes (teacher) evaluation, it's a good idea if the principal has actually watched the teacher teach a lesson."

Principals are required to conduct at least one formal observation of each teacher. Steele admitted finding the time is an issue for principals, and for that reason, the district allows assistant principals to conduct required additional administrative observations.

The change, Steele said, has led to principals having daily conversations with teachers about their planning, instructional strategies, and effective lessons, making them true instructional leaders.

The change also led the district to develop a new principal evaluation system based on 10 competencies derived from the state's educational leadership standards. Principals also are evaluated on student learning gains, area director assessment, school operation information, teacher retention, student attendance/discipline, and teacher evaluation accuracy compared with peer evaluators and teachers' value-added scores. The principal evaluation also incorporates teacher feedback.

STUDENT ACHIEVEMENT DATA

Steele said grant money was particularly helpful for the district to purchase software and hardware to collect evaluation data and student achievement data, and to hire help to determine value-added measures. Value-added models are "a collection of complex statistical techniques that use multiple years of students' test score data to estimate the effects of individual schools or teachers" (McCaffrey, Lockwood, Koretz, & Hamilton, 2003, p. xi).

The district uses value-added measures researched by the University of Wisconsin's Value-Added Research Center. The district collects student assessment data, links the data to individual students with information about their backgrounds, links to individual courses and teachers using unique identification numbers, and transmits the data for analysis to the university.

The university then uses the data to compute a value-added measure for each teacher indicating the growth of that teacher's students compared with that of an average district teacher for that subject.

"We truly want to find that best statistical measure of student growth and how it reflects on the teacher," Steele said. However, he continued, "Even within the value-added model, we are trying to get as many measures as possible. Our goal is that no teacher should have just one post-test that is used as a measure."

As data are collected, evaluations will be based on the three most recent years of information.

Because the district had been using student achievement data to pay teachers bonuses from 2002 on and had even earlier begun creating and validating hundreds of pre- and post-test instruments, it was ahead of many districts in which some teachers teach subjects or grades for which there is no exam.

LAUNCHING CHANGE

"Teacher evaluation is the centerpiece to the extent that it

gets publicized in that way, but any of us would say the professional development is the centerpiece; it's just evaluation that gets all the attention," Steele said.

The change to the new evaluation system began with extensive professional development in summer 2010 for principals, assistant principals, and peer and mentor evaluators. All needed to learn about the observation forms and become trained evaluators to use them.

The district worked with a consulting organization to make sure all those who would conduct observations measured each component the same way. After 40 to 50 hours of learning, nearly 700 observers were certified to use the process starting in 2010-11. After a year, their reliability was tested again to recalibrate, if necessary, and ensure that what one observer measured in one school with one teacher would be based on the same criteria as another observer in a different school with a different teacher.

The evaluators conduct formal, full-period observations. In

the first year of the new system, the most excellent teachers were observed three times and the most struggling were observed 11 times. For 2011-12, the district added informal unannounced observations, decreasing full-period formal visits and supplementing them with pop-in informal observations lasting 10 to 15 minutes.

The professional conversations that result from the observations are the most powerful form of professional learning teachers can experience, Steele said. The district also has made reference tables for teachers, Steele said, so if a teacher sees from an observation that he or she needs to strengthen a particular area, the teacher can look at the table for courses that will build that skill.

In addition, Steele said, the district's director of professional development has worked with each district trainer to deconstruct professional development course outlines and align each course to the new evaluation. Some courses may be eliminated, he said. The University of Wisconsin also helped the district identify some of its professional development courses that would benefit from improved pre- and post-tests, and the district's teachers worked to strengthen them.

"We want to make sure that our professional development is aligned with what we evaluate," Steele said. "We want to make sure our trainers are all on the same page when they are talking with teachers about effective pedagogical strategies."

The district has consistently dedicated $12 million to $15 million a year out of the operating budget to formal professional learning, Steele said. He said the district used the American Recovery and Reinvestment Act money it received for added professional learning both for trainers and courses and to offer teachers stipends for evening and summer work.

Another change in the system was to the final result of the observations and value-added data. Overall teacher ratings were changed from unsatisfactory, basic, proficient, and distinguished to: requires action, developing, accomplished, and exemplary. By specifying "requires action," the emphasis is on helping teachers improve their practice, Steele said.

In 2013-14, the district plans to have collected three years of data and to begin linking teacher evaluations to compensation and teacher promotions. Teachers who receive an unsatisfactory rating in two consecutive years face a dismissal process. Of the 250 lowest-evaluated teachers in 2010-11, 72 did not return to teach in 2011-12, Steele said. He said principals counsel those who are underperforming, and some leave.

In the first year of the new evaluation, only three teachers had two consecutive unsatisfactory ratings. Now, 100 have been identified with unsatisfactory performance and could receive a consecutive poor rating. However, Steele said, "It's not our goal to lose them; it's our goal to improve them."

"We think that too many districts caught up in the current evaluation craze around the country have the idea you're going to measure your success by how many ineffective teachers you

ADVANTAGES AND DISADVANTAGES OF VALUE-ADDED MEASURES

A comparison of value-added measures and classroom observation for teacher evaluation.

	Advantages	Disadvantages
Value-added measures	• Relatively inexpensive (after initial infrastructure costs). • Focuses solely and directly on student learning. • Relatively objective. • Comparable across schools, districts, and even states (if they are using the same statistical methods and achievement tests).	• Costly to build necessary data system; generally requires hiring experts to set it up and conduct the analyses. • No information about what effective teachers do in the classroom. • No information to help "bad" teachers improve. • No information for some teachers (e.g. special education, art, music, early elementary).
Classroom observation	• High face-validity and teacher buy-in. • Allows teachers to understand and participate in the evaluation process. • Useful for formative evaluation, particularly for novice teachers. • Based on "best practices."	• Higher personnel costs. • May not take student achievement into account. • Scores determined by evaluators with different levels of training. • May be affected by whether measures are used for high-stakes or low-stakes evaluation.

Source: Goe & Croft, 2009.

fire. Our approach is to measure our effectiveness by how many ineffective teachers we improve. You spend too much money recruiting and growing a teacher to discard them without trying professional development to give them a chance to improve."

As the district looks toward continuing to refine teaching practices that lead to student learning, the new system is making a difference. Steele said the principals advisory group has consistently reported over the past 15 months that "teachers have really raised their game. They understand what the expectations are, and they are teaching at a higher level than they've ever taught before."

REFERENCES

Goe, L. & Croft, A. (2009, March). *Methods of evaluating teacher effectiveness: Research-to-practice brief.* Washington, DC: National Comprehensive Center for Teacher Quality.

Hillsborough County Public Schools. (n.d.). *Leading change in Hillsborough County Public Schools.* Tampa, FL: Author. Available at http://communication.sdhc.k12.fl.us/eethome/casestudies.

Measures of Effective Teaching Project. (2010, December). *Learning about teaching: Initial findings from the Measures of Effective Teaching Project.* Seattle, WA: Author.

McCaffrey, D.F., Lockwood, J.R., Koretz, D.M., & Hamilton, L.S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND Corporation. Available online at www.rand.org/pubs/monographs/2004/RAND_MG158.pdf.

National Board for Professional Teaching Standards. (2011). *Student learning, student achievement: How do teachers measure up?* Arlington, VA: Author.

Rockoff, J.E. & Speroni, C. (2011, October). Subjective and objective evaluations of teacher effectiveness: Evidence from New York City. *Labour Economics, 18*(5), 687-696.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect.* Santa Cruz, CA: The New Teacher Project.

•

Valerie von Frank (valerievonfrank@aol.com) is an education writer and editor of Learning Forward's books. ■