education sector reports

Inside IMPACT: D.C.'s Model Teacher Evaluation System

By Susan Headden

EDUCATIONSECTOR
www.educationsector.org

ACKNOWLEDGEMENTS

I would like to thank all the DCPS teachers, principals, master educators, and administrators who somehow found time in their packed schedules to share their insights and experiences with me. Thanks also go to my wiser Education Sector colleagues for their helpful feedback and to Robin Smiles for her thoughtful editing and patience.

ABOUT THE AUTHOR

SUSAN HEADDEN is senior writer/editor at Education Sector. She can be reached at sheadden@educationsector.org.

ABOUT EDUCATION SECTOR

Education Sector is an independent think tank that challenges conventional thinking in education policy. We are a nonprofit, nonpartisan organization committed to achieving measurable impact in education, both by improving existing reform initiatives and by developing new, innovative solutions to our nation's most pressing education problems.

1201 Connecticut Ave., N.W., Suite 850, Washington, D.C. 20036
202.552.2840 • www.educationsector.org

For public school teachers, June is traditionally a time to exhale. The requisite tests have been given, the last lessons delivered, the artwork torn from the walls, rolled up, and sent home to parents. In the best cases, there is a sense that most of what students needed to learn they did, allowing the teacher, if not riches or public recognition, at least the personal satisfaction of having done a hard job well. But this year, as classes wind down in the District of Columbia Public Schools, teachers will not be breathing freely until they see one final judgment of their pedagogical efforts—a report that, it is no exaggeration to say, has the power to end careers.

The anxiety comes from the new teacher evaluation system known as IMPACT, a rigid, numerically based process that rates teachers primarily on classroom observations and student test scores. As one of the first in the nation to link teacher performance, pay, and job security to such measures, IMPACT is the most polarizing of the bold reforms initiated by ex-schools Chancellor Michelle Rhee. In the two years since this high-stakes report card was launched, it has led to the firing of scores of educators, put hundreds more on notice, and left the rest either encouraged and re-energized, or frustrated and scared. It almost certainly cost the local union president his job, and it helped force the mayor who supported it, as well as Rhee, out of office.

IMPACT sets clear expectations for effective teaching, from probing students' understanding to coming to work on time. Many teachers in the district welcome these standards and are motivated by salary bonuses of up to $25,000 to prove they can meet them. Others complain of being judged on elements of a craft that they insist can't be measured. But whether they are critics talking bitterly of being "impacted" or boosters talking about "getting great feedback on my 'Teach 1,'" D.C. teachers are speaking a new language—that of the rubric by which they are measured. And that is an unmistakable sign that IMPACT is changing the way many teachers teach.

As school districts around the country work to devise their own evaluation systems that include student test scores (so-called value-added measures) and classroom observations, they are closely watching how this high-profile prototype is playing out in the nation's capital. As they do, they will find encouraging lessons in how codifying best practices can be used to objectively assess teachers and help them improve, and how greater accountability can considerably enhance the public's faith in a school system. But they will also see how difficult it is to calibrate such a powerful tool so that it works in practice as intended. Nonetheless, multiple-measures teacher evaluation is the future of K-12 education. And in Washington, D.C., the future is happening now.

Defining Good Teaching

Anyone who has ever attended school or sent a child to one knows that some teachers are better than others. It's true in every other field of endeavor. But, as the organization known as The New Teacher Project reported in 2009, teacher evaluation systems fail to make these distinctions, treating all educators as if they're essentially the same.[1] So, before meaningful evaluations could take place, educators had to recognize that what teachers do, or don't do, has a profound effect on how much students learn.

At the time IMPACT was developed, even its staunchest opponents would have agreed that D.C. needed a new way to evaluate teachers. In 2007, when then-mayor Adrian Fenty assumed control of the city's vast school system, the district's scores on the National Assessment of Educational Progress were among the lowest in the nation, and its black-white achievement gap was the widest of 11 urban districts that reported their results. Those grim statistics came despite the fact that the city spent more money per pupil—nearly $13,000—than most of the largest public school systems in America.[2]

The data loudly suggested that D.C.'s teacher evaluation system, as with most others in the country, was ineffectual. Based on once-a-year observations, the system graded more than 3,000 teachers on a perfunctory checklist—allowing less than an inch of space for comments—and found, remarkably,

"I could have spent a whole class teaching nothing but the color yellow, and no one would have noticed."

that virtually all of them were doing a fine job: Fully 95 percent of teachers were rated "satisfactory" or above. One middle school teacher summed up the typical level of vigilance this way: "I could have spent a whole class teaching nothing but the color yellow, and no one would have noticed."

Reforms to the evaluation process took root under former superintendent Clifford Janey. But the push to raise teacher accountability went into overdrive with the arrival of Rhee, the blunt-spoken founder of the New Teacher Project who brought to the top job determination and energy along with an acknowledged shortage of public relations skills. Given wide latitude and full support by Fenty, Rhee shook up DCPS by closing schools, firing administrators, hiring new principals, and making countless enemies along the way.

At the core of all her efforts was improving the quality of instruction. And with a document known as the Teaching and Learning Framework, district officials worked to precisely define what good teaching was. As explained in a recent report by the Aspen Institute, the framework provided a way for principals, teachers, and administrators to work together to improve instruction.[3] Instead of focusing on what to teach, they concentrated on how to teach, with explicit directions that cut across different subject areas. "We focused first on pedagogy, whereas most other reforms focused on curriculum," says Scott Thompson, director of teacher effectiveness strategy for DCPS. "You could have the greatest curriculum in the world, but if the teachers are ineffective in conveying it, then it's not going to matter."[4]

Non-educators may be surprised to know that there is no universally accepted definition of good teaching. But the Teaching and Learning Framework is D.C.'s attempt to write one. And its nine commandments form the all-important rubric on which classroom performance is judged. They are as follows:

1. Lead well-organized, objective-driven lessons.
2. Explain content clearly.
3. Engage students at all learning levels in rigorous work.
4. Provide students with multiple ways to engage with content.
5. Check for student understanding.
6. Respond to student misunderstandings.
7. Develop higher-level understanding through effective questioning.
8. Maximize instructional time.
9. Build a supportive, learning-focused classroom community.

In the months since they were written, these directives and their related elements have been reduced to shorthand in the parlance of teachers—"Teach 1, Teach 2"—and, inevitably, committed to memory.

Overall, the IMPACT system rates teachers on a combination of factors, some weighted far more heavily than others. Classroom performance on the Teaching and Learning Framework counts for 35 percent of a teacher's overall rating; student test scores (so-called value-added data) for teachers in grades that take standardized tests count for 50 percent; commitment to the school community gets

10 percent; and school value-added data—a measure of the school's overall impact on student learning—is worth another 5 percent. On this last measure, all teachers in a school receive the same score. (See Figure 1.)

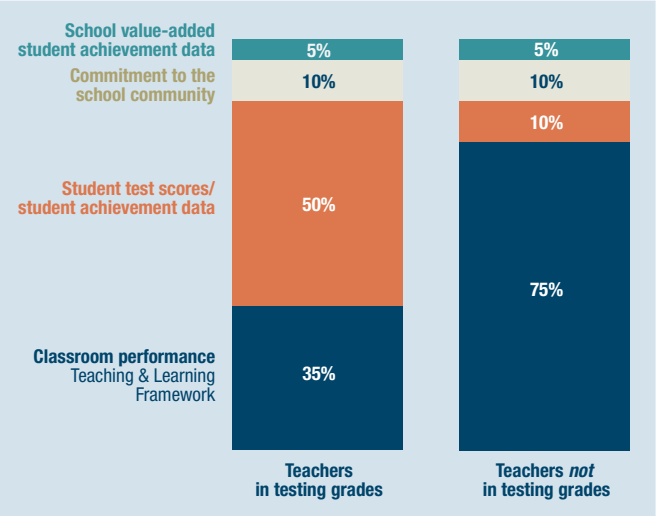
Teachers who are not in testing grades—whose students are not required to take standardized reading and math tests—do not receive value-added data, and so their classroom performance becomes even more important, counting for fully 75 percent of their score. For these teachers, a component called "teacher-assessed student achievement data" counts for 10 percent, and the other factors count the same as they do for the other teachers. For both categories of teachers, the final score is then adjusted based on a factor called "core professionalism," which covers things like respecting parents and coming to work reliably and on time. A less than satisfactory rating on this measure cuts 10 points off the teacher's overall score.

The value-added measure is, of course, controversial, tying as it does teacher performance to factors they say are very often beyond their control. And it has drawn further fire with recent reports of cheating by teachers and administrators on the tests on which it is largely based.[5] Yet, surprisingly, that is not what has teachers most agitated. What IMPACT really comes down to for the 86 percent who are not in testing grades is classroom observation. Even more than the test scores, it is this method of measuring teachers' on-the-job performance that critics say can treat them too subjectively and, by extension, misjudge them, mischaracterize them, and force them to teach in an overly prescriptive way.

The View From the Classroom

Every teacher in the district is observed five times a year: three times by a school administrator (usually the principal) and twice by a "master educator," an outside teacher trained in the same discipline who is seen as an impartial third party. The observations take 30 minutes—usually no more and never any less—and all but one of the administrator visits are unannounced. Based on these observations, teachers are assigned a crucial ranking, from 1 to 4. Combined with other factors, they produce an overall IMPACT score of from 100 to 400, which translates into

Figure 1. What Teachers Are Graded On



Note: Currently, the use of student test scores is limited to teachers who teach reading or math in grades four through eight.
Source: District of Columbia Public Schools.

"highly effective," "effective," "minimally effective," or "ineffective." A rating of ineffective means the teacher is immediately subject to dismissal; a rating of minimally effective gives him one year to improve or be fired; effective gets him a standard contract raise; and highly effective qualifies him for a bonus and an invitation to a fancy award ceremony at the Kennedy Center.

It is a measure of how weak and meaningless observations used to be that these pop visits can fill teachers, especially the less experienced ones, with the anxiety of a 10th-grader assigned an impromptu essay on this week's history unit for a letter grade. The stress can show up in two ways—the teacher chokes under the pressure, thereby earning a poor score, or she changes her lesson in a way that can stifle creativity and does not always serve students. Describing these observations, IMPACT detractors use words like "humiliating," "infantilizing," "paternalistic," and "punitive." "It's like somebody is always looking over your shoulder," said a high school teacher who, like most, did not wish to be named publicly for fear of hurting her career.

Teachers commonly protest that 30 minutes is an impossibly small window through which to view their ability to convey content and connect with students. Even though they recite the rubric in their heads and keep cheat sheets on Post-it notes around the classroom, they say their individual lessons cannot

possibly hit everything on the IMPACT checklist—a word that district officials would disavow—in that time frame. Making sure students understand the objective (Teach 1) is one directive they often miss. Sometimes the objective is implied; sometimes it's deliberately revealed slowly. Moreover, some of the requirements don't fit every lesson. The original framework called for "providing students multiple ways to engage with content." But if a teacher is instructing pre-kindergartners about texture, for example, she need only teach through touch. So, under the new framework, teachers can meet this standard even if they target just one learning style. The district also reduced the number of standards to assess behavior from three to one.

Another frequent complaint is that IMPACT fails to account for the stark differences in demographics among the district's schools—from those educating the children of U.S. senators to those serving the offspring of welfare recipients—and the unique challenges that confront teachers in the city's lower-income wards. The compensation system, however, does consider these factors: Teachers in low-income schools are eligible for higher salary bonuses. DCPS counts 62 percent of its 46,515 students as eligible for reduced-price lunch, a proxy for poverty. Low incomes can bring a number of social ills, including substance abuse, gang participation, and parental unemployment. Students who are acting out the effects of such problems can easily turn a good lesson sour, and it is the bad fortune of the instructor trying to conduct that lesson to be visited by a master educator on that day.

"Out of 22 students, I have five non-readers, eight with IEPs [individual educational plans, which are required by federal law for students with disabilities], and no co-teacher," says the middle school teacher. "The observers don't know that going in, and there is no way of equalizing those variables." The teacher said she wished to remain anonymous "because we are in this culture where acknowledging the truth of the challenge is misconstrued as having low expectations." Another teacher told the *Washington Post* that his students try to "sabotage" his class: "They deliberately play dumb so they can get you fired," he said.[6] Nathan Saunders, the president of the Washington Teachers Union, who was elected last fall on a platform of radically changing IMPACT, says that because the system doesn't accommodate such vagaries, it's no surprise that just 5 percent of district teachers rated "highly effective" last year were in the high-poverty Ward 8, whereas 22 percent were in the relatively affluent Ward 3.[7]

District administrators hear this objection routinely, and their response is both simple and frankly unsympathetic: If you are a good teacher—if your lessons are engaging, lively, and challenging—you will not have problems with classroom management. (Indeed, both of the teachers cited above were rated solidly "effective.") "Behavior and instruction always dovetail," says Cynthia Robinson-Rivers, a master educator specializing in early childhood instruction.[8] "When you hear a teacher say '1, 2, 3—eyes on me' (a common ditty for getting children's attention) then it's often too late. You are reacting to an action; you are

DCPS's Nine Commandments of Good Teaching

Teach 1
Lead well-organized, objective-driven lessons

Teach 2
Explain content clearly

Teach 3
Engage students at all learning levels in rigorous work

Teach 4
Provide students with multiple ways to engage with content

Teach 5
Check for student understanding

Teach 6
Respond to student misunderstandings

Teach 7
Develop higher-level understanding through effective questioning

Teach 8
Maximize instructional time

Teach 9
Build a supportive, learning-focused classroom community

not preventing it." This does not mean the evaluator can't adjust the score if she learns, for instance, that a hyperactive child has forgotten to take his medication. "We're not unreasonable," Robinson-Rivers says. But she says administrators are insistent about the larger goal: "We must have high expectations for all students, regardless of their home experiences."

A Receptive Audience

A case in point is the lively classroom of Andrea Stephens (not her real name), a first-grade teacher at a racially mixed elementary school in Northeast D.C. Master educator Robinson-Rivers is conducting an informal observation* as Stephens teaches a lesson about capital letters, punctuation marks, and the short "a." Stephens is kind, firm, and engaging, and she wins points for gestures like asking a reluctant pupil if she could "get one of his smiles," making him feel valued. But she is apparently not engaging enough. Several students are not paying attention; one is a mugger and a performer, and he can't sit still. After several attempts to quiet him, Stephens gently pulls him up next to her, holding his hand while she addresses the rest of the class. The general atmosphere suggests to Robinson-Rivers a need for better management. "The children weren't completely out of control," Robinson-Rivers says. "But if they aren't facing you it can suggest a lack of interest."

The session reveals other perceived shortcomings, despite Robinson-Rivers' respect for Stephens as "a warm, thoughtful practitioner." It was too teacher-directed, Robinson-Rivers says; it failed to make the objectives fully clear, and it didn't make the most of limited instructional time. "If the pacing is too slow, you can lose valuable time from the lesson," Robinson-Rivers says. "If in a 20-minute morning meeting the kids participate in a variety of engaging activities, it's much easier to maintain their interest and enthusiasm." Stephens also falls short on Teach 5—checking to see whether students actually understood her. "There was no way to know whether the shy girl or the boy who spoke little English understood or not," Robinson-Rivers says. Instead of having all the pupils answer in unison, she suggests that Stephens cold-call on individual students, or have all the boys or all the girls answer in some non-verbal way. "It's hard because teachers do think they are checking for understanding. But it's actually an easy one for professional development; you could just say there are three easy things you can do."

Stephens, whose overall score for the year was in the "effective" range, is open to evaluation and receptive to feedback—she even asked for an extra observation—and in this regard, master educators say she is fairly typical. Matt Radigan, another master educator specializing in elementary instruction, says he has been happily surprised by how willing teachers have been to engage with the evaluators even when the news is bad.[9] Robinson agrees, saying, "We expected more hostility [to the feedback sessions] but usually they go just fine. I evaluated 230 teachers last year, and I can only name four or five who were hostile." Radigan says he performed 220 observations last year and 170 this year "and maybe two per cycle are upset." With rare exceptions, teachers generally assess themselves the way the evaluators do, the IMPACT team has found. "It's not usually wildly different," Robinson-Rivers says. "When the class didn't go well, teachers know it didn't go well."

With rare exceptions, teachers generally assess themselves the way the evaluators do, the IMPACT team has found.

Teachers' outwardly gracious attitudes about their evaluations likely has to do with two very different factors. One is simply that the master educator holds all the cards—the teachers have virtually no input in the evaluation, and appeals of the scores are rarely successful. But teachers, most of whom work in relative isolation, are also hungry for meaningful feedback. They get it from these energetic, highly credentialed educators who are carefully screened not only for their technical skills but for their bedside manners. Of the 800 who applied for the job, only 32 were selected.

The teachers who spoke to Education Sector almost universally liked the people who evaluated them, finding them for the most part helpful, empathetic,

* Informal evaluation for feedback only.

and smart. Radigan says he always lets the teacher lead off the feedback session. "If they want to vent about how much they hate IMPACT," he says, "I let them vent." Master educators don't see any pattern in teachers' responses, particularly. "There is no generalizing or stereotyping that you can ever make," says Robinson-Rivers, "because every time you do, you are [wrong]. There are older veterans who may be super-open about getting a tough score and young, bubbly ones that you assume are going to be open, and they are really tough and question everything."

A Case of Inconsistency

Bill Rope is not young, or particularly bubbly, but he is a respected teacher who sees this unusual relationship from the confident perspective of an older man who went into education after a 30-year career in the foreign service. Rope, who now teaches third grade at Hearst Elementary School in an affluent neighborhood of Northwest D.C., was rated "highly effective" last year and awarded a bonus that he refused to accept in a show of union solidarity.

But a more recent evaluation served to undermine whatever validation the first one may have offered. In the later one, a different master educator gave him an overall score of 2.78—toward the low end of "effective." Although she gave Rope 3s and 4s on "higher-level understanding" and "correcting student misunderstanding," she rated him only minimally effective at "maximizing instructional time." As evidence, the master educator cited students "engaged in off-task conversations" and two who left their seats "to sharpen pencils when pencils were not required." (That was odd, Rope says, because the room has no pencil sharpeners.)[10] Rope was also downgraded for giving students only two ways to engage in content "when more would have been appropriate." And although his use of an illustrated anthology book matched the objective of the lesson, the evaluator said that "all students were not engaged or called on." The latter observation seemed to contradict her praise for Rope on another metric, which was that "students willingly raised their hands, and those who did not seemed comfortable responding to Mr. Rope." The evaluator also rated Rope only minimally effective at "engaging students at all learning levels in rigorous work."

As Rope sees it, several of these observations made little sense. "How can you [engage students at all levels] in 30 minutes and also put across challenging material?" he asks. "What about calling on one or more students more than once? If weak students are doing well, you might want to do that." The evaluator suggests, among other strategies, having the students fill out a worksheet, an activity Rope dismisses as one that would slow down dynamic discussion. To improve behavior, the evaluator suggests Rope prepare a poster-sized contract, evidently missing the big rules chart, signed by all students, that Rope has already displayed. In an unusual move, after objections from Rope, the master educator adjusted the scores on two measures, resulting in a higher rating.

Rope, who has been active in the teachers union, does not seem troubled by all this so much as he is

Table 1. How a Highly Effective Teacher Might Score*

Component	Component Score (Scale of 1–4)	Percentage of Score	Weighted Score
Individual Value-Added Student Achievement Data	3.5	x 50	= 175
Teaching and Learning Framework	3.7	x 35	= 130
Commitment to the School Community	3.5	x 10	= 35
School Value-Added Student Achievement Data	3.3	x 5	= 17
TOTAL			**357**

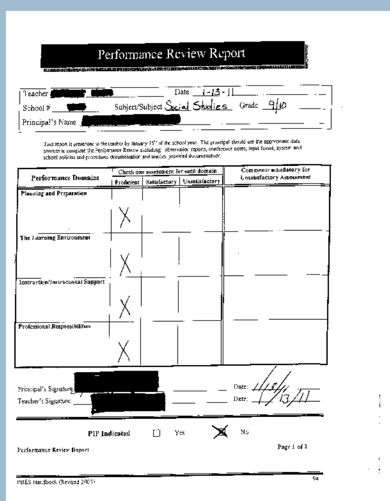
*Teacher in a testing grade.
Component Score Scale: 1=ineffective, 2=minimally effective, 3=effective, 4=highly effective.
Overall IMPACT Score Scale: 100–174=ineffective, 175–249=minimally effective, 250–349=effective, 350–400 highly effective.
Source: District of Columbia Public Schools.

irritated by its apparent pettiness and inconsistency. Perhaps most important, he says he worries about the system's effect on teaching. Last year, he says, he did his best to satisfy all of IMPACT's demands. "I would be hitting everything. I did everything you were supposed to do, and I hated it," he said. "It took so long to do everything you were supposed to do. The biggest problem is the narrowing of the curriculum." Says another teacher, who did not want to be named: "I am a worse teacher when I try to fit into [IMPACT's] scheme than when I am myself." Teachers, it seems, are now teaching to their own test.

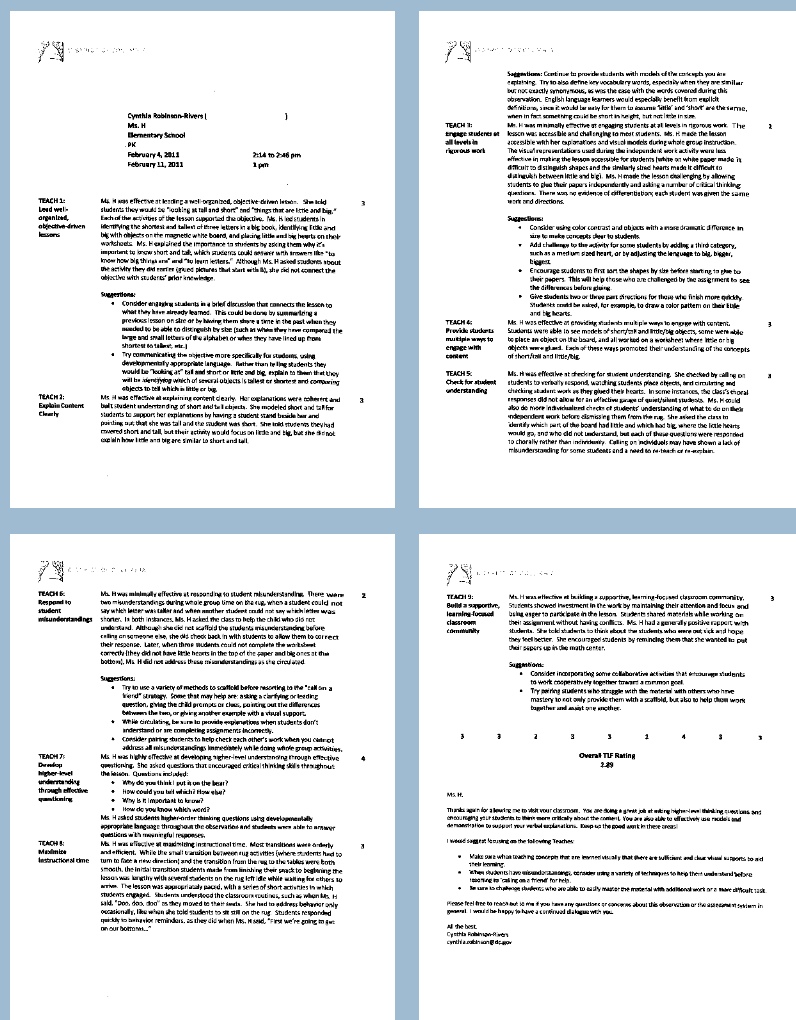
IMPACT's architects reject the argument that the system is overly prescriptive, especially since the rubric already has been streamlined in response to first-year concerns. Good teachers routinely demonstrate every element on the Teaching and Learning Framework without even thinking about it, they say, like touch-typists who don't look at the keys. "It's not as if this is a new way of teaching," insists Thompson. "Good teachers get high marks for doing what they are already doing." (Indeed, some principals complain that the IMPACT standards are not rigorous enough.)

Figure 2. Comparing Evaluations

Evaluation of a teacher in Baltimore City Public Schools:



IMPACT evaluation by a D.C. master educator:



Such reassurances, though, don't prevent teachers from keeping cheat sheets in their desks and from switching strategies or entire lesson plans at the last minute to impress an unexpected visitor. "Teachers aren't stupid. Do you think they are really doing these things? They do them only for the 30 minutes they are being observed," says Marni Barron, an instructional coach at Hearst. "They pull out a new lesson plan they have in their drawer for an occasion just like this. They say [about whatever they were doing] 'Oh kids, never mind. I think we are going to learn about the planets today.'"[11]

Predictably, D.C. teaching circles are abuzz with gripes and rumors about the perceived subjectivity of their scores—ratings that vary from one evaluator to the next, a master educator who didn't "get" a lesson,

IMPACT's architects reject the argument that the system is overly prescriptive, especially since the rubric already has been streamlined in response to first-year concerns.

or, as with Rope, being dinged for missing the mark on one aspect of the rubric. Barron talks of a teacher "so phenomenal that I would have her teach my kid from K through 12 if I could" who was rated "minimally effective" on her most recent evaluation. Teachers widely believe scores are lower this year than they were last year. (They are, but negligibly so.) One says her principal has a stated policy of never giving fours. "Four is a stretch because you have to show growth," says the teacher, who did not want to be named. Her belief that 3 is the new 4 prompts Barron to ask: "If we are telling our teachers to shoot for a B, why are we telling our students to shoot for an A?"

In fact, DCPS data does not support many of these arguments. In response to charges of inconsistency and grade "deflation," administrators have checked scores and found significant differences only in less than 1 percent of teacher observations. The district has found that the scores given by principals and master educators have been remarkably similar: In only five out of 3,500 evaluations was there a gap of larger than two points between master educator and principal scores. (The principal can see the master educators' scores, but not vice-versa. The thinking is that the principal is partly responsible for the teacher's growth, although the risk is that he will adjust scores up or down to compensate for ratings given by master educators.) To make sure that that everyone considers the same performance to be worth the same grade, the master educators "norm" the scores; they have spent hundreds of hours watching videos of teachers in action, role playing, and discussing what constitutes a 2, a 3, and so on. Teachers can appeal their observation scores, but they rarely do, and only 15 percent of appeals last year were successful.

So how did it all shake out? At the end of IMPACT's first year, 15 percent of teachers were rated highly effective, 67 percent were judged effective, 16 percent were deemed minimally effective, and 2 percent were rated ineffective and fired. Perhaps encouraging to both teachers and the general public, average scores given by both master educators and principals were right around 3—not bad. Based on preliminary scores, Thompson reports "a sizeable number" of teachers this year who appear to be moving from effective to highly effective. As to estimates of how many teachers appear to be moving in the other direction, he declines to say.

The Value of Test Scores

The beauty of the D.C. IMPACT system, as even its detractors agree, is that it includes multiple measures of effectiveness so that a teacher is not judged on just one thing. Teachers overwhelmingly told the district that this sort of diversification was what they wanted, and numerous studies support them. However fraught the classroom observations may seem, each visit by a master educator counts for just 14 percent. Says Robinson-Rivers: "You can get a 2 from me, a 3 from another ME, and a 3 from your principal and still come out strong." And in any case, for many teachers, the observations count for less than half of their score. The rest, for good or for ill, is based largely on student test scores.

Unlike teacher observations, which principals have long conducted to size up their teaching talent, if not to actually grade it, the use of value-added metrics

to judge teachers has emerged as a focus of intense debate. On the one hand, much research shows that the best predictor of teachers' future effectiveness is their past performance on just such measures. On the other, value-added scores can fluctuate from year to year, and from class to class, and they can't completely account for student characteristics—including learning disabilities—that make the jobs of some teachers especially hard. D.C.'s first two years with this controversial measurement puts a fine point on the issue, showing how harsh a measure it is in practice and suggesting ways it may need to be refined.

Specifically, the individual value-added (IVA) score is a measure of the influence a teacher has on student learning based on the D.C. Comprehensive Assessment System (DC CAS), the standardized test given to students every spring. For now, this data is available only for those teachers who teach reading and math in grades four through eight. But because the district plans to test more grades in the near future, the value-added score will become a key gauge for more and more teachers. In fact, Jason Kamras, chief of DCPS's Office of Human Capital Management, says the majority of D.C. teachers will be subject to value-added measures within the next five years. He calls the measure "the one solid anchor we have—more predictive of performance than the number of years you've taught or the number of degrees you have."[12]

District administrators have generated criticism for not providing more precise details on how the value-added measurement is calculated. But according to a report by Mathematica Policy Research, it measures the performance of school and teacher test scores and other data "in a statistical model designed to capture student test scores that are attributable to the school or teacher compared with the progress the student would have made at the average school or with the average teacher."[13] The measurement is called "value-added" because it attempts to isolate how much the school or teacher contributes to score improvements apart from factors outside the teacher's or school's control. Every April, the standardized test scores of a teacher's students are compared with the scores of those same students from the previous April. Taking into account the demographic makeup of the students, such as poverty and English language

Unlike teacher observations, which principals have long conducted to size up their teaching talent, if not to actually grade it, the use of value-added metrics to judge teachers has emerged as a focus of intense debate.

classifications, the district then scores the teacher from 1 to 4 on the students' growth.

Value-added is a relative measure, meaning that, as with sorting high school students by grade-point averages, it compares teachers to their peers and ranks them accordingly. The district has set the mean at 50 percent, so, by definition, no matter how effective the teachers may be, half of them will fall below the median and half will be above. (By contrast, the score from the observations is an absolute measure, which means it is theoretically possible for all the teachers to be ranked the same. Overall, the average scores for observations are a little higher than the value-added scores.)

Aaron Pallas, a professor of sociology and education at Teachers College, Columbia University, is among those who find flaws in the value-added methodology, questioning in particular why the threshold of competence is set at 50. "It's purely a matter of judgment why the average is 50 percent," he says. "They can set the threshold anywhere."[14] Pallas also notes that value-added measures carry statistical margins of error, and that IMPACT fails to take that uncertainty into account. What is now given as a precise number, he says, should instead be expressed as a range. "It really is a lot squishier," he says. "The mean could be from 50 to 90, or the single best estimate. Other values are possible, plausible, and can't be ruled out."

From all of this Pallas has concluded that "the system is rigged to label teachers as effective or minimally effective as a precursor to firing them." To which

Two Alternative Models: Cincinnati and Montgomery County

The IMPACT teacher evaluation system is testament to the belief that improving educational outcomes depends on the quality of teaching more than anything else. "Despite all the challenges, great teachers can close the achievement gap," says Jason Kamras, director of human capital management for the District of Columbia Public Schools. "We need to know who the great teachers are, who needs help, and who we need to transition out."

Before DCPS devised its system for doing that, officials conducted 150 focus groups with 1,500 individuals, taking inspiration from promising aspects of existing systems, or, in other cases, going a different route. The best evaluation systems, studies have shown, involve multiple measures, extensive professional development, reliable measuring instruments, and accountability.[1]

As successful models, educators often point to systems used by Cincinnati Public Schools, an urban and largely African-American district, and Montgomery County, Md., a large suburban district that is more affluent and increasingly diverse. Both feature elements that D.C. teachers often say they would like to see more of: early and aggressive intervention, true peer review, and input from teachers themselves.

Cincinnati's Teacher Evaluation System is all about early intervention and clear consequences. New teachers in that district, which has 33,000 students, most of whom are eligible for reduced-price lunch, get at least two formal and two informal evaluations before December of their first year. If they don't measure up, they are observed four more times that school year, with only one of the visits announced. New teachers who do meet the standards get only one more evaluation, again unannounced.

At the end of their fourth year, teachers receive a comprehensive evaluation. If they do well, they receive tenure. But tenure doesn't mean they are home free. If an administrator or fellow teacher believes a teacher is not effective, she can recommend the teacher get individual remediation. The principal then conducts two observations and draws her own conclusions. The case is then reviewed by a joint union-administration panel, which recommends either dismissal or "intervention"— a year of intense remediation with a fellow educator known as a Consulting Teacher.

Next door to D.C., Montgomery County, Md., is a district with 145,000 students and some schools ranked among the best in the country; it sends 84 percent of its students on to college. It also has a highly regarded teacher evaluation system based on a longstanding system in Toledo, Ohio, that Washington's teachers say gives teachers more and better professional help and more chances to redeem themselves.

Under the system known as Peer Assistance and Review, experienced teachers act as mentors for new ones, as well as helpers and counselors for more experienced educators who are having trouble. As with the Cincinnati system, if these interventions fail, a panel of teachers and principals can vote to dismiss the teacher. As in D.C. and elsewhere, the PAR system proves how ineffectual the previous evaluations were: In the 10 years before the program started, according to the county, five teachers were fired. In the 11 years it has been in place, 200 have been dismissed, and 300 more chose to leave rather than go through the intervention process.

Unlike the D.C. system, which was implemented with unusual speed, Montgomery County's system was rolled out over a number of years, with the full backing of the teachers' union. Also unlike the D.C. system, Montgomery County's teacher evaluations do not now include student test scores. Superintendent Jerry Weast, who will retire this year, has said that he does not believe the scores to be reliable.

Notes

1. Steven Glazerman, Dan Goldhaber, Susanna Loeb, Stephen Raudenbush, Douglas O. Staiger, and Grover J. Whitehurst, *Passing Muster: Evaluating Teacher Evaluation Systems* (Washington, DC: The Brookings Institution, Brown Center Task Force Task Force on Teacher Quality, April, 2011); *Building Teacher Evaluation Systems: Learning from Leading Efforts* (Washington, DC: The Aspen Institute Education & Society Program, March 2011).

Thompson responds, predictably: "It is not rigged. But, yes, we had to make a decision [on the mean], and we wrestled with where to put it." Given what Thompson calls the "huge disconnect" between past teacher evaluations and student achievement, he says "you would be hard-pressed to say that the mean belongs much higher than 2.5." The mean is not likely to move next year, but Thompson says it could change later. "If we see improvements in student achievement, we can recalibrate," he said, "but we don't want to shift the target every year."

Theoretically, a teacher's value-added score should show a high correlation with his rating from classroom observations. In other words, a teacher who got high marks on performance should also see his students making big gains. And yet DCPS has found the correlation between these two measures to be only modest, with master educators' evaluations only

slightly more aligned with test scores than those of principals.

In a perfect world, a high correlation would be .8 or .9. In fact, it is .34. The finding is perhaps not surprising given that tests measure limited competencies, whereas good schools teach a far broader set of skills. Indeed, noting that that high correlations are rare in the social sciences, Thompson calls the figure "moderately strong" and "relatively encouraging." As for variations, the district has found only a handful of cases in which the scores from classroom observations are much higher than the value-added scores. In fewer than 10 out of 434 cases was there a gap of more than two points between these two indicators. Elsewhere, researchers have surmised that gaps may have occurred because teachers performed well in individual classes but failed to present appropriate content overall or in the right sequence over the course of the year.

Assessing student learning in non-testing grades has proven more problematic for evaluators. The first iteration of IMPACT required teachers in this group to show data three times a year that proved student learning. Principals reviewed the information and scored the teachers from 1 to 4, a rating that accounted for 10 percent of teachers' overall IMPACT score. Although teachers were given guidance about how that learning could be measured, they sometimes disagreed with their principals about what should serve as the instruments—portfolios, reading tests?—and what reasonable goals should be. The district is now working to come up with a common assessment for teachers in these grades.

Cheating, of course, significantly distorts the playing field; the teacher who fudges the numbers on students' tests is judged against the teacher who doesn't—and often comes out ahead.

Many teachers say they are happy to be judged on the basis of value-added scores. "Bring it on," says a young teacher in a Northeast D.C. elementary school. "I am confident enough in my teaching that I would welcome being judged 100 percent by value-added." She would, that is, if she trusted the integrity of the tests on which the scores are based. And a recent national investigation seems to support her inclination not to. A March 2011 story in *USA Today* revealed that for the past three years, most of the classrooms at one particular school, Noyes Elementary, had an extraordinarily high number of erasures on the DC CAS, with a clear pattern of answers changed from wrong to right.[15] The story also noted that the number of students scoring at or above proficiency on the test increased from 10 to 58 percent in one year—a rate of increase far higher than the district average and virtually impossible statistically.

The findings of the investigation jibed with the experiences of this teacher and three of her colleagues, who also did not wish to be named. They told Education Sector of students whose test scores showed them to be proficient in reading or math in the grade before who suddenly were performing at a level of basic or below. The assumption was that the scores of these students in the previous year had somehow been inflated. Cheating, of course, significantly distorts the playing field; the teacher who fudges the numbers on students' tests is judged against the teacher who doesn't—and often comes out ahead. The teacher who gets the same students the following year is also hurt; because she is starting from an inflated baseline, she may not get credit for any growth she may have achieved.

Urging the public to "take a break" from the testing scandal, Kamras said that the questionable scores represented "only 2 percent of the data" and that "with that small amount, from a statistical standpoint, it doesn't throw off calculations in any material way"—meaning, among other things, that no teacher was fired as a result. Still, he said, "We take this very, very seriously. And if we find that improprieties led to a skewing, we will make modifications." In May, the district voided the test scores in the three Noyes classrooms. The D.C. inspector general continues an investigation. Meanwhile, the teachers' scores—and the IMPACT ratings on which they are based—stand.

The All-Important "Teach 2:" A Breakdown of the Rankings

Explaining content clearly, the second of the nine elements on the evaluation framework, is at the heart of good teaching. Here is what teachers generally demonstrate at each level.

Level 4: Highly Effective

Nearly all of the evidence listed under Level 3 is present, as well as some of the following:

- Explanations are concise, fully explaining concepts in as direct and efficient a manner as possible.
- The teacher effectively makes connections with other content areas, students' experiences and interest, or current events in order to make content relevant and build student understanding and interest.
- When appropriate, the teacher explains concepts in a way that actively involves students in the learning process, such as by facilitating opportunities for students to explain concepts to each other.
- Explanations provoke student interest in and excitement about the content.
- Students ask higher-order questions and make connections independently, demonstrating that they understand the content at a higher level.

Level 3: Effective

- Explanations of content are clear and coherent and build student understanding of content.
- The teacher uses developmentally appropriate language and explanations.
- The teacher gives clear, precise definitions and uses specific academic language as appropriate.
- The teacher emphasizes key points when necessary.
- When an explanation is not effectively leading students to understand the content, the teacher adjusts quickly and uses an alternative way to effectively explain the concept.
- Students ask relatively few clarifying questions because they understand the explanations. However, they may ask a number of extension questions because they are engaged in the content and eager to learn more about it.

Level 2: Minimally Effective

- Explanations are generally clear and coherent, with a few exceptions, but they may not be entirely effective in building student understanding of content.
- Some language and explanations may not be developmentally appropriate.
- The teacher may sometimes give definitions that are not completely clear or precise, or sometimes may not use academic language when it is appropriate to do so.
- The teacher may only sometimes emphasize key points when necessary so that students are sometimes unclear about the main ideas of the content.
- When an explanation is not effectively leading students to understand the concept, the teacher may sometimes move on or re-explain in the same way rather than provide an effective alternative explanation.
- Students may ask some clarifying questions showing that they are confused by the explanations.

Level 1: Ineffective

- Explanations may be unclear or incoherent, and they are generally ineffective in building student understanding of content.
- Much of the teacher's language may not be developmentally appropriate.
- The teacher may frequently give unclear or imprecise definitions or frequently may not use academic language when it is appropriate to do so.
- The teacher may rarely or never emphasize key points when necessary, such that students are often unclear about the main ideas of the content.
- The teacher may frequently adhere rigidly to the initial plan for explaining content even when it is clear that an explanation is not effectively leading students to understand the concept.
- Students may frequently ask clarifying questions showing they are confused by the explanations or students may be consistently frustrated or disengaged because of unclear explanations.

Source: District of Columbia Public Schools.

Development: The Missing Link?

IMPACT has three purposes: to outline clear performance expectations; provide clear feedback; and ensure that every teacher has a plan for getting better and receives guidance on how to do so. It is on this third goal that many teachers say IMPACT falls short.

In the conference that follows a classroom observation, the master educator explains to the teacher his scores, then offers concrete ideas on how he might improve. This sort of feedback came as a radical departure for Eric Bethel, a former elementary teacher at Marie Reed Learning Center who is now a master educator. He says he had never received instructional advice under the previous system, only a rating of "exceeds expectations"—a judgment that, however welcome, showed only how modest the expectations were. "I knew what excellence looked like," says Bethel.[16] "And in Montgomery County [the suburban district that adjoins D.C.], I don't even know that I could have kept my job." The master educator showed him, among other things, how he could use positive reinforcement to better control student behavior. "The observations allowed me to grow in very specific areas," he said.

As important, the master educator often serves to validate what the teacher is already doing, making a strong teacher even stronger. This is how it works when Radigan informally observes* Susan Haese, a first-grade teacher at Key Elementary School whom Radigan considers a "4." As Haese leads a small-group reading lesson, Radigan is frantically chronicling the event, filling up a grid with observations, quotes, and illustrations of teaching elements. Afterward, he tells her, "I want to celebrate what you did and repeat it." He gives her a "3" on Teach 1 because he's not convinced the students entirely understand her objective. "I hear ya," she says. But he gives her specific tips for building reading fluency, including having the students first read to themselves to build meaning, then read aloud as if they are on the radio. "I like that," says Haese enthusiastically. "I can have them talk into paper towel holders as microphones."

* Informal evaluation for feedback only.

But while this kind of advice is constructive, and while it certainly improves upon past practice, it is also limited. That's because, as Robinson-Rivers describes it, the job of the master educator is "80 percent evaluative and [only] 20 percent developmental." Radigan says administrators made it clear that they were not looking for instructional coaches when they hired master educators; each school already has at least one educator filling that role. Yet

One barrier to better development, both sides agree, is that, according to the union contract, the master educators may not share evaluations with instructional coaches, the teachers who work with their peers to help them improve their craft.

teachers, appreciative as they may be of the post-observation feedback, consistently say they want a stronger connection between support and evaluation. Specifically, they have asked for mentoring, along with actual demonstrations of precisely what is expected of them in the classroom. At the least, many say the district should not have held them to the teaching and learning standards without first giving them the full support they needed to meet them.

It's a familiar chicken-and-egg argument. But district officials were very deliberate in changing the protocol so that it is now up to the teachers to get themselves the help they need instead of making the principal responsible for providing it. "There is a shift," Thompson confirms. "Now we see the teacher as taking a more active role." The district calls this philosophy "empowerment." The teachers call it "sink or swim."

One barrier to better development, both sides agree, is that, according to the union contract, the master educators may not share evaluations with

instructional coaches, the teachers who work with their peers to help them improve their craft. Thus the coaches are deprived of some of the very data they need to diagnose areas targeted as weak spots. "It makes it hard for me to know where in the rubric they are falling short," says Barron. (The coaches, who fall in the category of teachers, come under the contract; the master educators, who work for the administration, do not.) There is nothing to prevent the teacher from sharing her IMPACT scores with the coach, of course, but the coach cannot ask her to, and many are reluctant to do so on their own. "Some of them are embarrassed to tell me," says Barron. "The whole psychology of this is so important. It's just as important for teachers as it is for kids."

From a policy standpoint, instead of spending valuable time that would best be directed to more promising instructors, it might be preferable to let this teacher sink and get fired.

This arrangement, which Thompson concedes is "not optimal," holds consequences for the instructional coaches, as well. As with principals (and custodians and administrative assistants) the coaches are subject to their own rubric, and 30 percent of their score is based on the professional growth of the teachers under their tutelage. Without the IMPACT data, that growth—at least as measured by the rubric—is harder to achieve. And there is the flip side. Take the case of a genuinely poor teacher who is appropriately rated minimally effective on all counts. A good coach may know that she is a lost cause. From a policy standpoint, instead of spending valuable time that would best be directed to more promising instructors, it might be preferable to let this teacher sink and get fired. That would be a good outcome, but it would count against a coach's score. "It's a game of the numbers now," says Barron.

Those numbers also translate into dollars, and, as with other aspects of IMPACT, the compensation system has brought some interesting, if not entirely unexpected, results. To be eligible for salary bonuses, teachers had to give up some protections and choices in the case they were "excessed," due to declining enrollment, for instance. It is hardly an academic question. In May, 384 teachers, librarians, and counselors were notified that they were losing their jobs because of school closings, budget cuts, and other factors.

One teacher who was willing to make the tradeoff—money in exchange for security—was Bethel. "I was good," he says, "but I knew what excellence looked like, and I thought I needed to raise my game." The money was not insignificant. Rated highly effective, and awarded extra points for teaching a high-need subject in a low-income neighborhood, Bethel earned a bonus of amounting to nearly 40 percent of his regular salary and plans to use it for a down-payment on a house. In the end, though, according to figures from DCPS, only 60 percent of eligible teachers last year proved willing to waive this protection, and it took more and more money to entice them. Nine of the 12 teachers who were eligible for $20,000 awards (75 percent) accepted the bonus, but only 57 percent accepted awards when they were less than $10,000. The maximum bonus a teacher can get is $25,000, for being highly effective and teaching a high-need subject (like high school physics), in a testing grade, in a high-poverty school. Two teachers were eligible for the top bonus last year, and both accepted it.

This pattern seems to be saying something about teacher motivation, and it suggests one more area for the district to study. To what degree are teachers motivated by money? Why ask the good teachers to give up job security? If these teachers are that good, and if their school is closed, wouldn't the district want to find a way for them to practice their craft elsewhere? Kamras says that district officials were not at all surprised by the number of teachers who turned down the bonuses. "Look, inherent in this whole thing is the opportunity to choose, and to guide your own career…you can get north of $130,000 in 10 years. But if accountability is not a good deal for you, it's your choice, and I completely respect that." Besides, Kamras says, "A lot of teachers didn't think we were actually going to pay."

Toward a Better IMPACT

Even as teachers await their final scores for the school year that's drawing to a close, IMPACT administrators are waiting for a report on the system's implementation by an independent consultant group. The report is expected to make new recommendations for changes to the system. Washington, D.C.'s new mayor, who campaigned against some aspects of IMPACT and won the support of the teachers union, says that more improvements are needed. "[IMPACT] is a step in the right direction," the mayor, Vincent Gray, recently told a group of constituents, "but it has a long way to go to be a fair evaluator of our teachers."[17]

To ensure objectivity and consistency, teachers and others have suggested some of the following changes:

1. Making the master educator observations longer or extending them over a few days in the same week.
2. Having teachers write an evaluation of their own classroom performance.
3. Meeting with the teacher prior to the evaluation so that the master educator can learn about any special issues with the class.
4. Taking better account of difficult classroom situations.
5. Making sure that master educators and school administrators are grading the same way.

Many teachers also say they want evaluators to calculate the value they add over more than one school year.

Thompson says the district is "committed to making the changes that are necessary," but after already making substantial adjustments this year, he doesn't expect large-scale changes in the next. "Teachers need time to get comfortable and develop mastery of the rubric," he says. Besides, Kamras says of the revised rubric, "I think we have pretty much hit the sweet spot." Instead, the district's big push next year will be connecting evaluation to development, as well as providing teachers with better academic and curricular support. Among other tools, the district is producing an online video library it calls "Reality P.D."—more than 120 clips of DCPS teachers demonstrating various aspects of the rubric and sharing their tips.

The district is also starting to use data generated by IMPACT to improve instruction. In the first year, teachers districtwide consistently scored lowest on measures of rigor and probing for higher-level understanding. That finding led the district to further clarify and emphasize these skills in the revised framework and in professional development. The information drives improvements at individual schools, as well. Reviewing a spreadsheet that helpfully breaks down scores by teacher and by each element of the rubric, Dwan Jordon, the principal at Sousa Middle School, noticed that his teachers scored lowest in Teach 2—delivering content clearly—and, as with the district overall, in Teach 7—probing for higher understanding. So he and his fellow administrators went into action, collaborating on a PowerPoint presentation called "How to Get a 4 on IMPACT." As a result, he says, two teachers who had been rated minimally effective boosted their scores to 3.75 and 3.89 respectively.[18]

IMPACT may be an imperfect measuring tool, but, as many experts see it, it may be the best one out there right now.

As to IMPACT improvements down the road, Kamras says the district is "seriously looking into" student evaluations of teachers because new research sponsored by the Bill & Melinda Gates Foundation has shown that pupils themselves are remarkably good judges of effective instruction.[19] Also being considered are ways for teachers to submit assessments of themselves, although Kamras says such evaluations would not likely factor heavily into an overall score. Finally, as IMPACT enters its third year, Kamras says he is determined to calm teachers' fears. "There is still a perception that IMPACT is a 'gotcha,'" he says. "But I think the big thing has been getting over the hump. We went from zero accountability right to 100 percent accountability. So without changing the fundamentals, I want to reduce the anxiety level."

IMPACT may be an imperfect measuring tool, but, as many experts see it, it may be the best one out there right now. It is the product of a desperate problem crying out for an immediate, dramatic solution—a solution that DCPS says couldn't wait to be piloted. The net may drag in teachers who didn't deserve to be caught. But district administrators, along with a fed-up public, have essentially decided that it's better that one teacher lose her job unfairly than many bad ones undeservedly keep theirs. "If teachers are anxious because they have low scores, I empathize," says Kamras, "but at the end of the day, we have to hold the line on quality. I believe with every fiber of my being that we can't have different standards for other people's children than we have for our own." Evaluation has raised those standards. Thus, it's no longer a question of whether teachers will be judged by an intensive system of test scores and classroom observation—only how.

Notes

1. Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling, *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (Brooklyn, NY: The New Teacher Project, 2009).
2. Rachel Curtis, *District of Columbia Public Schools: Defining Instructional Expectations and Aligning Accountability and Support* (Washington, DC: The Aspen Institute Education and Society Program, 2011).
3. Rachel Curtis, *District of Columbia Public Schools: Defining Instructional Expectations and Aligning Accountability and Support.*
4. Scott Thompson, in discussion with author, Spring 2011.
5. Jack Gillum and Marisol Bello, "When Standardized Test Scores Soared in D.C., Were the Gains Real?" *USA Today,* March 30, 2011.
6. Bill Turque, "Gray: IMPACT Teacher Evaluation System Has 'a Long Way To Go' for Fairness." D.C. Schools Insider blog, *Washington Post,* Jan. 17, 2011.
7. Nathan Saunders, in discussion with author, Jan.–Feb. 2011.
8. Cynthia Robinson-Rivers, in discussion with author, Jan.–Feb. 2011.
9. Matt Radigan, in discussion with author, Jan.–Feb. 2011.
10. Bill Rope, in discussion with author, Jan.–Feb. 2011.
11. Marni Barron, in discussion with author, Spring 2011.
12. Jason Kamras, in discussion with author, Spring 2011.
13. Eric Isenberg and Heinrich Hock, *Measuring School and Teacher Value Added for IMPACT and TEAM in D.C.* (Washington, DC: Mathematica Policy Research, Inc. August 20, 2010).
14. Aaron Pallas, in discussion with author, 2011.
15. Jack Gillum and Marisol Bello, "When Standardized Test Scores Soared in D.C., Were the Gains Real?"
16. Eric Bethel, in discussion with author, Spring 2011.
17. Bill Turque, "D.C. Mayor Offers Most Explicit Criticism of IMPACT Teacher Evaluation System," *Washington Post,* Jan. 18, 2011.
18. Dwan Jordon, in discussion with author, Spring 2011.
19. Education Sector receives funding from the Gates Foundation, but the findings in this report are those of the author alone.