Carnegie Foundation
for the Advancement of Teaching

DECEMBER
2014

ISSUE BRIEF

Adding Eyes: The Rise, Rewards, and Risks of Multi-Rater Teacher Observation Systems

BY TAYLOR WHITE

NEW TEACHER EVALUATION SYSTEMS have emerged as the cornerstone of the recent movement to improve public school teaching. Fueled by incentives from the federal government, state and local policymakers have sought to replace the often-cursory evaluation models of the past with more comprehensive ones. In contrast to past evaluations, which often relied on a single classroom visit by an untrained administrator, new models evaluate teachers on the basis of their students' achievement, on surveys that capture students' perceptions of their teachers' practice, and on improved classroom observations.

The inclusion of student-performance measures has been highly controversial. But research conducted by the Brookings Institution in four urban districts found that only 22 percent of teachers had test-score gains factored into their evaluations.[1] In these districts and elsewhere, observations of teachers' work in their classrooms continue to generate the majority of the performance information under the new evaluation systems. Policymakers have sought to improve upon traditional observation systems by tying observation to rigorous rubrics, intensifying observer training, and increasing the number of required observations. Between 2011 and 2013 alone, the number of states requiring teachers to undergo multiple observations each year increased from under 10 to 25.[2]

But as these new systems roll out, there is mounting evidence that principals alone cannot bear the time burden they impose. Nor can a single principal be depended upon to deliver effective feedback across content areas to teachers with vastly different strengths, weaknesses, and teaching assignments. In response to these challenges, a growing number of districts have adopted multi-rater systems, in which several observers watch teachers at work, score their performance, and provide feedback. Sometimes the raters observe together, sometimes independently. And more and more, they come to the process from different vantage points: Many districts now rely on combinations of peer teachers, master teachers, and administrators from different schools. By adding more eyes to these evaluations, districts aim not only to relieve principals but, more important, to lend new perspectives, deeper expertise, and greater objectivity to the evaluation process.

> “As these new systems roll out, there is mounting evidence that principals alone cannot bear the time burden they impose. Nor can a single principal be depended upon to deliver effective feedback across content areas to teachers with vastly different strengths, weaknesses, and teaching assignments.”

This report explores the use of multi-rater evaluation systems in 16 districts with widely varying student populations, resources, and policy priorities. The districts range from New York City, the nation's largest school system, to Transylvania County, NC, which educates just 3,500 students each year. Drawing on document reviews and interviews with district officials, it examines the districts' varying aspirations for multi-rater models, as well as how the models are designed, how they operate, and the challenges they pose. This report is not intended to be a technical assessment of these systems, nor does it take into account processes related to non-scored observation, such as observation by coaches that provide only formative feedback. Although this brief does not explore the training of raters in depth, training is an increasingly important topic with significant implications for the quality and equity of evaluation processes both within and across districts. Rater training will be the topic of a future Carnegie brief.

The Problems with Observing Teaching

It is increasingly clear from research and practice that relying solely on principals to conduct classroom observations is problematic. Time demands, for one thing, present a substantial burden to principals. After the first year of district-wide implementation of Chicago's new evaluation system, for instance, 66 percent of administrators said that the system's increased observation requirement took too big a chunk of their already overscheduled time.[3] Administrators in Tennessee felt similarly after the first year of that state's evaluation system. After conducting nearly 300,000 classroom observations, school administrators declared the time burden “unmanageable.”[4,5] Says Vince Botta, a former principal who is now the director of performance management in Georgia's Gwinnett County Public Schools: “There is no way as a principal you can do it all alone.”[6]

Principals who do try to “do it all,” research suggests, may end up hurting the process. A 2013 study from the Consortium on Chicago School Research observed that the heavier workloads caused by the demand for teacher evaluation have “contributed to lower [than required] engagement in the new system for some principals.”[7] More specifically, researchers from Stanford University find that some overburdened administrators "cut corners...typically doing fewer [evaluations] than desired or even required."[8] While this behavior is understandable given the responsibilities administrators juggle, it could significantly undermine the improvement aims of evaluation by cutting the time a principal needs to conduct high-quality observations and to prepare and deliver targeted, actionable feedback.

> "Teachers who believed that administrators did not spend enough time in their classrooms questioned the validity of their evaluators' assessment, doubted that the appraisal reflected an understanding of their daily work, and complained that the evaluation process lacked credibility."

Quick, cursory observations are also likely to damage teachers' trust in the evaluation system, further undermining efforts to improve their instruction. In a study of six urban schools, Harvard University researcher Stefanie Reinhorn found that "teachers who believed that administrators did not spend enough time in their classrooms questioned the validity of their evaluators' assessment, doubted that the appraisal reflected an understanding of their daily work, and complained that the evaluation process lacked credibility."[9] And even if evaluators do spend enough time in teachers' classrooms, Reinhorn found, teachers felt strongly that "their evaluators need to have [content knowledge and pedagogical knowledge] in order to validly assess their practice, provide valuable feedback, and support individual teachers' growth." In other words, they needed expertise that principals can't be expected to have across a full range of grade levels and subjects.[10]

Although there were some skilled evaluators in Reinhorn's study, additional research suggests that these evaluators have been the exceptions rather than the norm. According to a research synthesis done by Heather Hill of Harvard University and Pamela Grossman of Stanford University (now of the University of Pennsylvania), "many principals lack the knowledge and expertise to provide content-specific feedback," especially in math.[11]

A Potential Solution: Multiple Raters

Multi-rater systems show promise in addressing these challenges. For one thing, they are designed to give teachers more time with more evaluators. If calibrated carefully, says Reinhorn, multi-rater systems can "address the shortage of time or specialized expertise" that lead to teachers' concerns.[12]

A growing body of research suggests that multi-rater designs might improve the reliability of evaluation scores, too. The Measures of Effective Teaching project, funded by the Bill & Melinda Gates Foundation, found that "if a school district is going to pay the cost (both in money and time) to observe two lessons for each teacher it gets greater reliability when each lesson is observed by a different person."[13] Likewise, a 2014 Mathematica study on Pittsburgh's evaluation system found that ratings might be improved if more than one observer rated each teacher. This finding was due partly to inconsistencies in how principals applied the rating rubric.[14] And when systems match classroom teachers with observers who are experts in the same grade or subject, the result may be more nuanced assessments of teachers' practice, as well as more precise feedback to drive improvement—not only in teachers' practice, but in principals' evaluative and feedback skills as well.[15]

> “When systems match classroom teachers with observers who are experts in the same grade or subject, the result may be more nuanced assessments of teachers' practice, as well as more precise feedback to drive improvement.”

Consistent with the literature examining multi-rater systems, nearly every district in our review cited reduced workloads for principals as a key reason for incorporating additional raters. Most also believed that adding observations and observers would produce more and more varied information on which to evaluate a teacher's practice. And several expressed hope that this better data, paired with added content or grade-level expertise, would lead to more accurate evaluation scores and richer, more tailored feedback. Despite broadly consistent goals for their multi-rater approaches, the districts' varying budgets, populations, and local and state contexts yield a wide range of designs. In general, the systems vary along three key dimensions: teacher type (which teachers must be scored by multiple raters); rater type (who can serve as raters); and rater role (how raters share responsibilities for observation).

What Teachers Get Rated?

Districts that use multiple raters must decide which teachers or subgroups of teachers will be observed by multiple raters. This decision is based on two factors: what the district hopes to accomplish by using multiple raters and what resources it has to support the design. Districts primarily concerned with improving the precision and validity of scoring in high-stakes cases might require multiple raters only for those teachers on track to receive the highest or lowest scores—teachers who may be headed for recognition or sanction. This is the policy, for instance, in New Haven, CT. Requiring that all teachers be observed by multiple raters is a more resource-intensive approach, because it generally requires training and compensating additional raters and more careful coordination than a targeted design. Nonetheless, five of the 16 districts—Desoto Parish, LA;[16] Eagle County, CO; Santa Fe, NM; Maricopa County, AZ;[17] and Hillsborough County, FL—have taken this approach, many citing that it is necessary to build a common culture around evaluation and professional improvement. However, because of resource concerns, some of these districts plan to eventually differentiate how often teachers are observed. The District of Columbia Public Schools (DCPS), whose IMPACT evaluation system initially required that all teachers be observed by a combination of master educators and school administrators, has already taken this step. Teachers who have earned five consecutive years of “highly effective” or “effective” ratings—expert teachers—are now observed only by administrators. This arrangement allows the district to vary the intensity of observation and support teachers receive.[18]

As with New Haven, five other districts—Transylvania County, Greenville County, Greene County, Baltimore City, and New York City—use multiple raters only for novices or teachers who have earned very low performance ratings in the past. And in Greenville County, in accordance with state law, both probationary teachers and teachers at risk for termination must be observed by teams of evaluators. These strategies direct the benefits of multiple raters to those teachers most in need of supervision and support.

Four districts allow schools to use multiple raters but do not require they be used generally or for any particular subgroup of teachers. Most of these districts have determined that requiring the practice would be unfeasible due to logistical challenges, particularly at schools with just one administrator or with administrators who oversee multiple sites.[19] Officials in each of these four districts believe some schools have adopted multi-rater approaches, but none could say for certain. Jana Burk, executive director of Tulsa's Teacher/Leader Effectiveness Initiative, says that "it's only when really high-caliber principals are in charge [that they] find time to make it happen."

Who Does the Rating?

Another design consideration for districts is the question of who should serve as raters. Some districts use only administrators, both district- and school-based; other districts use administrators and other expert raters, such as master teachers and mentor teachers. Others use administrators and teachers' peers.[20]

More than half of the districts rely solely on administrators to serve as raters. For several, this approach is the only option since state requirements or local bargaining agreements restrict high-stakes observations to people at this level. Although teachers in these districts are usually observed primarily by their principals or assistant principals, some districts also tap central office staffers or others with administrative credentials. This arrangement is often the case in districts where schools may have only one on-site administrator.

> "Some districts use only administrators, both district- and school-based; other districts use administrators and other expert raters, such as master teachers and mentor teachers. Others use administrators and teachers' peers."

Six districts use distinguished teachers—called master teachers, mentor teachers, peers, or validators—to rate teachers and (in some cases) provide feedback, supplementing administrators' observations. Districts varied in their approaches to recruiting these expert raters. While some, such as DeSoto Parish, Eagle County, and Hillsborough County, hire most or all of their teacher-leaders from within the district, others prefer to recruit from the outside. New Haven Public Schools, for example, employs a cadre of evaluators called "third-party validators," most of whom have worked in various capacities in other nearby districts, but none of whom are teachers or administrators in the New Haven district. This approach ensures that the third-party validators are objective raters (they do not provide feedback), an important quality given that their primary role is to validate principals' scores for the highest and lowest performers. DC recruits both locally and nationally for its master educators, a strategy that officials say allows the district to hire the best talent.

> "She and her team filled 41 peer-evaluator spots with educators who had deep expertise in specific grades and content areas. These peers ensure not only that all teachers receive ratings and feedback from relevant experts, but also that principals have the resources to improve their own observation and feedback skills."

While relying on expert raters to alleviate administrators' workloads, districts also depend on them to provide content and grade-level expertise that administrators may lack. Lori Renfro, who directs the new teacher evaluation system used by several districts in Maricopa County, knew that with higher expectations for feedback, "principals would need help with content."[21] With this in mind, she and her team filled 41 peer-evaluator spots with educators who had deep expertise in specific grades and content areas. These peers ensure not only that all teachers—even music and physical education teachers—receive ratings and feedback from relevant experts, but also that principals have the resources to improve their own observation and feedback skills.

Only two districts, Greenville County and Transylvania County, include actual peers—full-time classroom teachers—among raters of colleagues' practice. State policies in both North Carolina and South Carolina require that peers help evaluate provisional early career teachers. Both states require that peer observers undergo district-provided training, but they do not hold special contracts or take on other leadership duties as is common for master teachers elsewhere. As do many districts with more formal master-teaching positions, both Greenville County and Transylvania County aim whenever possible to pair teachers with peer evaluators who teach in the same grade or content areas.

How Are Raters Deployed?

Even when districts have similar policies about who can rate teachers, they often deploy raters differently. In several of the districts where rating responsibilities are restricted to administrators, building-level teams are free to divide and share the observation duties as they see fit. Several districts—including Transylvania County and New York City, the report's smallest and largest districts, respectively—provide this flexibility.

A handful of other districts provide more specific guidance on administrators' observational roles. In Santa Fe, for example, all teachers are now observed by their own principals or assistant principals and by an administrator from another school within the district; teachers rated minimally effective are also observed by the assistant superintendent. Almi Abeyta, the district's chief academic officer, reports that the new model, in just its second year, has improved calibration between principals and assistant principals. And when there have been questions about a teacher's performance, it has often confirmed what the school-based administrator had initially observed.[22]

Similarly, in Boston Public Schools, administrators serve as primary evaluators for some teachers and secondary evaluators for others. Teachers are assigned a primary evaluator, but they might also have multiple administrators serving as secondary evaluators so that school leadership teams can establish trends in teachers' performance over time. While both types can conduct observations and submit evidence to teachers' online portfolios, raters score evidence and provide a final performance rating only for those teachers in their primary caseload. Angela Rubenstein, a district implementation specialist, says this division of labor "distributes workload so caseloads are more manageable and so teachers get a lot of opportunities to generate data for their evaluation." It also ensures that evaluators "have enough interaction with teachers to provide meaningful feedback for development," she says.[23]

> "Teachers are assigned a primary evaluator, but they might also have multiple administrators serving as secondary evaluators so that school leadership teams can establish trends in teachers' performance over time."

Greene County, a large rural school system, was the only district in which administrators frequently observed classrooms in tandem. In a pilot project supported by the state's federal Race to the Top grant, principals and assistant principals co-observe with counterparts from similar schools within the district. They rate lessons separately, then meet to reach a consensus on scores and feedback. Though currently limited only to the district's lowest-performing teachers, the pilot has provided three key benefits, Greene county officials say: greater objectivity in scoring, broader expertise in providing feedback, and professional development for principals, who, according to Superintendent Vicki Kirk, have "really, really enjoyed breaking down [teachers'] lessons and norming their interpretations of the rubric with colleagues."[24]

Co-observation also occurred in districts in which master, mentor, or peer teachers serve as expert peer raters, although the practice is required only in New Haven and New York City, where some teachers are observed by both their principals and external validators. In New York City, the policy applies only to the lowest performing teachers; New Haven uses validators to co-observe both its lowest- and highest-scoring teachers.

Most other districts using expert peer raters simply divide observation responsibilities among combinations of raters, with master teachers often shouldering half or more of the observation and feedback requirements for each teacher.

The workloads of the expert raters depend largely on districts' needs and the nature of their contracts. In DeSoto Parish, DC, Hillsborough County, and Maricopa County, master teachers (DeSoto and DC) or peers (Hillsborough and Maricopa) devote all of their time to observing, rating, coaching, and leading professional development sessions for teachers and sometimes principals. In Eagle County, on the other hand, master teachers maintain about 30 percent of their teaching duties, devoting the other 70 percent to evaluation and staff development, with each carrying a caseload of about 20 teachers. Eagle's 20 master teachers work under 201-day contracts—20 days longer than the contracts held by the district's classroom teachers.[25]

Both Eagle County and DeSoto Parish also employ mentor teachers who perform many of the same duties as masters (doing observations, providing feedback, etc.) but carry heavier teaching loads and devote less time to observation and feedback responsibilities.

Transylvania County, one of the two districts that include peer raters in their system, also distributes responsibilities across raters, but not quite so evenly. Administrators in Transylvania conduct three observations for each teacher up for promotion to career contract status; peers contribute just one.

Challenges to Using Multi-Rater Systems

In making decisions about how to design multi-rater systems, districts had to consider their goals for evaluation, the capacity of their existing systems, and the resources available to improve or expand them. Many districts faced tradeoffs: adding raters may bring much-needed expertise and reduced workloads for principals, but such improvements were often costly and frequently introduced new challenges.

Rater Reliability

An especially big problem for districts is how to train observers to properly rate classroom practice. Training is important both for the technical challenge of making sure that ratings are consistent, common, precise, and reliable and for the significance that accurate ratings hold for an entire school system.

It is essential, first, that each rater judge every teacher by the same standard—that a score of 5, for instance, represents the same caliber of instruction each time it is applied. At the same time, when several raters are judging, they must all agree on what each of the rating levels looks like. To make sure their ratings are reliable in this way, they must meet several times a year to recalibrate their scoring. Ensuring this kind of quality is expensive, often more costly than the initial training.

At DCPS, Stephanie Aberger, the director of IMPACT's training platform, says that robust introductory training includes about 10 hours of preliminary work online and another 10 hours of in-person support. "This is a significant time commitment," Aberger says, "especially for a busy new school leader." The work is funded with a $2.2 million dollar grant from the Bill & Melinda Gates Foundation.

> “It is essential that each rater judge every teacher by the same standard—that a score of 5, for instance, represents the same caliber of instruction each time it is applied.”

Yet thorough and repeated training is essential to the process of accurately identifying good teaching and to building a shared understanding of performance. Just as previous "drive-by" teacher evaluations were essentially meaningless, and signaled as much to the teaching profession, the new generation of evaluations can help build a shared sense of what good teaching looks like in a district—or breed mistrust when scores for comparable teachers are radically different. Teachers lose faith in the system when scores are obviously out of sync. Administrators, when guided by inaccurate ratings, can make poor judgments, sometimes in high-stakes cases. Training raters effectively, and recalibratig them periodically to ensure the quality of ratings, is an important topic that Carnegie will address in an upcoming brief.

Financial Cost

The cost of multi-rater systems depends on the specifics of their design. Hillsborough County, a district that increased both the frequency of observation and the number of eligible raters, spent $11.9 million on its evaluation system in 2011-2012.[26] More than 85 percent of that total—nearly $10.4 million—went to support its multi-rater observation model, according to a 2013 report from the RAND Corporation and the American Institutes for Research.

Costs associated with the four key components of Hillsborough's observation model:

Design and implementation *Includes cost of training and calibrating raters*	$525,580
Peer and mentor observers *Includes salaries and benefits for 189 peer and mentor teachers and the teachers hired to take over their full-time teaching loads*	$8,122,558
Management and communications *Includes salaries for central office staff managing the system and fees to contractors for communications and management support*	$316,740
Technology and data systems *Includes an online platform to house and aggregate data on teachers' performance and laptops for peer and mentor observers*	$1,432,988
Total costs associated with Hillsborough observation in 2011-2012	$10,396,865[27]

Although this sum may be a relative drop in the bucket of Hillsborough's total annual operating budget of $2.3 billion, it's worth noting that nearly two thirds of the $24.8 million Hillsborough spent on its evaluation system from 2009 to 2012 came from external grant funding.[28]

Hillsborough is not alone in this regard. Significantly, Eagle County, DC, Maricopa County, and others received significant external support from state Race to the Top awards, federal Teacher Incentive Fund grants, and private foundations to design, launch, and initially sustain their intensive and costly multi-rater systems. Most of the remaining districts relied primarily on local and state funding to get their less-intensive systems up and running.

Perhaps because of these funding differences, most districts pursued less resource-intensive designs than did Hillsborough or DC—particularly models that use multiple raters for only lowest-performing teachers or those that recommend rather than require multiple raters. Only one district—Santa Fe—implemented a multi-rater requirement for all teachers without increasing the number of raters in the district. The district will employ two retired principals in 2014-2015 to help cover increased observation duties when it moves to a more differentiated system.[29] But several districts—Santa Fe included—said they would reconsider elements of their designs if they had more money to train and compensate more raters.

Staffing Considerations

Even if resources exist to employ additional raters, many districts have little opportunity to do so, since state policies or local contractual language sometimes restricts evaluation tasks to licensed administrators.

In Chicago, an effort to allow department chairs to serve as evaluators was rejected by the Chicago Teachers Union which, like unions elsewhere, raised concerns about allowing teachers from the same collective bargaining unit to evaluate one another.[30] DC, which faces similar contractual restrictions, was able to use its master educators as raters by including them in the local administrators' union.

Districts with greater flexibility still had to weigh the pros and cons of engaging non-administrative raters in their evaluation processes. Though they can expand the system's bandwidth and expertise, non-administrative raters—whether they are external raters, master teachers, or peers—are not always welcomed by principals and teachers, who may doubt their motives and value. When Maricopa County introduced expert peer evaluators, teachers and principals experienced an "initial shock," says Lori Renfro; they did not immediately trust the peers or their motives in the evaluation process.[31] To foster more personal relationships between educators and the peer evaluators, many Maricopa County peer evaluators now meet with principals and teachers before the evaluation cycles begin in informal, one-on-one meetings and school-level "meet and greets." And, to ensure that peers aren't spread too thinly across the district, Maricopa has reconfigured their assignments, giving them more time to spend on fewer campuses. Maricopa officials hope the move will build stronger relationships between the roving peers and school-based educators.

Though Maricopa, DC, New Haven, and other districts hire non-administrative evaluators primarily from outside the district (an expensive process in itself), many other districts hire master or mentor teachers from within—both to thwart trust issues associated with "external" raters and to provide meaningful leadership opportunities for their own high-performing teachers. But hiring raters from within may also introduce complex interpersonal dynamics, as teacher leaders adjust to their new authority and redefine relationships with former colleagues.[32]

> “Hiring raters from within may also introduce complex interpersonal dynamics, as teacher leaders adjust to their new authority and redefine relationships with former colleagues.”

Logistics and Coherence

Adding new raters—whether newly certified assistant principals, external raters, or master teachers—requires more careful coordination, both between the district and schools and within schools themselves. Systems must be in place to schedule and monitor classroom visits, to collect and store observation data from multiple raters, and, most important, to provide time for raters to calibrate their scoring practices and ensure they are providing consistent, coherent feedback.

Implementing such systems, however, is often easier said than done. Building and maintaining databases to house observation data can be time-consuming, as well as costly, as evidenced by Hillsborough's budget on page 9. And it can be difficult in a busy school week to carve out time for raters to discuss their scores and align feedback with each other.

The districts cited these challenges almost unanimously, but several have found ways to mitigate their effects. They now rely on a range of online applications to aggregate and store evaluation data from various raters. Many districts allow teachers to access these systems and upload artifacts of their teaching practice, making the process more transparent and interactive than ever before.

Several districts have also hired full-time staff to help manage their evaluation processes. Along with Patty Fox, who coordinates Greenville County's evaluation system, the district also employs seven full-time lead teachers who, in addition to serving as raters on the district's evaluation teams, schedule visits for their teams, collect and input data from team members, and generate official reports. Based in the district's central office, these lead teachers handle caseloads of about 50 teachers at a time and, according to Fox, have been "absolutely huge" in making the process manageable for school leaders. Similar positions existed in several of the districts, but smaller districts employ fewer full-time staff in such roles and seemed to rely more heavily on principals to coordinate the process.

With multiple raters, it is important for districts to ensure that raters are carefully calibrated, particularly as some districts found that non-principal raters produced lower scores, on average, than principals (a finding consistent with several recent studies). In two districts in which this difference was most pronounced—both districts in which expert peers serve in full-time evaluation roles—officials attributed the pattern largely to the fact that expert peer raters were hand-selected for their roles and deployed exclusively to evaluate and provide feedback (and to perform accompanying reporting duties). Many administrators, on the other hand, were not hired for their ability to rate teachers' practice; some were even reluctant to do so, either because they lacked confidence in their evaluative skills or simply because they were skeptical of the process itself. And all had other administrative duties occupying their time.

A small number of districts (including one of those with large gaps in administrator and expert peers' scores) also noted that expert raters undergo more intensive training than principals, in the hope that they will coach and assist the principals where necessary. Though this strategy may conserve resources, it also raises concerns about inadequate and inconsistent training. Raters who do not receive proper training and support may introduce issues of reliability into evaluation scores, compromising the quality of data used to make judgments about teachers' practice, even threatening their employment. Such issues could plague any evaluation system, but multi-rater systems that intentionally apply different training standards to raters may be particularly vulnerable. In practice, the strategy also received mixed reviews: though some districts using this approach found it beneficial for improving administrators' abilities and the collegiality between the raters, others noted that it created tension, as when principals felt they were being "policed" unfairly by other, more highly-trained raters.

A small number of districts mentioned deliberate efforts to track and coordinate the feedback teachers receive from various raters, but few had clear, established systems to ensure that feedback was consistent and coherent. DeSoto Parish, a standout in this area, has worked hard to define raters' various roles and enlists master teachers to coordinate the information and support teachers receive from principals, mentor teachers, and one another. Kathy Noel, DeSoto Parish's director of student learning, says the master teachers are "the maestros of the whole [evaluation and support] system."[33]

"A small number of districts mentioned deliberate efforts to track and coordinate the feedback teachers receive from various raters, but few had clear, established systems to ensure that feedback was consistent and coherent."

In some cases, districts' attempts to better coordinate feedback may be hindered by local policy. In DC, for example, an information "firewall" prevents the master educators and the district's fleet of instructional coaches from freely sharing any details about their interactions with specific teachers. Put in place to allay concerns from the local union (such as members of the administrative union sharing evaluation data with non-administrative coaches), the firewall makes it challenging to ensure that teachers receive consistent feedback from the various instructional leaders. This is the case even though the district has taken steps to assess and improve coherence among various feedback providers.

A handful of other districts have sought to coordinate feedback by allowing various raters to exchange information through an online evaluation portal. It is unclear, however, how often raters in each district use this function, or whether other feedback providers (such as coaches) can access these portals or input their own information. Without that functionality, or other efforts to coordinate between the various providers, teachers are likely to be overwhelmed by the amount and range of support they receive. And this outcome may have the unintended effect of undermining systems' aims for instructional improvement.

Conclusion

Though it is too early to determine the full impact of multi-rater evaluation systems, it is clear that districts have high hopes for their ability to gather better information, produce more accurate, objective ratings, and improve the quality of feedback teachers receive, all while reducing principals' workloads. Emerging research suggests that many of these expectations are well-warranted.

Districts' unique aspirations for their systems, along with their policy contexts, led to designs that varied along three main dimensions: the type of teachers evaluated by multiple-raters, the type of raters used, and how raters are deployed. Districts also faced certain common challenges in implementing their designs. Chief among these were issues related to cost, raters' capacity, and system coherence.

Perhaps as a result of these challenges, multi-rater systems are not yet commonplace in the nation's schools.[34] Only four states—North Carolina, South Carolina, New Jersey, and Maryland—require that teachers be observed by multiple raters, but these requirements apply only to novices or the lowest-performing teachers. In one survey of middle school math teachers in Missouri, fewer than 23 percent reported having been rated by more than one observer.[35] But, given a recent assertion by the Brookings Institution that "nearly all the opportunities for improvement to teacher evaluation systems are in the area of classroom observations rather than test score gains," it seems likely that policymakers and district officials will soon double down on efforts to strengthen observation processes.[36] And when they do, multi-rater systems stand to become a prevalent strategy for improvement.

■■■

Taylor White is a former associate for public policy engagement at the Carnegie Foundation for the Advancement of Teaching. She now serves as the deputy director of education policy and university research at the Australian Embassy in Washington, DC.

	Transylvania County, NC	DeSoto Parish, LA	Eagle County, CO	Greene County, TN	Santa Fe, NM	Maricopa County, AZ (REIL Only) (2)	New Haven, CT
DISTRICT DATA (1)							
Students	3,557	5,040	6,344	7,467	14,071	48,347	20,554
Schools	9	13	21	18	30	100	49
Teachers	261	379	429	470	949	2,420	1,615
EVALUATION SYSTEM INFORMATION							
System Name	NC Educator Effectiveness System	TAP	Professional Excellence, Appraisal, and Recognition (PEAR)	Tennesee Educator Acceleration Model (TEAM)	NM TEACH	REIL	TEVAL
Implementation Date	2011-2012	2008-2009	2009-2010	2011-2012	2013-2014	2010-2011	2009-2010
Components of Rating	Teachers scored by principal on 5 domains of practice. Sixth component is based on student achievement.	Composite of Skills, Knowledge, and Responsibility Score (converted from TAP rubric scale to LA state sysem's scale) and Student Growth Measure	50% Measures of Professional Practice + 50% Measures of Student Learning	35% Student Growth + 15% Academic Achievement + 50% Observation	25% Observation + 25% Multiple Measures + 50% Student Achievement	50% Observation + 40% Individual Growth + 5% Team Growth + 5% School Growth	Matrix: Instructional Practice and Professional Value Score + Student Learning Growth Score
Effectiveness Levels	4	5	5	5	5	4	5
Minimum number of required, summative observations (annually)	3 (career status) / 4 (probationary)	3	3	4 (>3 years experience)/2 (3+ years experience)/1 (teachers with past "5" ratings)	3	4	3
RATERS POLICIES							
What is the state policy on multiple raters?	Probationary teachers undergo 3 observations from principal and one from a peer. Career teachers undergo 3 observations from principal.	No requirement	No requirement	No requirement	No requirement	No requirement	No requirement
Who, in addition to administrators, can contribute obesrvation scores to teachers' summative evaluations?	Actual Peers (3)	Master Teachers	Master Teachers	Administrators only	Administrators	Expert Peers	Third Party Evaluators
Which teachers are required to be observed by multiple raters?	Probationary Teachers/ Contract Teachers	Yes	All	Teachers who scored a "1" on previous evaluation.	All	All	Highest and Lowest Performing Teachers
May other teachers be observed by mulitple raters?	Yes	Yes	Yes	Yes	Yes	Yes	Yes
How do various raters contribute to teachers' summative score(s)?	For probationary teachers, peers contribute 1 of the 4 required observations scores. For Contract status teachers, administrators may divide the required observations.	Frequent scored and unscored observation	Mentor teachers' ratings account for 35% of teachers' professional practice score. Principals' scores make up the remaining 65%. Mentor teachers also conduct a required observation for each teacher, but these are not counted in final summative evaluation scores.	Administrators co-observe low-perfomers. Observation "teams' consist of a teacher's own principal/AP and a principal/AP from another school in the district. The home principal issues the final summative score.	Third required observation must be conducted by administrator who did not conduct first two observations.	Frequent scored observation.	Administrator and TPV co-observe 3 lessons. TPVs' scores are kept separate and reviewed only if evaluation scores become consequential.

(1) All district data from National Center for Education Statistics, based on CCD Public school district data for the 2011-2012, 2012-2013 school years.

(2) The REIL project involves twelve districts in Maricopa County in greater Phoenix, AZ. This report examines only policies in participating REIL districts.

(3) The report uses the term "actual peer" to refer to peer observers who occasionally observe, rate, and provide feedback to other teachers, but do not spend the bulk of their time in this capacity. The report uses the term "expert peer" to refer to the master teachers, master educators, and third party evaluators employed as full- or nearly full-time in an evaluative capacity.

Tulsa, OK	Washington, DC	Boston, MA	Greenville, SC	Baltimore City, MD	Gwinnett County, GA	Hillsborough County, FL	Chicago, IL	New York, NY
41,199	44,618	55,027	72,153	84,212	132,370	194,041	403,004	968,143
94	134	135	96	195	134	326	647	1,523
2,457	3,472	4,261	4,376	5,532	10,323	13,862	22,460	62,368
TULSA Model	IMPACT	Educator Effectiveness System	Performance Appraisal System for Teachers	Teacher Effectiveness Evaluation	Teacher Effectiveness System	Empowering Effective Educators	Recognizing Educators Advancing Chicago's Students (REACH)	Advance
2011-2012	2009-2010	2012-2013	2006-2007	2013-2014	2013-2014	2010-2011	2012-2013	2013-2014
100% Professional Practice (observations, student surveys, etc.)	40-75% Teaching and Learning + 15-50% Student Achievement Data + 10% Commitment to School Community (4)	Matrix: Instructional Practice and Professional Values Score & Student Learning Growth Score	100% Professional Practice (observations, student surveys, etc.)	85% Observations + 15% Professional Expectations Measure	50% Observation & Survey Scores + 50% Student Growth & Academy Achievement	35% Pricipal Appraisal + 25% peer/Mentor Appraisal + 40% Student Achievement Gains	0-25% Student Growth + 75-100% Teacher Practice (varies by grade-level, subject)	40% Measures of Student Learning + 60% Measures of Teacher Practice
5	5	4 (5)	4	4	4	4	4	4
2 (contract teachers)/4 (probationary teachers)	1 to 5 (varies based on prior ratings)	1 to 5 (varies based on prior ratings)	0 (first year teachers)/ 6 (second year teachers)/ 1 (all others)	2	2 full-length + 4 short visits	3 to 11 (varies based on prior ratings)	2 (tenured teachers)/4 (probationary teachers)	4 to 6 (varies by effectiveness level and teacher preference)
No requirement	Required for all teachers except "highly effective"	No requirement	Required for novice & lowest performers	Required for low performers	No requirement	No requirement	No requirement	No requirement
Administrators only	Master Educators	Department Heads, Specialists, & Peers (6)	Actual Peers	Administrators only	Administrators only	Expert Peers, Supervisors	Administrators only (7)	Administrators and Validators
No requirement	All teachers rated "effective" or below	All	Probationary teachers eligible to move to continuing contracts	Teachers previously rated "ineffective"	No requirement	All	No requirement	Teachers rated "ineffective" in previous year
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
A single administrator must conduct mandatory observations. Other building administrators can contribute additional observation or evidence.	Administrators conduct two mandatory observations and collect evidence throughout school year. Master teachers conduct two observations.	Primary Evaluator issues summative rating. Secondary Evaluator conducts observations and collects evidence to be considered in primary rater's summative rating.	Observation scores	Observation scores	Administrators share observation duties and any other duties related to evaluation, but principals must sign off on summative ratings.	Observation scores	Principals and Assistant Principals share observation duties and other duties related to evaluation.	Validators score teachers rated "ineffective" in the previous evaluation cycle. Administrators may share other evaluation duties.

(4) In 2014-2015, all DCPS IMPACT scores will be calculated using the 75% Teaching and Learning + 15% Student Achievement + 10% Commitment to School Community formula because the district will not be calculating value-added scores during the first year of PARCC implementation. Teacher-assessed student achievement data (TAS) will be make up the 15% Student Achievement data component.

(5) Evaluations in Boston include two separate components: a summative performance rating (4 possible ratings) and a student impact rating (3 possible ratings).

(6) Only one or two campuses in Boston have established peer observation systems through special arrangements with the district. This report does not examine those cases, but acknowledges they exist.

(7) Chicago also employs Instructional Effectiveness Specialists to support principals in the evaluation process. They regularly co-observe with principals to help calibrate scoring, but their scores do not count toward teachers' summative ratings.

ENDNOTES

[1] Grover J. Whitehurst, Matthew M. Chingos, and Katharine M. Lindquist, "Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts," May 2014, The Brown Center on Education Policy at the Brookings Institution.

[2] "State of the States 2011: Trends and Early Lessons on Teacher Effectiveness and Evaluation Policies," October 2011, National Center on Teacher Quality; Kathryn M. Doherty and Sandi Jacobs. "State of the States 2013: Connect the Dots," October 2013, National Center on Teacher Quality.

[3] Susan E. Sporte, W. David Stearns, Kaleen Healey, Jennie Jiang and Holly Hart, "Teacher Evaluation in Practice: Implementing Chicago's REACH Students," September 2013, The University of Chicago Consortium on Chicago School Research, p. 25.

[4] TN Board of Education. "Teacher Evaluation in Tennessee: A Report on Year 1 Implementation," 2012: pg. 20.

[5] Ibid.

[6] Author Interview with Vince Botta, June 2014.

[7] Sporte et al, "Teacher Evaluation in Practice," p. 6.

[8] Jennifer Goldstein, "Making Observations Count: Key Design Elements for Meaningful Teacher Observation," December 2013, Policy Analysis for California Education, Stanford Graduate School of Education: p. 4.

[9] Stefanie Reinhorn, "Working Paper: Seeking Balance Between Assessment and Support: Teachers' Experiences of Teacher Evaluation in Six High-Poverty Schools," December 2013, The Project on the Next Generation of Teachers, Harvard Graduate School of Education, p. 23.

[10] Ibid. p. 23.

[11] Heather Hill and Pamela Grossman. "Learning from Teacher Observations: Challenges and Opportunities Posted by New Teacher Evaluation Systems," 2013, Harvard Education Review, Vol. 83, No. 2.

[12] Reinhorn, "Working Paper," p. 46.

[13] Andrew D. Ho and Thomas J. Kane, "The Reliability of Classroom Observations by School Personnel," 2013. Seattle, WA: The Bill & Melinda Gates Foundation.

[14] Chaplin, Duncan, Brian Gill, Allison Thompkins and Hannah Miller "Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in Pittsburgh Public Schools," July 2014, Institute of Education Sciences, U.S. Department of Education.

[15] Goldstein, 2013; Hill and Grossman, 2013.

[16] DeSoto Parish uses the TAP System, designed and supported by the National Institute for Excellence in Teaching.

[17] This report examines only those districts in Maricopa County that are participating in the Rewarding Excellence in Instruction and Leadership (REIL) initiatives, two TIF-funded project administered by the Maricopa County Education Service Agency. Twelve districts are currently participating in REIL and using the observation policies described in this report.

[18] Teachers who have not yet achieved "expert" designation but have a sustained history of effective/highly effective performance generally undergo a reduced number of observations and, in some cases, may not be observed by a master educator.

[19] Author Interview with David Weiner, March 2014.

ENDNOTES

[20] This report uses the term "expert peers" to refer to current or former teachers who devote a significant amount of time to observing, rating, and providing feedback to other teachers. They may carry reduced teaching loads or or have retired from the field but receive compensation for their engagement as observers. They may be called master teachers, master educators, or peers, depending on local policy. The report differentiates between these "expert peers," and "actual peers" who are also teachers only occasionally deployed in rating capacities. "Actual peers" spend nearly all of their time teaching, but occasionally observe, rate, and provide feedback to other teachers.

[21] Author Interview with Lori Renfro, April 2014.

[22] Author correspondence with Almi Abeyta, September 2014.

[23] Author Interview with Angela Rubenstein, April 2014.

[24] Greene County is using Race to the Top money to pay small stipends to the principals and to cover travel expenses administrators incur when traveling to other schools, some of which are more than 40 miles apart.

[25] Eagle County Public Schools. Professional Excellence, Accountability and Recognition. http://www.eagleschools.net/index.aspx?page=649

[26] Chambers, J. Brodziak de los Reyes, I., and Caitlin O'Neil, "How Much Are Districts Spending to Implement Teacher Evaluation Systems?," 2013, American Institutes for Research.

[27] Ibid.

[28] Ibid.

[29] Beginning in the 2014-2015 school year, Santa Fe will differentiate its observation process. Teachers who earned "effective" ratings last year will be observed twice, once by a building administrator and once by an administrator from another school. Teachers who are "minimally effective" will be observed three times: once by an administrator from their building, once by an administrator from another school, and once by the assistant superintendent. To help with these third observations, the district has hired two retired principals. Previously, all teachers were observed three times by a combination of administrators from their own and other schools.

[30] Author interview with Sue Sporte and Jennie Jiang, April 2014.

[31] Author interview with Lori Renfro, April 2014.

[32] Melinda Mangin and Sara Ray Stoelinga. "Peer? Expert? Teacher Leaders Struggle to Gain Trust While Establishing Their Expertise" June 2011, JSD. P 48-51. Vol 32. No 3.; Park, Sandra, Sola Takahashi and Taylor White. 2014. Developing an Effective Teacher Feedback System. Carnegie Foundation for the Advancement of Teaching.

[33] Author Interview with Kathy Noel, April 2014.

[34] Herlihy, Corinne, Ezra Karger, Cynthia Pollard, Heather C. Hill, Matthew Kraft, Megan Williams, and Sara Howard. State and Local Efforts to Investigate the Validity and Reliability of Scores. 2014 Teachers College Record, Vol 116, No 1.

[35] Liang, Guodong. "Teacher Evaluation Policies in the United states: Implementation and Impact on Constriuctivist Instruction." International Perspectives on Education and Society 19: 191.

[36] Grover J. Whitehurst, Matthew M. Chingos, and Katharine M. Lindquist, May 2014.

INTERVIEWS & PERSONAL CORRESPONDENCE

- Almi Abeyta, chief academic officer, Santa Fe Public Schools
- Stephanie Aberger, director, Align TLF training platform, IMPACT, Office of Human Capital District of Columbia Public Schools
- Vince Botta, director of performance management, Gwinnett County Public Schools
- Jana Burk, executive director of teacher/leader effectiveness initiative, Tulsa Public Schools
- Patty Fox, coordinator, performance assessment systems, Greenville County Public Schools
- Jeremy Gibbs, director of human resources and professional development, Transylvania County Schools
- Anne Heckman, director of educator quality, Eagle County Public Schools
- Michelle Hudacsko, deputy chief of human capital, IMPACT, District of Columbia Public Schools
- Jennie Jiang, research analyst, University of Chicago Consortium on Chicago School Research
- Vicki Kirk, superintendent, Greene County Public Schools
- Luke Kohlmoos, director of TEAM, Tennessee Department of Education
- Kathy Noel, director of student learning, DeSoto Parish Public Schools
- Paulette Poncelet, executive director of educator effectiveness, Chicago Public Schools
- Kim Procell, principal, DeSoto Parish Public Schools
- Lori Renfro, assistant superintendent for performance-based management systems, Maricopa County Educational Service Agency
- Bill Ripley, assistant director of academics and human resources, Greene County Public Schools
- Angela Rubenstein, implementation specialist, Boston Public Schools
- Michele Sherban, director of teacher evaluation and development, New Haven Public Schools
- Sarah Silverman, program director, education division, National Governors Association
- Sue Sporte, director for research operations, University of Chicago Consortium on Chicago School Research
- David Weiner, former deputy chancellor, New York City Department of Education
- Marie Whelan, director of evaluation, Hillsborough County Public Schools

SOURCES

Baltimore City Public Schools. 2013. *A Guide to City Schools' Teacher Effectiveness Evaluation.* Retrieved from: http://www.baltimorecityschools.org/cms/lib/MD01001351/Centricity/Domain/7955/20131018-TeacherEvaluationGuide-FINAL.pdf.

Chambers, Jay, Iliana Brodziak de los Reyes, and Caitlin O'Neil. 2013. *How Much Are Districts Spending to Implement Teacher Evaluation Systems?* Washington, DC: American Institutes for Research. Retrieved from: http://www.air.org/sites/default/files/downloads/report/What_Districts_Are_Spending_to_Implement_Teacher_Evaluation_Systems_Final_0.pdf.

Chaplin, Duncan, Brian Gil, Allison Thompkins, and Hannah Miller. 2014. *Professional practice, student surveys, and value-added: Multiple measures of teacher effectiveness in the Pittsburgh Public Schools.* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Education Laboratory Mid-Atlantic.

Chicago Public Schools. 2014. *Recognizing Educators Advancing Chicago Students.* Accessed 16 September 2014. http://www.cps.edu/pages/reachstudents.aspx.

Chikoore, Heather. 2012. *Making Teacher Evaluation Matter: District Strategies for Selecting and Training Evaluators.* Denver, CO: Colorado Legacy Foundation.

Cicarella, David. "A Fine Balance." *American Educator 23* (1): 18-21.

District of Columbia Public Schools. 2012. FY 2013 DCPS Budget Guide. Retrieved from: http://dcps.dc.gov/DCPS/Files/downloads/ABOUT%20DCPS/Budget%20-%20Finance/FY13%20documents/DCPS-Budget-Guide_FY2013.pdf.

Doherty, Kathryn and Sandi Jacobs. 2013. *State of the States 2013: Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice.* Washington, DC: National Center on Teacher Quality.

Greenville County Schools. 2014. *Teacher Evaluation –ADEPT/PAS-T.* Accessed 16 September 2014. Retrieved from: http://www.greenville.k12.sc.us/Employees/main.asp?titleid=teacheval.

SOURCES

Hillsborough County Public Schools. 2011. *Teacher Evaluation Handbook: Empowering Effective Teachers.* Retrieved from https://www.fldoe.org/profdev/pdf/pa/Hillsborough.pdf.

Goldstein, Jennifer. 2013. *Making Observations Count: Key Design Elements for Meaningful Teacher Observation.* Policy Analysis for California Education at Stanford Graduate School of Education.

Herlihy, Corinne, Ezra Karger, Cynthia Pollard, Heather C. Hill, Matthew Kraft, Megan Williams, and Sara Howard. "State and Local Efforts to Investigate the Validity and Reliability of Scores." *Teachers College Record 116* (1).

Hill, Heather C. and Pamela Grossman. "Learning from Teacher Observations: Challenges and Opportunities Posted by New Teacher Evaluation Systems." *Harvard Educational Review 83* (2): 371-41.

Ho, Andrew and Thomas J. Kane. 2013. *The Reliability of Classroom Observations by School Personnel.* Seattle, WA: Bill and Melinda Gates Foundation.

Liang, Guodong. "Teacher Evaluation Policies in the United States: Implementation and Impact on Constructivist Instruction." *International Perspectives on Education and Society 19:* 179-206.

Mangin, Melinda and Sara Ray Stoelinga. "Peer? Expert? Teacher Leaders Struggle to Gain Trust While Establishing Their Expertise" JSD 32 (3): 48-51.

New York City Department of Education. 2014. *Advance Guide for Educators.* Accessed 27 October 2014. http://www.uft.org/files/attachments/advance-guide-2014-15.pdf.

New York City Department of Education. 2014. *Teacher Evaluation and Development in NYC.* Accessed 16 September 2014. http://schools.nyc.gov/Offices/advance/default.htm.

Park, Sandra, Sola Takahashi, and Taylor White. 2014. *Developing an Effective Teacher Feedback System.* Carnegie Foundation for the Advancement of Teaching.

SOURCES

"Professional Excellence, Accountability and Recognition." Eagle County Public Schools. Accessed 16 September 2014. http://www.eagleschools.net/index.aspx?page=649

Reinhorn, Stefanie. 2013. *Working Paper: Seeking Balance Between Assessment and Support: Teachers' Experiences of Teacher Evaluation in Six High-Poverty Urban Schools.* Cambridge, MA: The Project on the Next Generation of Teachers. Retrieved from: http://isites.harvard.edu/fs/docs/icb.topic1231814.files/Seeking%20Balance%20Between%20Assessment%20and%20Support.pdf.

Sporte, Susan E., David Stevens, Kaleen Healey, Jennie Jiang and Holly Hart. *Teacher Evaluation in Practice: Implementing Chicago's REACH Students.* Chicago, IL: The University of Chicago Consortium on Chicago Schools Research.

State of the States 2011: Trends and Early Lessons on Teacher Effectiveness and Evaluation Policies October 2011. Washington, DC: National Center on Teacher Quality.

Tennessee Board of Education. 2012. Teacher Evaluation in Tennessee: A Report on Year 1 Implementation. Retrieved from: https://www2.ed.gov/programs/racetothetop/communities/tle2-year-1-evaluation-report.pdf.

Toch, Thomas and Robert Rothman. 2008. *Rush to Judgment: Teacher Evaluation in Public Education.* Washington, DC: Education Sector.

Tulsa Public Schools. 2013. TLE Observation & Evaluation Handbook for Evaluators using the Tulsa Model. Retrieved from: http://www.tulsaschools.org/4_about_District/_documents/TLE/Handbook_TLE_Observation_and_Evaluation_System_8-7.pdf.

Whitehurst, Grover J., Matthew M. Chingos, and Katharine M. Lindquist. 2014. E*valuating Teachers with Classroom Observations: Lessons Learned in Four Districts.* Washington, DC: Brown Center on Education Policy at the Brookings Institution.

Carnegie Foundation
for the Advancement of Teaching

Carnegie Foundation for the Advancement of Teaching
51 Vista Lane
Stanford, California 94305
650-566-5100

Carnegie Foundation for the Advancement of Teaching is committed to developing networks of ideas, individuals, and institutions to advance teaching and learning. We join together scholars, practitioners, and designers in new ways to solve problems of educational practice. Toward this end, we work to integrate the discipline of improvement science into education with the goal of building the field's capacity to improve.

www.carnegiefoundation.org

This publication was written with funding from The Joyce Foundation. The statements made and views expressed are solely the responsibility of the author. Carnegie is also grateful to the Institute of Education Sciences for supporting the Foundation's work on teacher evaluation and the improvement of teaching.