BROWN CENTER on
Education Policy
at BROOKINGS

May 2014

Evaluating Teachers with Classroom Observations

Lessons Learned in Four Districts

Reuters

Grover J. (Russ) Whitehurst, Matthew M. Chingos, and
Katharine M. Lindquist

Executive Summary

The evidence is clear: better teachers improve student outcomes, ranging from test scores to college attendance rates to career earnings. Federal policy has begun to catch up with these findings in its recent shift from an effort to ensure that all teachers have traditional credentials to policies intended to incentivize states to evaluate and retain teachers based on their classroom performance. But new federal policy can be slow to produce significant change on the ground. The Obama Administration has pushed the creation of a new generation of meaningful teacher evaluation systems at the state level through more than $4 billion in Race to the Top funding to 19 states and No Child Left Behind (NCLB) accountability waivers to 43 states. A majority of states have passed laws requiring the adoption of teacher evaluation systems that incorporate student achievement data, but only a handful had fully implemented new teacher evaluation systems as of the 2012-13 school year.

As the majority of states continue to design and implement new evaluation systems, the time is right to ask how existing teacher evaluation systems are performing and in what practical ways they might be improved. This report helps to answer those questions by examining the actual design and performance of new teacher evaluation systems in four urban school districts that are at the forefront of the effort to meaningfully evaluate teachers.

Although the design of teacher evaluation systems varies dramatically across districts, the two largest contributors to teachers' assessment scores are invariably classroom observations and test score gains. An early insight from our examination of the district teacher evaluation data is that *nearly all the opportunities for improvement to teacher evaluation systems are in the area of classroom observations rather than in test score gains.*

Despite the furor over the assessment of teachers based on test scores that is often reported by the media, in practice, only a minority of teachers are subject to evaluation based on the test gains of students. In our analysis, only 22 percent of teachers were evaluated on test score gains. All teachers, on the other hand, are evaluated based on classroom observation. Further, classroom observations have the potential of providing formative feedback to teachers that helps them improve their practice, whereas feedback from state achievement tests is often too delayed and vague to produce improvement in teaching.

Improvements are needed in how classroom observations are measured if they are to carry the weight they are assigned in teacher evaluation. In particular, we find that the districts we examined do not have processes in place to address the possible biases in observation scores that arise from some teachers being assigned a more able group of students than other teachers. Our data confirm that such a bias does exist: teachers with students with higher incoming achievement

Grover J. (Russ) Whitehurst is a senior fellow in Governance Studies and director of the Brown Center on Education Policy at the Brookings Institution.

Matthew M. Chingos is a fellow in the Brown Center on Education Policy at the Brookings Institution.

Katharine M. Lindquist is a research analyst in the Brown Center on Education Policy at the Brookings Institution.

levels receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels.

We should not tolerate a system that makes it hard for a teacher who doesn't have top students to get a top rating. Fortunately, there is a straightforward fix to this problem: adjust teacher observation scores based on student demographics. Our analysis demonstrates that a statistical adjustment of classroom observation scores for student demographics is successful in producing a pattern of teacher ratings that approaches independence between observation scores and the incoming achievement level of students. Such an adjustment for the makeup of the class is already factored into teachers' value-added scores; it should be factored into classroom observation scores as well.

We make several additional recommendations that will improve the fairness and accuracy of these systems:

- The reliability of both value-added measures and demographic-adjusted teacher evaluation scores is dependent on sample size, such that these measures will be less reliable and valid when calculated in small districts than in large districts. We recommend that states provide prediction weights based on statewide data for individual districts to use when calculating teacher evaluation scores.
- Observations conducted by outside observers are more valid than observations conducted by school administrators. At least one observation of a teacher each year should conducted by a trained observer from outside the teacher's school who does not have substantial prior knowledge of the teacher being observed.
- The inclusion of a school value-added component in teachers' evaluation scores negatively impacts good teachers in bad schools and positively impacts bad teachers in good schools. This measure should be eliminated or reduced to a low weight in teacher evaluation programs.

Overall, our analysis leaves us optimistic that new evaluation systems meaningfully assess teacher performance. Despite substantial differences in how individual districts designed their systems, each is performing within a range of reliability and validity that is both consistent with respect to prior research and useful with respect to improving the prediction of teacher performance. With modest modifications, these systems, as well as those yet to be implemented, will better meet their goal of assuring students' access to high-quality teachers.

Background

The United States is in the middle of a transformation in how teacher quality is characterized and evaluated. Until recently, teachers were valued institutionally in terms of academic credentials and years of teaching. This approach is still embedded in the vast majority of school districts across the country that utilize the so-called single salary schedule. Under this pay system, a regular classroom teacher's salary is perfectly predictable given but three pieces of information: the district in which she works, the number of years she has worked there continuously, and whether she has a post-baccalaureate degree. This conception of teacher quality based on credentials and experience is the foundation of the highly qualified teacher provisions in the present version of the federal Elementary and Secondary Education Act (No Child Left Behind, or NCLB), which was enacted by Congress in 2001, and is still the law of the land. A highly qualified teacher under NCLB must hold a bachelor's degree, be fully certified in the state in which she works, and have demonstrated competence in subject knowledge and teaching by passing a written licensure examination.[1]

Almost everyone who has been to school understands through personal experience that there are vast differences in the quality of teachers—we've all had really good, really bad, and decidedly mediocre ones, all of whom were deemed highly qualified in terms of paper credentials. In the last decade, the research community has been able to add substantially to the communal intuition that teachers differ markedly in quality by quantifying teacher performance, i.e., putting numbers on individual teachers' effectiveness. With those numbers in hand, researchers have been able to measure the short- and long-term consequences for students of differences in the quality of the teachers to which they are assigned, determine the best predictors of long-term teacher effectiveness, and explore the impact of human resource systems that are designed to attract, retain, and place teachers based on their performance rather than their years of service and formal credentials.

Among the yields from the new generation of research is evidence that having a better teacher not only has a substantial impact on students' test scores at the end of the school year, but also increases their chances of attending college and their earnings as adults. The difference in effectiveness between a teacher at the 84th percentile of the distribution and an average teacher translates into roughly an additional three months of learning in a year.[2] In turn, these differences in teacher quality in a single grade increase college attendance by 2.2 percent and earnings at age 28 by 1.3 percent.[3]

A consequence of these research findings has been a shift at the federal level from policies intended to ensure that all teachers have traditional credentials and are fully certified, to policies intended to incentivize states to evaluate and retain teachers based on their classroom performance. In the absence of a reauthorization of NCLB, the Obama administration has pushed this policy change forward using: first, a funding competition among states (Race to the Top); and, second, the availability of state waivers from the NCLB provisions for school accountability, conditional on a federally approved plan for teacher evaluation. Eighteen states and the District of Columbia won over $4 billion in Race to the Top funding and 43 states and the District of Columbia have been approved for NCLB waivers, in each case promising to institute meaningful teacher evaluation systems at the district level.[4] All told, 35 states and the District of Columbia have passed laws requiring the adoption of teacher evaluation systems that incorporate student achievement data, but as of the 2012-2013 school year, only eight states and the District of Columbia had fully implemented these systems. All other states were still in the process of establishing new systems.[5]

States face many challenges in implementing what they promised, undoubtedly including how to design the systems themselves. Ideally, a system for evaluating teachers would be: 1) practical in terms of the resources required for implementation; 2) valid in that it measures characteristics of teacher performance that are strongly related to student learning and motivation; 3) reliable in the sense of producing similar results across what should be unimportant variations in the timing and circumstances of data collection; 4) actionable for high-stakes decisions on teacher pay, retention, and training; and 5) palatable to stakeholders, including teachers, school and district leaders, policymakers, and parents.

None of these five characteristics of an ideal evaluation system is easy to accomplish given current knowledge of how to build and implement such systems. Expecting states to accomplish all five in short order is a federal Hail Mary pass.

Consider New York's not atypical approach to designing a statewide teacher evaluation system to comply with the promises it made to Washington. Based on a law passed by the state legislature, 20 to 25 percent of each teacher's annual composite score for effectiveness is based on student growth on state assessments or a comparable measure of student growth if such state growth data are not available, 15 to 20 percent is based on locally selected achievement measures, and the remaining 60 percent is based on unspecified locally developed measures.[6]

Notice that under these terms each individual school district, of which there are roughly 700 in New York, nominally controls at least 75 percent of the features of its evaluation system. But because the legislative requirement to base 25 percent of the teacher's evaluation on student growth on state assessments can only be applied directly to approximately 20 percent of the teacher workforce (only those teachers responsible for math and reading instruction in classrooms in grades and subjects in which the state administers assessments), this leaves individual districts in the position of controlling all of the evaluation system for most of their teachers. There is nothing necessarily wrong in principle with individual school districts being captains of their own ship when it comes to teacher evaluation, but the downside of this degree of local control is that very few districts have the capacity to develop an evaluation system that maximizes technical properties such as reliability and validity.[7]

Imagine two districts within New York that are similar in size and demographics. As dictated by state law, both will base 25 percent of the evaluation score of teachers in self-contained classrooms in grades 4-6 on student growth in math and reading over the course of the school year (the teachers of math and English in grades 7-8 may be included as well). The rest of the evaluation scores of that 20 percent of their workforce will be based on measures and weights of the district's own choosing. All of the measures for the rest of the workforce will also be of the district's choosing, as will the weights for the non-achievement measures.

There is an almost infinite variety of design decisions that could be made at the district level given the number of variables that are in play. As a result, the evaluation scores for individual teachers in two similar districts may differ substantially in reliability and validity, but the extent of such differences will be unknown. Many practical consequences flow from this. For example, given two teachers of equivalent actual effectiveness, it may be much easier for the one in District A to get tenure or receive a promotion bonus than it is for the one in District B next door.

Depending on one's political philosophy, such variation across districts in how teachers are evaluated and the attendant unevenness in how they are treated in personnel decisions could be good or bad. But bridging these philosophical and political differences in the value placed on local autonomy should be a shared interest in evaluation systems that provide meaningful differentiation of teacher quality as contrasted with the all too prevalent existing systems in which almost all teachers receive the same high rating if they are evaluated at all.[8] Further, no one, regardless of their political views, should want wholesale district-level disasters in instituting new teacher evaluation systems because this would have serious negative impacts on students. In that regard, the high probability of design, rollout, and operational problems in the many districts that do not have the internal capacity or experience to evaluate "by the numbers" is a ticking time bomb in terms of the politics of reforming teacher evaluation.

The Four District Project

To inform this report and contribute to the body of knowledge on teacher evaluation systems, we examined the actual design and performance of new teacher evaluation systems in four moderate-sized urban school districts scattered across the country. These districts are in the forefront of the effort to evaluate teachers meaningfully. Using individual level data on students and teachers provided to us by the districts, we ask whether there are significant differences in the design of these systems across districts, whether any such differences have meaningful consequences in terms of the ability to identify exceptional teachers, and whether there are practical ways that districts might improve the performance of their systems.

Our goal is to provide insights that will be useful to districts and states that are in the process of implementing new teacher evaluation systems, and to provide information that will inform future decisions by policymakers at the federal and state levels. We believe that our careful empirical examination of how these systems are performing and how they might be improved will help districts that still have the work of implementing a teacher evaluation system in front of them, and will also be useful to states that are creating statewide systems with considerable uniformity.

An early insight from our examination of the district teacher evaluation data was that most of the action and nearly all the opportunities for improvement lay in the area of classroom observations rather than in the area of test score gains. As we describe in more detail below, only a minority of teachers are subject to evaluation based on the achievement test gains of students for whom the teachers are the primary instructors of record, whereas all teachers are subject to classroom observations. Further, classroom observations have the potential of providing formative feedback to teachers that helps them improve their practice, whereas the summative feedback to teachers from state achievement tests is too delayed and nonspecific to provide direction to teachers on how they might improve their teaching and advance learning in their classrooms.

The weighting of classroom observations in the overall evaluation score of teachers varies across the districts in question and within districts depending on whether teachers can or cannot be linked to test score gains. But in no case is it less than 40 percent. A great deal of high-level technical attention has been given to the design of the "value-added" component of teacher evaluations. Value-added measures seek to quantify the impact of individual teachers on student learning by measuring gains on students' standardized test scores from the end of one school year to the end of the next. The attention given to value-added has led to the creation of a knowledge base that constrains the variability in how value-added is calculated and provides information on the predictive validity of value-added data for future teacher performance.[9] The technical work on classroom observations as used for teacher evaluation pales by comparison, even though it plays a much greater role in districts' overall teacher evaluation systems. The intention of this report is to correct some of that imbalance by providing a detailed examination of how four urban districts use classroom observations in the context of the overall teacher evaluation system and by exploring ways that the performance of classroom observations might be improved.

Methods

Overview of districts and data

Our findings are based on individual student achievement data linked to individual teachers from the administrative databases of four urban districts of moderate size. Enrollment in the districts ranges from about 25,000 to 110,000 students and the number of schools ranges from roughly 70 to 220. We have from one to three years of data from each district drawn from one or more of the years from 2009 to 2012. The data were provided to us in de-identified form, i.e., personal information (including student and teacher names) was removed by the districts from the data files sent to us.

Because our interest is in district-level teacher evaluation writ large rather than in the particular four districts that were willing to share data with us, and because of the political sensitivities surrounding teacher evaluation systems, which are typically collectively bargained and hotly contested, we do not provide the names of districts in this report. Further, we protect the identity of the districts by using approximations of numbers and details when being precise would allow interested parties to easily identify our cooperating districts through public records. No conclusions herein are affected by these efforts to anonymize the districts.

Findings

1. **The evaluation systems as deployed are sufficiently reliable and valid to afford the opportunity for improved decision-making on high stakes decisions by administrators, and to provide opportunities for individual teachers to improve their practice.**

When we examine two consecutive years of district-assigned evaluation scores for teachers with value-added ratings—meaning teachers in tested grades and subjects with whom student test score gains can be uniquely associated (an important distinction we return to in the next section)—we find that the overall evaluation scores in one year are correlated 0.33 to 0.38 with the same teachers' value-added scores in an adjacent year. In other words, multi-component teacher evaluation scores consisting of a weighted combination of teacher- and school-level value-added scores, classroom observation scores, and other student and administrator ratings have a statistically significant and robust predictive relationship with the ability of teachers to raise student test scores in an adjacent year. This year-to-year correlation is in keeping with the findings from a large prior literature that has examined the predictive power of teacher evaluation systems that include value-added.[10] Critics of reforms of teacher evaluation based on systems similar to the ones we examined question whether correlations in this range are large enough to be useful. In at least two respects, they are. First, they perform substantially better in predicting future teacher performance than systems based on paper credentials and experience. Second, they are in the range that is typical of systems for evaluating and predicting future performance in other fields of human endeavor, including, for example, the type of statistical systems used to make management decisions on player contracts in professional sports.[11]

We also examine the districts' evaluation systems for teachers without value-added scores—meaning teachers who are not in tested grades and subjects. We do so by assigning teachers with value-added scores the overall evaluation scores they would have received if they instead did not have value-added scores (i.e., we treat teachers in tested grades and subjects as if they are actually in non-tested grades and subjects). We calculate the correlation of these reassigned scores with the same teachers' value-added scores in an adjacent year. The correlations are lower than when value-added scores are included in the overall evaluation scores as described above, ranging from 0.20 to 0.32. These associations are still statistically significant and indicate that each district's evaluation system offers information that can help improve decisions that depend on predicting how effective teachers will be in a subsequent year from their present evaluation scores.

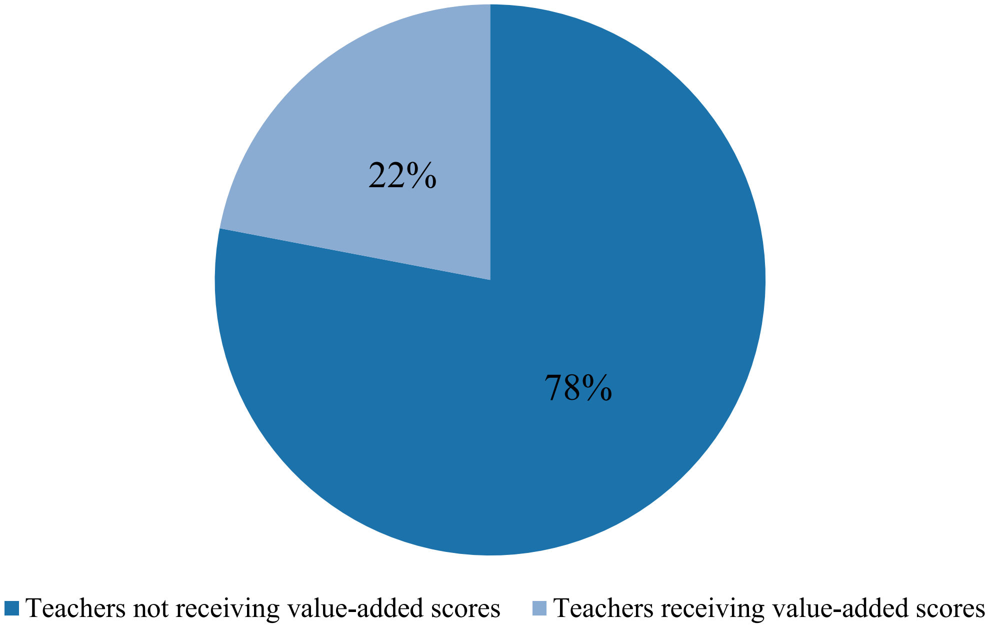
We calculate the year-to-year reliability of the overall evaluation scores as the correlation between the scores of the same teachers in adjacent years. "Reliability" is something of a misnomer here, as what is really being measured is the stability of scores from one year to the next. The reliability generated by each district's evaluation system is respectable, ranging from a bit more than 0.50 for teachers with value-added scores to about 0.65 when value-added is not a component of the evaluation score.

In summary, the districts on which we're reporting have created systems that are performing solidly within the expected range of reliability and validity for large scale teacher evaluation systems. The systems they've created, although flawed and improvable in ways we will discuss subsequently, can improve prediction of teachers' future performance compared to information based on teacher credentials and experience, and in so doing, provide for better administrative decisions on hiring, firing, and promotion. They are also capturing valid information through classroom observations that could be useful in the context of mentoring and professional development.

2. Only a small minority of the teacher workforce can be evaluated using gains in student test scores.

You would think from the media coverage of teacher evaluation and the wrangling between teacher unions and policy officials that the new teacher evaluation systems being implemented around the country are principally about judging teachers based on their students' scores on standardized tests. Value-added has been a bone of contention in almost every effort to replace the existing teacher evaluation systems that declare everyone a winner, with new systems that are designed to sort teachers into categories of effectiveness.

Figure 1. Percentage of all teachers receiving value-added scores in our four districts

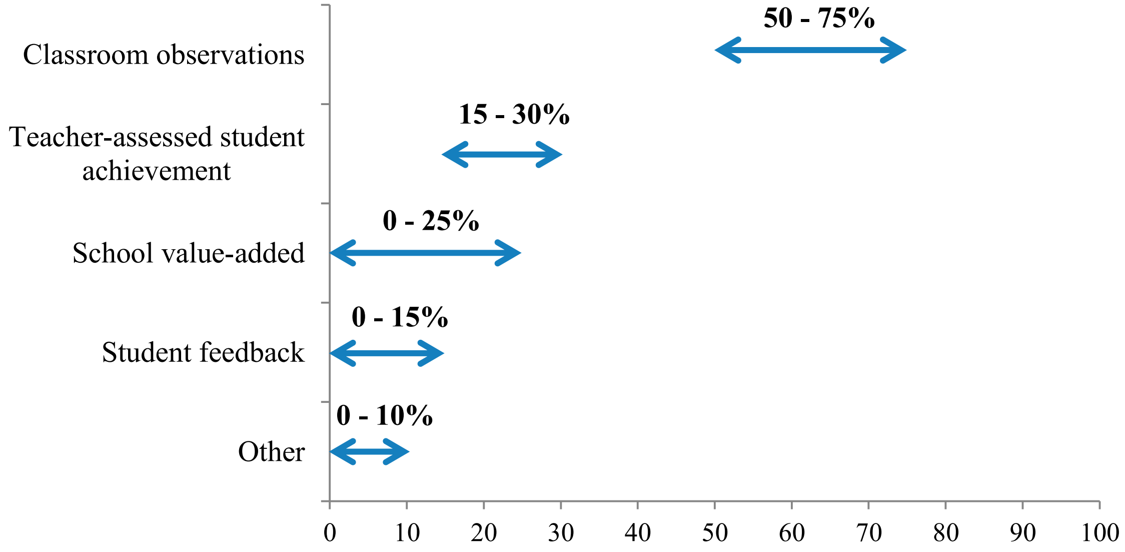


In our four districts, only about one fifth of teachers can be evaluated based on gains in their students' achievement test scores. Why? Under NCLB, states have to administer annual tests in mathematics and language arts at the end of grades 3 through 8. Third graders haven't been tested before the end of 3rd grade, so, with only a score at the end of the year and no pretest, their gains can't be calculated. Gain scores can be computed for 4th through 8th graders by subtracting their score at the end of their previous grade from their score at the end of their present grade. But by 6th or 7th grade, students are in middle school, which means that they typically have different teachers for different subjects. Thus, their gain scores in math and language arts can only be allocated to the teachers of those subjects. This means that only 4th, 5th, and 6th grade teachers in self-contained classrooms or 6th, 7th, and 8th grade teachers of math and language arts who remain the teacher of record for a whole year can be evaluated based on the test scores of the students in their classrooms.[12] Every other teacher has to be evaluated some other way, including, in our districts, by basing a portion of the teacher's evaluation score on classroom observations, achievement test gains for the whole school ("school value-added"), performance on non-standardized tests chosen and administered by each teacher to that teacher's students, and by some form of "team spirit" rating handed out by administrators.

Figure 2, below, represents how the four districts we have examined go about evaluating the roughly 80 percent of teachers who are in non-tested grades and subjects. As can be seen, classroom observations carry the bulk of the weight, comprising between 50 and 75 percent of teachers' overall evaluation scores. This is probably as it should be given: the uncertain comparability across teachers of gains on achievement tests selected by individual teachers; the weak relationship between the school's overall performance and that of any individual teacher within the school; and the possible dangers of having administrators rate teachers on ambiguous measures such as "commitment" or "professionalism." But, whereas there is logic to emphasizing classroom observations in the design of evaluation systems for teachers when value-added can't be computed, there has been little attention to how such classroom observations perform in teacher evaluation systems such as the ones in the four districts on which we focus. Our findings are both supportive of the continued use of classroom observations, and indicative of the need for significant improvements.

It is important to note that the true weight attached to each component in an evaluation system is determined not just by the percentage factor discussed above, but also by how much the component varies across teachers relative to other components. For example, if all teachers receive the same observation score, in practice it will not matter how much weight is assigned to the observations: the effect is like adding a constant to every teacher's evaluation score, which means that teachers will be ranked based on the non-observational components of the evaluation system.

Figure 2. Weight given to components of the overall teacher evaluation scores for teachers in non-tested grades and subjects across our four districts



3. Observations scores are more stable from year to year than value-added scores.

As shown in Table 1, observation scores have higher year-to-year reliability than value-added scores. In other words, they are more stable over time. This may be due, in part, to observations typically being conducted by school administrators who have preconceived ideas about a teacher's effectiveness. If a principal is positively disposed towards a particular teacher because of prior knowledge, the teacher may receive a higher observation score than the teacher would have received if the principal were unfamiliar with her or had a prior negative disposition. If the administrator's impression of individual teachers is relatively sticky from year-to-year, then observation scores will have a high level of reliability but will not be reflective of true teacher performance as observed. Indeed, we find evidence of this type of bias. As we report subsequently, classroom observers that are from outside the building provide more valid scores to teachers than administrators (such as principals) from within the teachers' schools. For this reason, maximizing reliability may not increase the effectiveness of the evaluation system.

Table 1. Year-to-year within- and cross-measure correlations of evaluation components

Correlation between:	Correlation
Observation in year one and observation in year two	0.65
Value-added in year one and value-added in year two	0.38
Observation in year one and value-added in year two	0.20
Value-added in year one and observation in year two	0.28

This leaves districts with important decisions to make regarding the trade-off between the weights they assign to value-added versus observational components for teachers in tested grades and subjects. The

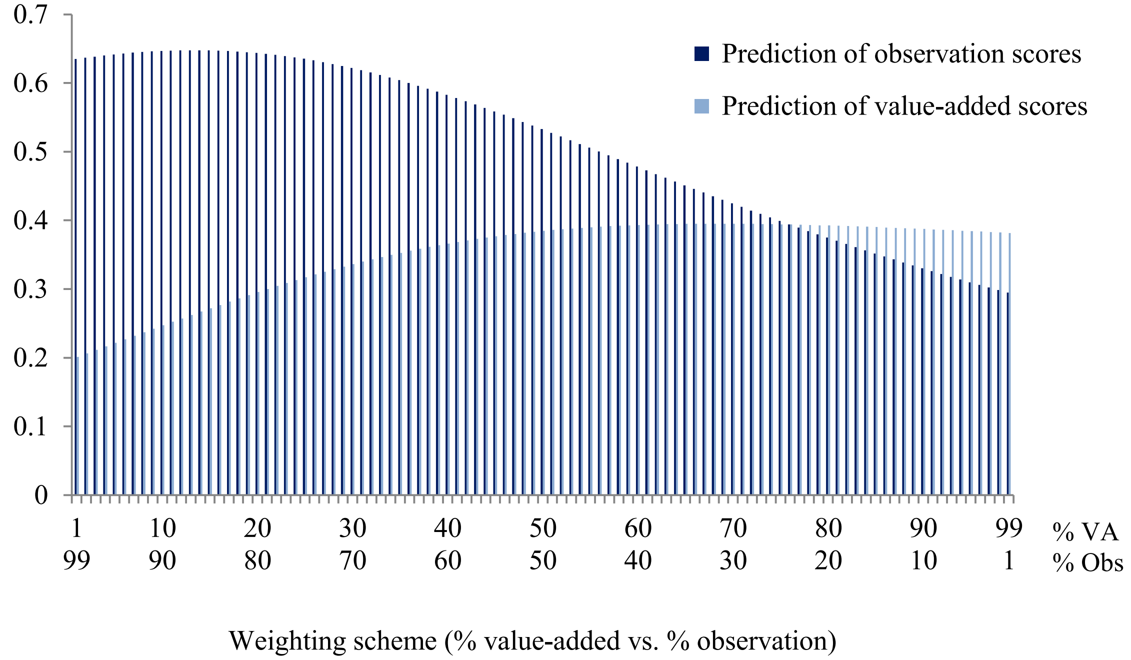
Gates Foundation Measures of Effective Teaching project (MET) has explored this issue and recommends placing 33 to 50 percent of the weight on value-added and splitting the rest of the weight between multiple components (including classroom observations and student evaluations), in order to provide a reasonable tradeoff between reliability and the ability to predict teachers' future value-added scores.[13] This is relevant information, but the Gates work is premised on the supremacy of an evaluation system with the ability to predict teachers' future test score gains (which many districts don't buy into), and it uses a student evaluation component that few districts have adopted.

We approach this issue agnostically with respect to the importance of observational versus value-added components of teacher evaluation systems. We find that there is a trade-off between predicting observation scores and predicting value-added scores of teachers in a subsequent year. Figure 3 plots the ability of an overall evaluation score, computed based on a continuum of different weighting schemes, to predict teachers' observation and value-added scores in the following year. The optimal ratio of weights to maximize predictive power for value-added in the next year is about two to one (value-added to observations) whereas maximizing the ability to predict observations requires putting the vast majority of weight on observations. A one-to-one weighting implies a significantly greater preference for the prediction of value-added than for the prediction of observation scores.

We do not believe there is an empirical solution for the ideal weights to assign to observation versus value-added scores. The assignment of those weights depends on the a priori value the district assigns to raising student test scores, the confidence it has in its classroom observation system as a tool for both evaluation and professional development, and the political and practical realities it faces in negotiating and implementing a teacher evaluation system.

At the same time, there are ranges of relative weighting—namely between 50 and 100 percent value-added—where significant increases in the ability to predict observation scores can be obtained with relatively little decrease in the ability to predict value-added. Consequently, most districts considering only these two measures should assign a weight of no more than 50 percent to value-added—i.e. a weight on observations of at least 50 percent.

Figure 3. Predictive power of alternative weighting schemes for prediction of observation versus value-added scores, by prediction type



It is important to ensure that the observation system in place makes meaningful distinctions among teachers. As discussed previously, an observation system that gives similar ratings to most teachers will not carry much weight in the overall evaluation score, regardless of how much weight is officially assigned to it. An important consequence of this fact is that an observation system with little variation in scores would make it very difficult for teachers in non-tested grades and subjects to obtain very high or low total evaluation scores—they would all tend to end up in the middle of the pack, relative to teachers with value-added scores.

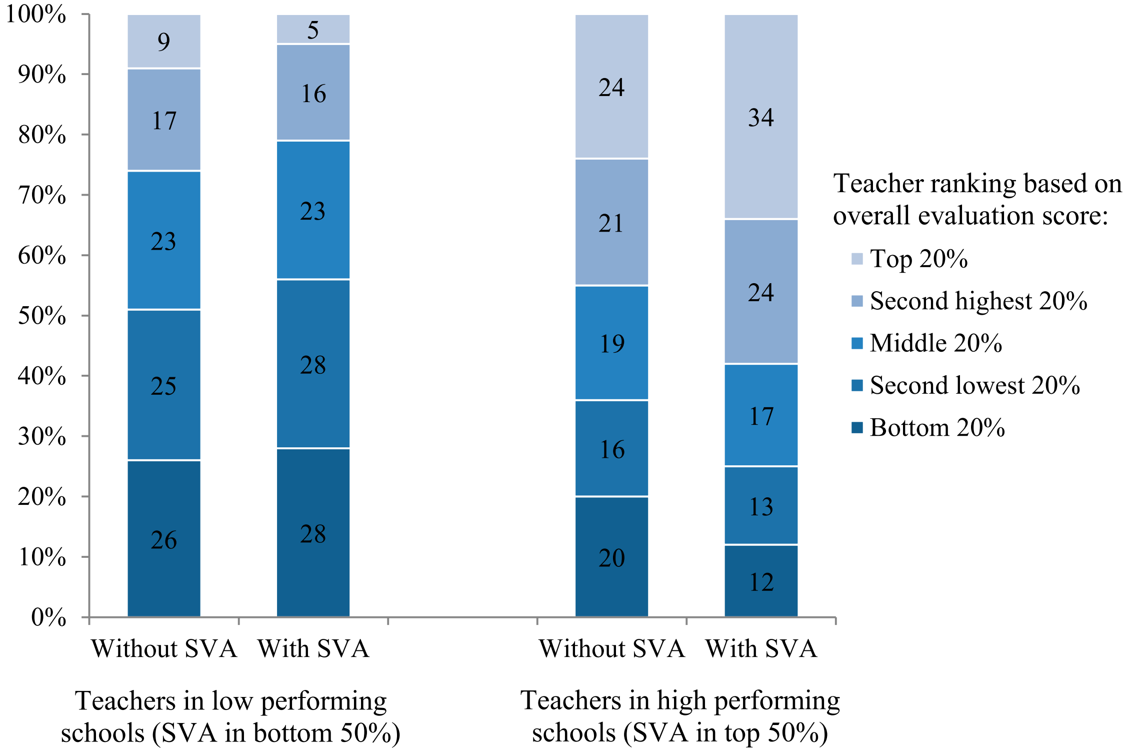
Recommendation: Classroom observations that make meaningful distinctions among teachers should make up at least 50 percent of the overall teacher evaluation score.

4. The inclusion in individual teacher evaluation scores of a school value-added component negatively impacts good teachers in bad schools and positively impacts bad teachers in good schools.

Some districts include a measure of school value-added in the evaluation scores for individual teachers—a measure of student gains on achievement tests averaged across the whole school. This is often for teachers without individual value-added scores (meaning the teachers are not in tested grades and subjects), and the measure is meant to capture the impact that every staff member in a building has on student achievement. In practice, the inclusion of such a measure brings all teachers' scores closer to the school value-added score in their building. The overall scores of the best performing teachers in a building are brought down and the overall scores of the worst performing teachers in a building are brought up.

Figure 4 shows the percentage of teachers who receive scores in each quintile when all teachers in the district are ranked by their overall evaluation scores, both with the inclusion of school value-added and without. The two left-most bars show teachers in low performing schools (those with school value-added in the bottom half of all schools) and the two right-most bars show teachers in high performing schools (those with school value-added in the top half of all schools). When school value-added is included as a component of overall teacher effectiveness measures, the percentage of teachers in low performing schools who score in the top 20 percent of all teachers in the district is cut almost in half, dropping from nine to five percent (lightest blue sections of the two left-most bars). Conversely, the percentage of teachers in high performing schools who score in the bottom 20 percent of all teachers is also cut almost in half, dropping from 20 to 12 percent (darkest blue sections of the two right-most bars).

Figure 4. Distribution of teachers ranked into quintiles by evaluation scores that include and exclude measures of school value-added, divided into teachers in low performing and high performing schools



School value-added measures partly reflect the overall quality of teaching in the school, and separating the teaching and non-teaching components is difficult.[14] But including school value-added in an overall evaluation measure has an important practical impact on high-performing teachers who are in low-performing schools (and vice versa). It creates an incentive for the best teachers to want to work in the best schools—if they move to a low performing school, their effectiveness will be rated lower and they could miss out on bonuses and other preferential treatment given to top teachers. It also creates a system that is demonstrably and palpably unfair to teachers, given that they have little control over the performance of the whole school.

Recommendations: Eliminate or reduce to a low weight the contribution of school-wide value added to individual teacher evaluation scores. If the goal is to motivate teachers to cooperate with one another, consider other means of achieving that end, such as modest school-based bonuses or evaluation of individual teachers based on their contributions to activities that benefit instruction and learning outside of their individual classrooms.

5. Teachers with (initially) higher performing students receive higher classroom observation scores.

Considerable technical attention has been given to wringing the bias out of value-added scores that arises because student ability is not evenly distributed across classrooms. It is obvious to everyone that a teacher evaluation system would be intolerable if it identified teachers in the gifted and talented program as superior to other teachers because students in the gifted and talented program got higher scores on end-of-year tests. Value-added systems mitigate this bias by measuring test score gains from one school year to the next, rather than absolute scores at the end of the year, and by including statistical controls for characteristics of students and classrooms that are known to be associated with student test scores, such as students' eligibility for free and reduced price lunch. Thus, in a system in which teacher effectiveness is calculated by measuring demographically adjusted test score gains, the teacher in the gifted and talented program has no advantage over the teacher of regular students. And, likewise, a teacher of regular students who happens to get assigned more than his fair share of lower performing students relative to his colleague in the classroom next door is not at a disadvantage compared to that colleague if his evaluation score is based on test score gains.

But as we have previously described, classroom observations, not test score gains, are the major factor in the evaluation scores of most teachers in the districts we examined, ranging from 40 to 75 percent of the total score depending on the district and whether the teacher is responsible for a classroom in a tested grade and subject. Neither our four districts, nor others of which we are aware have processes in place to address the possible biases in these observation scores that arise from some teachers being assigned a more able group of students than other teachers.

Imagine a teacher who, through the luck of the draw or administrative decision, gets an unfair share of students who are challenging to teach because they are less well prepared academically, aren't fluent in English, or have behavioral problems. Now think about what a classroom observer is asked to judge when rating a teacher's ability. For example, in a widely used classroom observation system created by Charlotte Danielson, a rating of "distinguished" on questioning and discussion techniques requires the teacher's questions to consistently provide high cognitive challenge with adequate time for students to respond, and requires that students formulate many questions during discussion.[15] Intuitively, the teacher with the unfair share of students that are challenging to teach is going to have a tougher time performing well under this rubric than the teacher in the gifted and talented classroom.

This intuition is borne out in our data: teachers with students with higher incoming achievement levels receive classroom observation scores that are higher on average than those received by teachers whose incoming students are at lower achievement levels. This finding holds when comparing the observation scores of the same teacher at different points in time. The latter finding is important because it indicates that the association between student incoming achievement levels and teacher observation scores is not due, primarily, to better teachers being assigned better students. Rather, it is due to bias in the observation system—when observers see a teacher leading a class with higher ability students, they judge the teacher to be better than when they see that same teacher leading a class of lower ability students.

Figure 5. Distribution of teachers ranked by classroom observation scores, divided into quintiles by the average incoming achievement level of their students

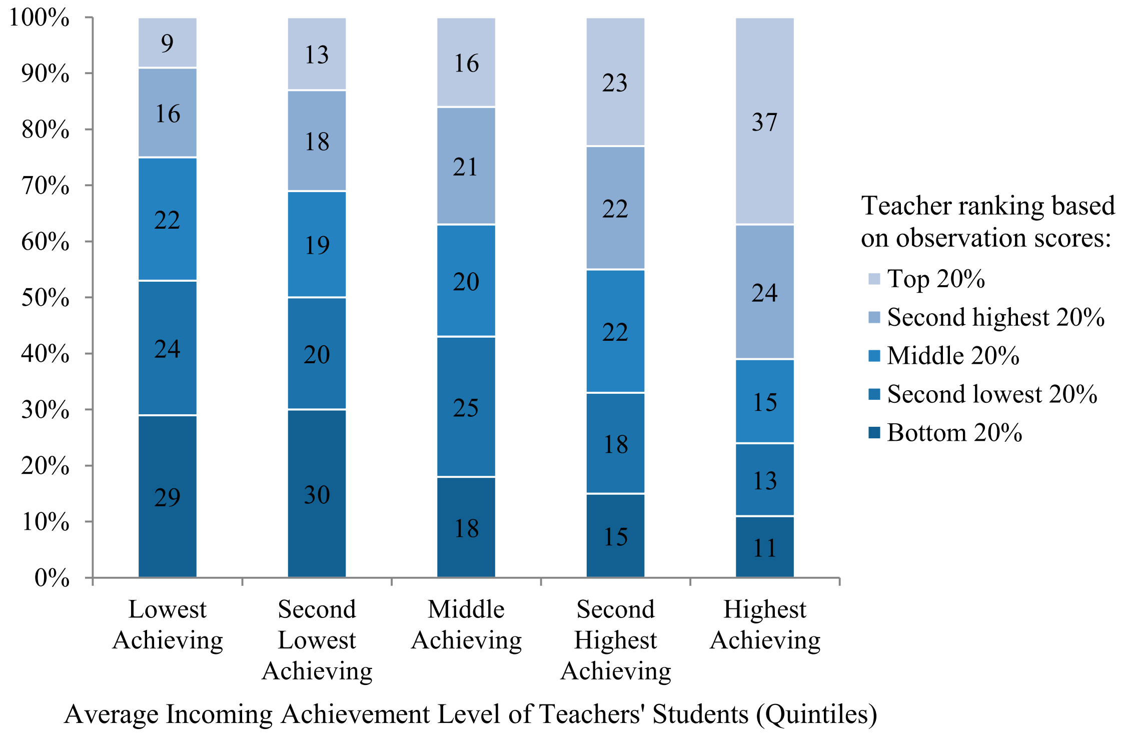
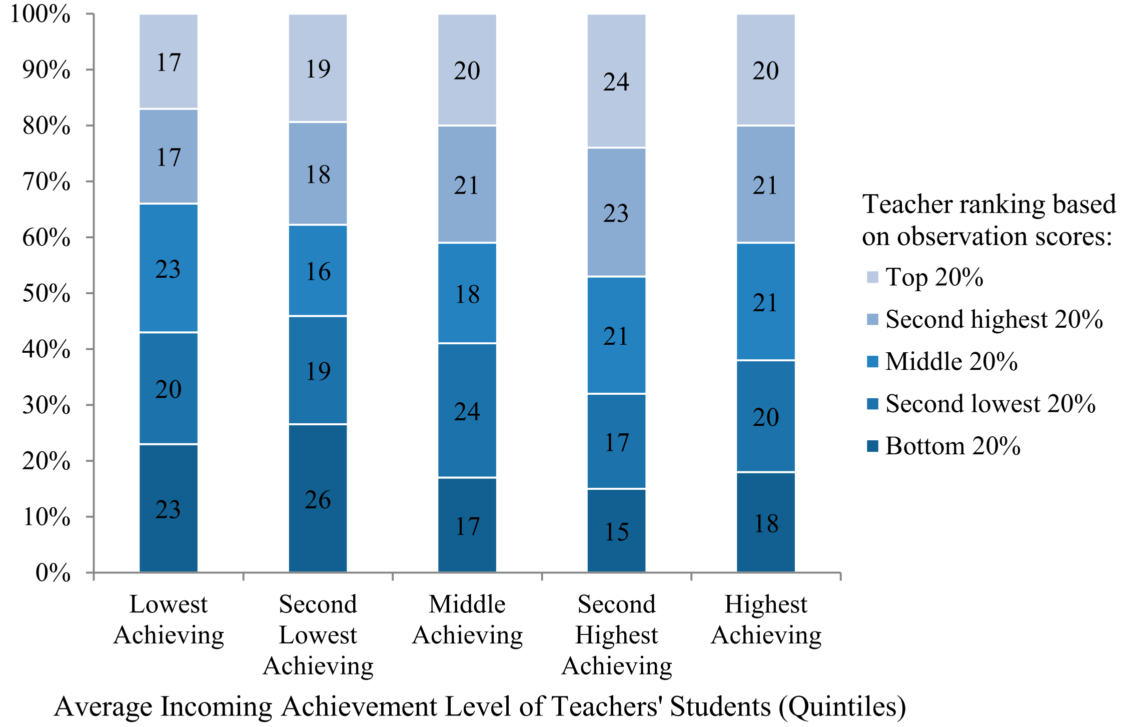


Figure 5, above, depicts this relationship using data from teachers in tested grades and subjects, allowing us to examine the association between the achievement levels of students that teachers are assigned and the teachers' classroom observation scores. Notice that only about 9 percent of teachers assigned a classroom of students who are "lowest achieving" (in the lowest quintile of academic performance based on their incoming test scores) are identified as top performing based on classroom observations, whereas the expected outcome would be 20 percent if there was no association between students' incoming ability and a teacher's observation score. In contrast, four times as many teachers (37 percent) whose incoming students are "highest achieving" (in the top quintile of achievement based on incoming test scores) are identified as top performers according to classroom observations. Again, the expectation under an assumption of independence between student ability and observation score is 20 percent.

Considering Figure 5 as a whole, there is a very strong statistical association between the incoming achievement level of students and teacher ranking based on observation scores.[16] This represents a substantively large divergence from what might be expected from a "fair" system in which teacher ratings would be independent of the incoming quality of their students.

We believe this represents a very serious problem for any teacher evaluation system that places a heavy emphasis on classroom observations, as nearly all current systems are forced to do because of the lack of measures of student learning in most grades and subjects. We should not tolerate a system that makes it hard to be rated as a top teacher unless you are assigned top students.

Figure 6. Distribution of teachers ranked by classroom observation scores adjusted for student demographics, divided into quintiles by the average incoming achievement level of their students



Fortunately, there is a straightforward fix to this problem: adjust teacher observation scores based on student demographics. Figure 6 is based on the same students and teachers as Figure 5, but in this case, each teacher's observation score is adjusted statistically for the composition of her class, using percent white, percent black, percent Hispanic, percent special education, percent free or reduced price lunch, percent English language learners, and percent male.

Each of these student demographic features is correlated to a statistically significant degree with achievement test scores. Table 2, below, shows the differences in test scores of students with different characteristics, holding constant the other characteristics in the table. For example, the first coefficient (-0.297) indicates that the average white student scores roughly 30 percent of a standard deviation lower than the average Asian student. Likewise, the average black student scores more than a full standard deviation below the average Asian student. Achievement gaps related to English language learner and special education status are nearly as large.

These data, coupled with the previously demonstrated relationship between prior-year scores and observation scores, make clear that adjusting statistically for these variables gives teachers who receive a disproportionate share of lower-performing students a boost in their classroom observation score. Likewise, the statistical adjustment lowers the observation scores of teachers who receive a disproportionate share of students whose demographic characteristics are positively associated with achievement, such as Asians and females. We do not adjust for prior test scores directly because doing so is only possible for the minority of teachers in tested grades and subjects.

Table 2. Relationship between student demographics and students' average math and English language arts standardized test scores from the prior year

Demographic Variable	
White (relative to Asian)	-0.297*** (0.009)
Black (relative to Asian)	-1.021*** (0.008)
Hispanic (relative to Asian)	-0.536*** (0.009)
Special education (relative to non-special education)	-0.922*** (0.004)
Eligible for free or reduced price lunch (relative to ineligible)	-0.423*** (0.002)
English language learner (relative to non-learner)	-0.987*** (0.006)
Male (relative to female)	-0.160*** (0.002)

Note: *** indicates that the coefficient is statistically significant at the 1 percent level. Standard errors appear in parentheses. The dependent variable (student achievement test scores) is standardized to have mean of 0 and a standard deviation of 1. This model includes controls for year and district.

Such an adjustment for the makeup of the class is already factored in for teachers for whom value-added is calculated, because student gains are also adjusted for student demographics. However, these teachers are also evaluated in large part based on observation scores, which make no adjustment of the kinds of students in the class. For the teachers in non-tested grades and subjects, whose overall evaluation score is based primarily on classroom observations, there is no adjustment for the make-up of their class in any portion of their evaluation score.

A comparison of the two figures presented above indicates that the statistical adjustment of classroom observation scores for student demographics is successful in producing a pattern of teacher ratings that is much closer to independence between observation scores and the incoming achievement level of students than is the case when raw classroom observations scores are used. Thus, in the second figure, teachers who are assigned students with incoming achievement in the lowest quintile have close to the same chance of being identified as a high performing teacher as teachers who are assigned students in the top quintile of performance. There remains a statistical association between incoming student achievement test scores and teacher ratings based on classroom observations, but it is reduced substantially by the adjustment for student demographics.[17]

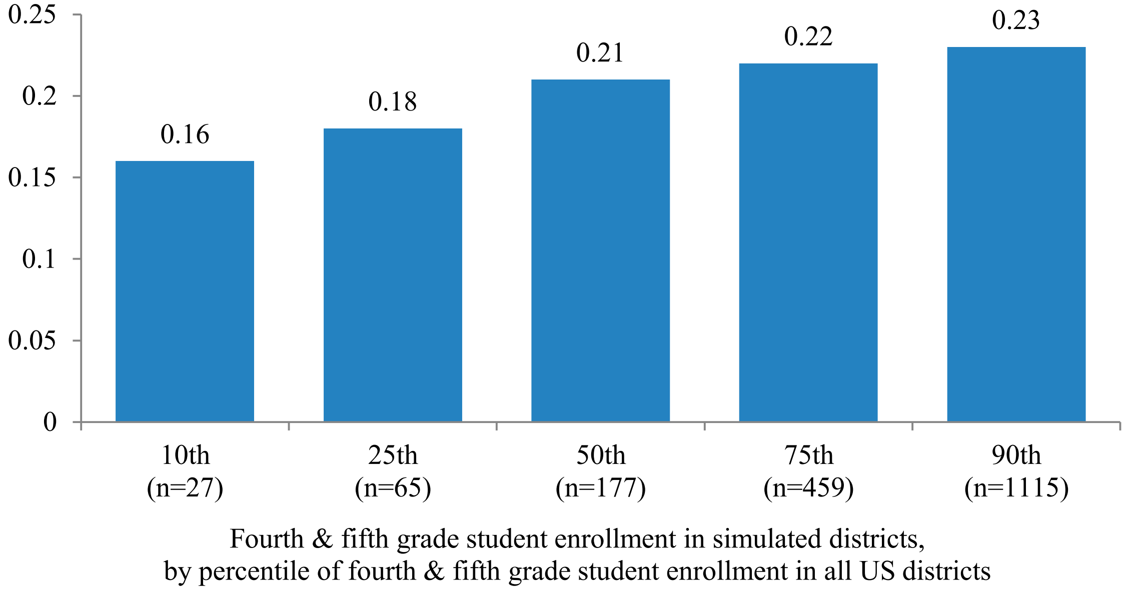
This adjustment would have significant implications if it were used in teacher evaluation systems. When we divide teachers into deciles based on both unadjusted and adjusted observation scores, we find that roughly half of teachers move out of the decile of performance to which they would have been assigned based on their unadjusted classroom observation score when the adjusted score is used.

Recommendations: Adjust teacher observation scores for the demographic characteristics of their classrooms. Do not use raw observation scores to calculate teacher effectiveness.

6. The reliability of demographic-adjusted teacher evaluation scores is dependent on sample size such that these measures will be less reliable and predictive of future teacher performance when calculated in small districts than in large districts.

Using data from one of our largest urban districts, we simulated the creation of value-added measures for 4th and 5th grade teachers in small districts to test the reliability of the estimates (defined as the year-to-year correlation within measures). These measures use information on student subgroups to adjust the scores assigned to teachers. We simulated the creation of these measures in districts at the 10th, 25th, 50th, 75th, and 90th percentiles of 4th and 5th grade student enrollment in all U.S. school districts, by pulling a set number of students from our target district's full database and calculating value-added measures for a "district" of that size.[18] The coefficients on student demographics and prior year test scores were pulled from each adjustment model for small "districts" and applied to the target district's full database to keep the sample size consistent when computing reliability (about 13,000 4th and 5th graders). In other words, we are allowing the sample size of the adjustment model to vary according to simulated district size but holding constant the sample used to calculate reliability. Figure 7, below, shows that as student enrollment increases, the reliability of value-added measures also increases until it converges on the reliability of measures in the largest districts.

Figure 7. Reliability (year-to-year within-measure correlation) of value-added measures in simulated districts, by size of simulated district

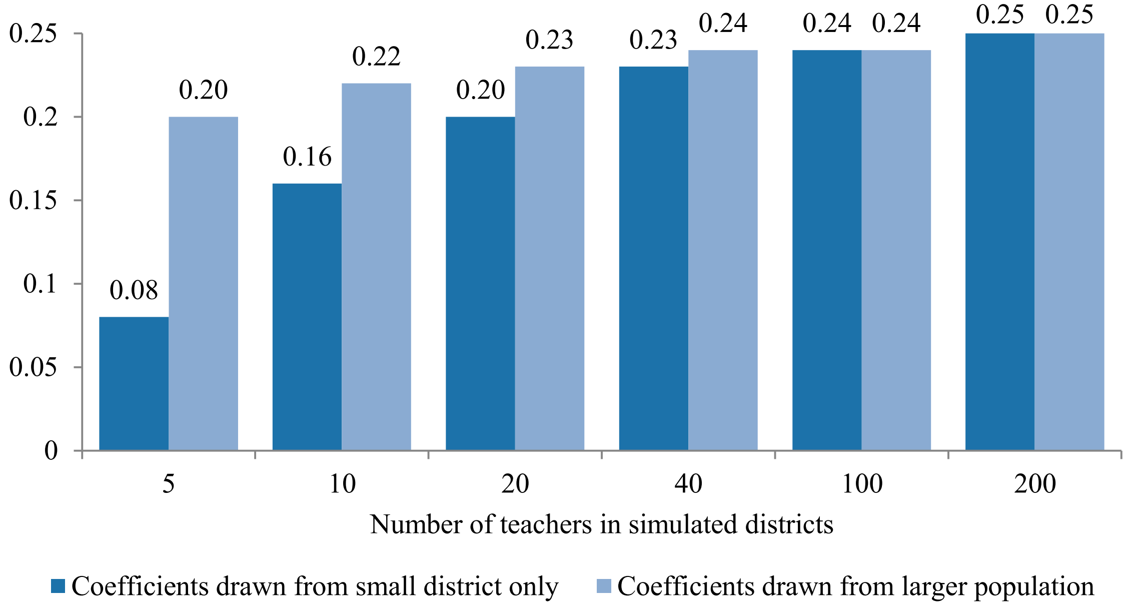


We also simulated districts with different numbers of 4th and 5th grade teachers, starting at five teachers and working up to 200. Again, we find that as the number of 4th and 5th grade teachers (and thus students) increases, the reliability of value-added measures calculated for those teachers increases. In our simulation, the value-added measures in a district with only five 4th and 5th grade teachers have an average reliability below 0.10 due to the small sample sizes, but the average reliability for a district with 200 teachers is close to 0.25. This pattern is illustrated below in Figure 8 by the dark blue bars.

Through simulation, we tested a mechanism for increasing the reliability of value-added measures in very small districts by treating those districts as part of larger geographic areas with more students (e.g. neighboring districts or the whole state). Due to higher sample sizes, the performance of groups of students in larger geographic areas is expected to be more stable over time. When the coefficients from an adjustment model calculated for 4th and 5th grade teachers in our large school district are applied to smaller simulated districts, the reliability of demographic-adjusted measures in the small districts increases significantly. As can be seen in Figure 8, below, this process increases the average reliability of value-added measures in simulated districts with five 4th and 5th grade teachers from 0.08 to 0.20 (the two left-most bars). In practice, the coefficients from larger districts could be provided to smaller school districts for use in their evaluation systems.

We repeated this analysis using demographic-adjusted observation scores as the evaluation measure of interest (instead of value-added scores), and came to the same conclusion: the number of observations in the adjustment model has an important impact on the reliability of the resulting evaluation measures.

Figure 8. Reliability (year-to-year within-measure correlation) of value-added measures in districts of different sizes, by adjustment method



The small number of states that already have evaluation systems in place address this issue by calculating value-added scores centrally and providing them to local school districts. This should remain the norm, but it is usually only done for value-added based on state tests, not for district-specific tests or for observation scores. For example, in New York state, districts are required to base 20 to 25 percent of all teachers' evaluation scores on student growth data; for teachers outside of the grades and subjects covered by compulsory state-wide assessments, districts have local control over which assessments are used and whether or not adjustments are made for student demographics.[19] The part of the evaluation score provided by the state—value-added on state tests—thus only constitutes a small part of the set of measures that would benefit from large-sample coefficients.

Recommendations: States should supply grade-specific coefficients for districts to use on prior year test scores and student demographics when calculating teacher evaluation scores, including both value-added and demographically-adjusted observation scores. These coefficients can be based either on statewide data or on groups of demographically-similar districts. This provision of large sample coefficients will even out the reliability of teacher evaluation systems across districts of different size and different student composition.

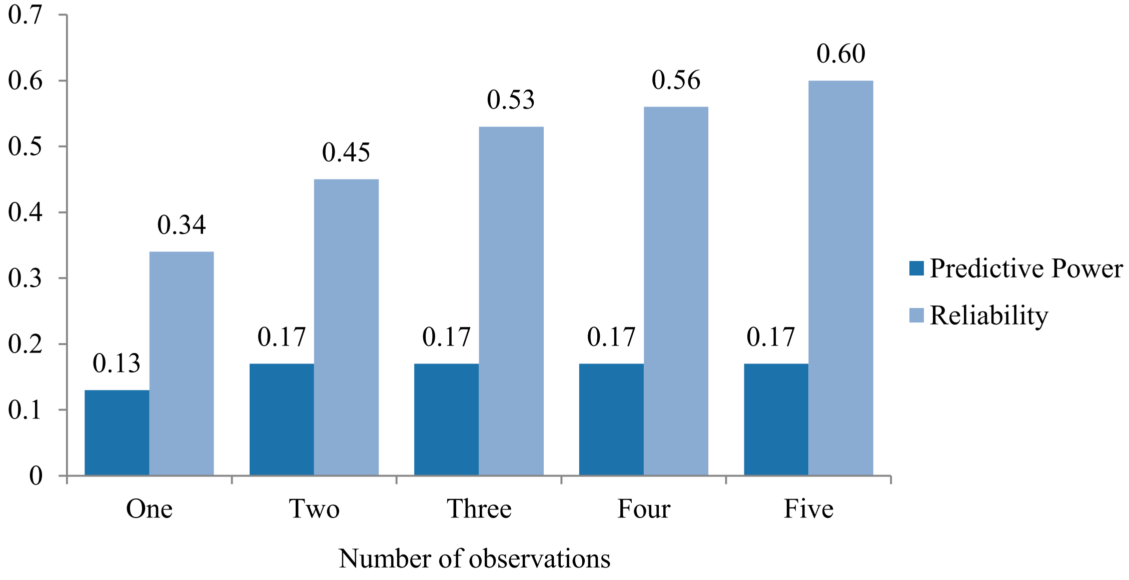
7. Observations conducted by outside observers are more predictive than observations conducted by school administrators.

In all of our districts, the quality of information garnered from classroom observations depends on how many are conducted. In districts that use a mix of building leaders and central administration staff to conduct classroom observations, the quality of the information also depends on who conducts the observations. Moving from one to two observations (independent of who conducts them) increases both the reliability of observation scores and their predictive power for value-added scores in the next year. Adding additional observations continues to increase the reliability but has no further effect on the predictive power of future value-added scores. This pattern is illustrated in Figure 9, below.

Observations conducted by in-building administrators, e.g., the principal, are more reliable (0.61) than those done by central administration staff (0.49), but observations conducted by evaluators from outside the building have higher predictive power for value-added scores in the next year (0.21) than those done by administrators in the building (0.15).

We interpret the higher year-to-year stability of observations conducted by the principal or assistant principal compared to out-of-building observers as reflecting the influence of the principal's opinion about a given teacher in addition to the information that is derived from the classroom observation itself. In this sense, the higher reliability may not be desirable if the subsequent judgments made by the principal are not directly related to the teacher's performance, but rather on a pre-existing impression. This interpretation is consistent with the higher predictive power to value-added of the observations conducted by out-of-building observers (as stated above).

Figure 9. Reliability and predictive power of observation measures, by number of observations



Classroom observations are expensive, and, for the majority of teachers, they are the most heavily weighted contributor to their individual evaluation score. Observations also have a critical role to play for principals who intend to be instructional leaders, as they present a primary point of contact between the school leader and classroom teaching and learning. It is important to balance what are, in part, the competing demands of empowering the school leader to lead, spending no more than necessary in staff time and money to achieve an effective observation system, and designing a system in which observation scores are based on what is observed rather than extraneous knowledge and prior relationships between the observer and the teacher.

Our data suggest to us that three observations squeeze about as much reliability and validity out of the process as five, and that the system either needs unbiased observers from outside the building as a validity check on principal observations, or, as is recommended in the Gates MET project, training and reliability checks on the principal or other in-building observers.[20]

Recommendations: Conduct two-to-three annual classroom observations for each teacher, with at least one of those observations being conducted by a trained observer from outside the teacher's school without substantial prior knowledge of, or conflict of interest with respect to the teacher being observed. Arrange for an additional classroom observation by another independent observer in cases in which there are substantial and potentially consequential differences between the observation scores generated by the primary observers.

Conclusions

A new generation of teacher evaluation systems seeks to make performance measurement and feedback more rigorous and useful. These systems incorporate multiple sources of information, including such metrics as systematic classroom observations, student and parent surveys, measures of professionalism and commitment to the school community, more differentiated principal ratings, and test score gains for students in each teacher's classrooms.

Although much of the impetus for new approaches to teacher evaluation comes from policymakers at the state and national levels, the design of any particular teacher evaluation system in most states falls to individual school districts and charter schools. Because of the immaturity of the knowledge base on the design of teacher evaluation systems, and the local politics of school management, we are likely to see considerable variability among school districts in how they go about evaluating teachers.

That variability is a double-edged sword. It offers the opportunity to study and learn from natural variation in the design of evaluation systems, but it also threatens to undermine public support for new teacher evaluation systems to the extent that the natural variation suggests chaos, and is used by opponents of systematic teacher evaluation to highlight the failures of the worst performing systems. The way out of this conundrum is to accelerate the process by which we learn from the initial round of district experiences and to create leverage points around that learning that will lift up the weakest evaluation systems.

Our examination of the design and performance of the teacher evaluation systems in four medium-sized urban districts provides reasons for optimism that new, meaningful evaluation systems can be designed and implemented by individual districts: despite substantial differences in how our individual districts designed their systems, each is performing within a range of reliability and validity that is both consistent with respect to prior research and useful with respect to improving the prediction of teacher performance.

At the same time, we find that our districts share, to one degree or another, a number of design decisions that contribute to lower than possible performance in terms of the reliability, validity, and consequences of the systems, and that will probably be seen as mistakes by stakeholders as more experience with the systems accrues. The most important of these, we believe, is the failure to adjust classroom observation scores for the degree to which the students assigned to a teacher create challenging conditions for the teacher. Put simply, the current observation systems are patently unfair to teachers who are assigned less able and prepared students. When this bias is combined with a system design in which school-wide value-added contributes significantly to individual teachers' overall evaluation score, the result is an unintended but strong incentive for good teachers to avoid teaching low-performing students and to avoid teaching in low-performing schools.

A prime motive behind the move towards meaningful teacher evaluation is to assure greater equity in students' access to good teachers. A teacher evaluation system design that inadvertently pushes in the opposite direction is clearly undesirable. We have demonstrated that these design errors can be corrected with tools in hand.

End Notes

1. For more information on this topic, see Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling, *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (New York: The New Teacher Project, 2009). http://widgeteffect.org/downloads/TheWidgetEffect.pdf.
2. Eric A. Hanushek and Steven G. Rivkin, "Generalizations about Using Value-Added Measures of Teacher Quality," *American Economic Review* 100, no. 2 (2010): 267-71. We translate effects of 0.11 and 0.15 standard deviations in reading and math, respectively, into years of learning using the average annual gains reported for grades 4-6, assuming a 10-month school year, in: Carolyn J. Hill, Howard S. Bloom, Alison Rebeck Black, and Mark W. Lipsey, "Empirical Benchmarks for Interpreting Effect Sizes in Research," Child Development Perspectives 3, no. 2 (2008): 172-177.
3. Raj Chetty, John N. Friedman, and Jonah E. Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-added and Student Outcomes in Adulthood" (NBER Working Paper, 2013). http://www.nber.org/papers/w19424.pdf.
4. *Race to the Top Fund*, U.S. Department of Education, http://www2.ed.gov/programs/racetothetop/awards.html; ESEA Flexibility, U.S. Department of Education, http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html.
5. States fully implementing by the 2012-13 school year (in addition to the District of Columbia) were Delaware, Florida, Indiana, Louisiana, North Carolina, Rhode Island, Tennessee, and Virginia. We categorize New York as a partial implementer because New York City, which accounts for approximately one third of teachers in New York State, did not fully implement by 2012-13. All other states were still in the process of establishing new systems, were piloting their systems in select districts only, or had delayed implementation of one or more components of their systems. Kathryn M. Doherty and Sandi Jacobs, "Connect the Dots: Using Evaluations of Teacher Effectiveness to Inform Policy and Practice," (National Council on Teacher Quality, 2013); U.S. Government Accountability Office, "Race To The Top: States Implementing Teacher and Principal Evaluation Systems despite Challenges," No. GAO-13-777 (2013).
6. New York Assembly Bill 11171 (New York State Senate, 2009), http://open.nysenate.gov/legislation/bill/A11171.
7. Reliability is the stability of evaluation scores for the same teachers across multiple years. Validity is the ability of the evaluation system to correctly predict teacher performance in future years.
8. Weisberg et al, *The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness* (New York: The New Teacher Project, 2009). http://widgeteffect.org/downloads/TheWidgetEffect.pdf.
9. Douglas N. Harris. *Value-Added Measures in Education: What Every Educator Needs to Know* (Cambridge: Harvard Education Press, 2011).
10. Dan Goldhaber and Michael Hansen, "Is it Just a Bad Class? Assessing the Long-term Stability of Estimated Teacher Performance," *Economica* 80, no. 318 (2013): 589–612.

11. Glazerman et al. "Evaluating Teachers: The Important Role of Value-Added," (The Brookings Institution, 2010). http://www.brookings.edu/research/reports/2010/11/17-evaluating-teachers.
12. Some states administer tests in additional subjects, such as science, and some districts have their own tests in subjects not covered by state assessments. The percentage of teachers for whom value-added can be calculated ranges from 15 to 35 percent in our four districts.
13. "Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study" (Bill & Melinda Gates Foundation, 2013). http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.
14. Mark Elhert et al., "Selecting Growth Measures for School and Teacher Evaluations" (Working paper, Department of Economics, University of Missouri–Columbia, 2013), http://economics.missouri.edu/working-papers/2012/WP1210_koedel.pdf.
15. Charlotte Danielson, "Evaluations that Help Teachers Learn," *Educational Leadership* 68, no. 4 (2011): 34-39.
16. There is a strong association between the quintile of incoming student achievement level and the quintile of teacher observation score (Chi square = 152.02, $p < 0.0001$).
17. The statistical association between the quintile of incoming student achievement and the quintile of teacher observation score adjusted for student demographics is statistically significant (Chi square = 34.68, $p < 0.005$), but much weaker than when student demographics aren't accounted for. The demographic adjustment also reduces the correlation between observation scores and value-added in the adjacent year from 0.20 to 0.09 (but the latter estimate is still statistically significant from zero at the five percent level).
18. We are able to calculate results for a district with enrollment at the 90th percentile of all U.S. districts because our medium-sized urban district is still among the largest districts in the country.
19. New York State Department of Education, "Guidance on New York State's Annual Professional Performance Review for Teachers and Principals" (2014). http://www.engageny.org/sites/default/files/resource/attachments/appr-field-guidance.pdf.
20. "Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study" (Bill & Melinda Gates Foundation, 2013). http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Brown Center on Education Policy
The Brookings Institution
1775 Massachusetts Ave., NW
Washington, DC 20036
Tel: 202.797.6090
Email: browncenter@
brookings.edu

Editing, Production, and Layout
Liz Sablich

ABOUT BROOKINGS

The Brookings Institution is a private nonprofit organization devoted to independent research and innovative policy solutions. For more than 90 years, Brookings has analyzed current and emerging issues and produced new ideas that matter—for the nation and the world. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.

ABOUT THE BROWN CENTER ON EDUCATION POLICY

Raising the quality of education in the United States is imperative for society's well-being. With that goal in mind, the purpose of the Brown Center on Education Policy at Brookings is to carry out research and analysis that informs decision-making on policies intended to impact American education. For more information, see our website: www.brookings.edu/about/centers/brown.