BROWN CENTER on
Education Policy
at BROOKINGS

April 26, 2011

Passing Muster:
Evaluating Teacher Evaluation Systems

Yellow Dog Productions

The Brookings Brown Center Task Group on Teacher Quality

Steven Glazerman, Mathematica Policy Research
Dan Goldhaber, University of Washington
Susanna Loeb, Stanford University
Stephen Raudenbush, University of Chicago
Douglas O. Staiger, Dartmouth University
Grover J. Whitehurst, The Brookings Institution

With the assistance of
Michelle Croft, The Brookings Institution

EXECUTIVE SUMMARY

U.S. public schools are in the early stages of a revolution in how they go about evaluating teachers. In years past there was little more than intuition and anecdote to support the view that teachers vary in their quality. The little data that was available came from ratings of teachers carried out by school principals, a process that typically resulted in nearly all teachers receiving uniformly high ratings. It is nearly impossible to discover and act on performance differences among teachers when documented records show them all to be the same.

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. The latter indicator, test score gains, typically incorporates a variety of statistical controls for differences among teachers in the circumstances in which they teach. Such a measure is called teacher value-added because it estimates the value that individual teachers add to the academic growth of their students.

The impetus to improve teacher evaluation systems comes from federal and state policymakers but the design falls to 16,000 independent school districts.

Value-added has a prominent role in new evaluation systems for several reasons, including a burgeoning research literature that demonstrates that value-added measures predict future teacher ability to raise student test scores better than principal ratings and teacher attributes such as years of experience or advanced coursework. Further, federal law and policy in the George W. Bush and the Obama administrations has incentivized states to develop the assessment systems and databases that allow value-added to be calculated, and to incorporate value-added information as a significant factor in evaluating teacher performance. For example, a commitment to developing new teacher evaluation systems incorporating value-added information was required of states competing for the billions of dollars of Race to the Top funds that were available under the American Recovery and Reinvestment Act during the first year of the Obama administration.

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 falls to the roughly 16,000 school districts and 5,000 independent public charter schools in the U.S. that have the responsibility for developing human resource policies and procedures for their instructional staff. 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—even as most move to new systems that are intended to be more informative than those used in the past.

If an individual state or the federal government wishes to require or incentivize local education agencies to evaluate teachers more rigorously and meaningfully, how can they do so while honoring each district's authority to do it its own way? And how can individual school districts benchmark the performance of their teacher evaluation system against the performance of evaluation systems in other districts or against the previous version of their own evaluation system? In other words, how can teacher evaluation systems be compared, one to another?

This report addresses the comparison of teacher evaluation systems in the context of a particular administrative and legislative challenge: How a state or the federal government could achieve a uniform standard for dispensing funds to school districts for the recognition of exceptional teachers without imposing a uniform evaluation system on those districts. We address and provide practical procedures for determining the reliability of local teacher evaluation systems. We then demonstrate that the reliability of the evaluation system determines the proportion of teachers that a system can identify as exceptional. Thus a school district wanting to accurately recognize the top quartile of teachers as highly effective would only be able to identify some portion of the top 25 percent with confidence given the lack of perfect reliability in the measures of teacher effectiveness that are deployed by the district. Further the portion of the top quartile that could be identified would be greater in an identically sized district that has more reliable measures.

In our approach, school districts with more reliable evaluation systems would be able to identify a greater proportion of their teachers as exceptional.

Our approach to answering the question of how the federal government or a state could dispense funds to local districts to reward exceptional teachers is that the amount of funding should be scaled to the reliability of each district's evaluation system. Thus school districts with more reliable systems would be able to accurately identify a greater proportion of their teachers as exceptional and would receive funding in line with those numbers. The procedures we propose by which the reliability of a district-level teacher evaluation system could be reported to and evaluated by state or federal officials are straightforward and simple. They do not necessitate an intrusive federal or state role.

Although we provide a worked solution to a specific administrative challenge, i.e., state or federal funding for district-level recognition programs for exceptionally effective teachers, the underlying approach we offer has more general uses in a variety of circumstances in which decisions have to be made about teachers based on imperfect data. For example our approach is easily adapted to the district-level task of identifying low-performing teachers for intensive professional development or to the state- or federal-level task of setting minimal standards for the reliability of teacher evaluation systems. Further, by demonstrating how the reliability of non value-added measures of teacher performance such as classroom observations is an important component of the overall reliability of a teacher evaluation system, our approach provides a spur to the development of multi-faceted methods for evaluating how well teachers are doing their jobs.

Background

The vast majority of school districts in the U.S. presently use teacher evaluation systems that result in nearly all teachers receiving uniformly high ratings. For instance, a recent study by The New Teacher Project of twelve districts in four states revealed that more than 99 percent of teachers in districts using binary ratings were rated satisfactory whereas 94 percent received one of the top two ratings in districts using a broader range of ratings.[1] As Secretary of Education Arne Duncan put it, "Today in our country, 99 percent of our teachers are above average."[2]

The reality is far different from what these evaluations would suggest. We know from a large body of empirical research that teachers differ dramatically from one another in effectiveness. The failure of today's evaluation systems to recognize these differences means that human resource decisions are not as productive or fair as they could be if they incorporated data that meaningfully differentiated among teachers. To put it plainly, it is nearly impossible to act on differences between teachers when documented records show them all to be the same.

A new generation of teacher evaluation systems seeks to make performance measurement and feedback more rigorous and useful. As such, the measures should demonstrate meaningful variation that reflects the full range of teacher performance in the classroom. New evaluation systems typically incorporate several sources of information on teacher performance. For example, the Hillsborough County Public School District in Florida utilizes classroom observations of teacher performance, student ratings of their teachers, direct assessments of teacher knowledge, and student test score gains in each teacher's classrooms as components of their teacher evaluation system.[3] The District of Columbia Public Schools evaluates teachers based on student test score growth in each teacher's classroom, a classroom observation measure, a rating of commitment to the school community, student test score gains for the whole school, and a measure of professionalism that takes into account factors such as unexcused teacher absences and late arrivals.[4]

Many of these new systems incorporate student test score growth in ways that aim to capture the contribution teachers make toward student achievement. This contribution is often referred to as teacher value-added. There are various methods for estimating teacher value-added, but all typically entail some variant of subtracting the achievement test scores of a teacher's students at the beginning of the year from their scores at the end of the year, and making statistical adjustments to account for differences in student learning that might result from student background or school-wide factors outside the teacher's control. These adjusted gains in student achievement, also known as value-added, are compared across teachers.

The prominence of value-added in new evaluation systems is a result of several influences. Among them is the commonsensical view that because the principal role of teachers is to enhance student learning, a central measure of their job performance should be how much their students learn. Another influence is a burgeoning research literature on teacher classroom effectiveness that has focused on value-added measures and demonstrated that those measures predict future teacher performance, as measured by value-added in subsequent years, better than teacher attributes such as years of experience or advanced coursework.

The broader reform community in education has taken up the cause of meaningful teacher evaluation grounded in value-added measures of effectiveness. Incentives for school districts to evaluate teachers based on value-added are central to the Teacher Incentive Fund that was authorized and funded during the George W. Bush administration and also to the Obama administration's proposed replacement, the Teacher and Leader Innovation Fund. Further, the Obama administration made a state's commitment to measuring teacher performance using value-added a requirement for success in the competition for $4.3 billion in the Race to the Top fund. The appeal of teacher value-added measures is further strengthened by their wide availability as a result of the No Child Left Behind Act's requirement for testing of all students in reading and mathematics in grades 3-8 coupled with federal funding to states to develop longitudinal data systems to serve as central state repositories of the resulting student assessment data.

Value-added measures of teachers' contribution to student growth have come to be central to popular and powerfully driven efforts to improve U.S. public schools.

The availability of student test scores allows, within each state, a common and face-valid yardstick for measuring teacher effectiveness across all schools in the state—an attribute that is not present for the other presently available sources of information on teacher effectiveness that individual school districts might employ, e.g., supervisor ratings or classroom observations. Thus for a variety of reasons value-added measures of teachers' contribution to student growth have come to be central to popular and powerfully driven efforts to improve U.S. public schools.

Researchers have pointed out that value-added estimates for individual teachers fluctuate from year to year and can be influenced by factors over which the teacher has no control. We have previously issued a report that describes some of the imperfections in value-added measures while documenting that: a) they provide one of the best presently available signals of the future ability of teachers to raise student test scores; b) the technical issues surrounding the use of value-added measures arise in one form or another with respect to any evaluation of complex human behavior; and c) value-added measures are on par with performance measures in other fields in terms of their predictive validity.[5] Our report recommended the use of value-added measures as a part of teacher evaluation but in the context of continuous improvement of those measures and awareness of their imperfections and limitations.

The present report offers advice on how to determine the degree to which an evaluation system is successful in the face of those imperfections and limitations. We address the connection between the reliability of an evaluation system and its ability to accurately identify exceptional teachers for special action, e.g., for a salary bonus if they are exceptionally good or for remedial action or dismissal if they are exceptionally bad. Reliability is not the only issue arising from the use of value-added measures. In particular, designers of evaluation systems and policymakers have to address biases that are introduced by differences in the contexts in which different teachers work. However, in this report, we focus on the issue of reliability.

We build our presentation around a proposal we put forth in a previous report, *America's Teacher Corps*, calling for the creation of national recognition for teachers deemed effective based on approved state and local evaluation systems.[6] The three design features of that proposal were:

- Promoting the use of teacher evaluation systems to identify and reward excellence

Whereas most of the focus of teacher evaluation systems using value-added has been on the identification and removal of ineffective teachers, we believe that such systems can also have a major impact by identifying and promoting excellence through recognition of exceptionally strong classroom performance.

- Flexibility on the components that would need to be a part of a teacher evaluation system and how those components would be weighted

There is no consensus on the degree to which teacher performance should be judged based on student gains on standardized achievement tests. Supporters of test-based measures would seek to expand standardized testing to virtually all grades and subjects and weight the results heavily in personnel decisions about teachers. Opponents question the validity of state assessments as measures of student learning and the accuracy and reliability value-added indicators at the classroom level. They typically prefer observational measures, e.g., ratings of teachers' classroom performance by master teachers. Our proposal for a system to identify highly-effective teachers is agnostic about the relative weight of test-based measures vs. other components in a teacher evaluation system. It requires only that the system include a spread of verifiable and comparable teacher evaluations, be sufficiently reliable and valid to identify persistently superior teachers, and incorporate student achievement on standardized assessments as at least some portion of the evaluation system for teachers in those grades and subjects in which all students are tested.

- Involving a light hand from levels of government above the school district

A central premise of our previous report is that buy-in from teachers and utilization of their expertise are most likely if the design of an evaluation system occurs at a level at which they feel they have real influence. In most cases this will be the local school district where they work. We expect wariness from teachers, even with respect to a system intended only to identify and reward excellence, if the design of that system is subject to considerable control from Washington or the state level. Further, we doubt that there is much of an appetite within Congress for the creation of a federal bureaucracy devoted to the fine-grained oversight of state and local teacher evaluation systems. And we doubt there is sufficient capacity within state-level education bureaucracies to carry out such oversight even if there is a political will to do so.

Suppose a state or the federal government wanted to fund a program whereby individual school districts could provide a bonus or other rewards to their exceptionally effective teachers. This requires a system of evaluation that meaningfully differentiates among teachers based on their performance. Similarly, suppose that a state wanted to encourage districts to differentiate the teaching profession so that new teachers started with one set of responsibilities but could be promoted into more complex and challenging roles as they demonstrated capability in the job. This reform, again, requires evaluations to determine different levels of teaching performance. Given the great variation in design and quality of district evaluation systems and the practical and political constraints on states or the federal government producing uniformity in those systems, how could state or federal funds for such recognition programs be fairly distributed?

Given the great variation in the quality of district evaluation systems, how could state or federal funds be fairly distributed to recognize exceptional teachers?

In this report we address the question of how a state or the federal government could achieve a sufficiently uniform standard for dispensing funds for the recognition of exceptional teachers without imposing a uniform evaluation system on participating school districts. In particular, we address the role of the state or federal government in assessing the reliability of local evaluation systems. We demonstrate that the quality of the measures and the quantity of data affect reliability and determine the number of teachers a system can identify as exceptional. Instead of a school district wanting to recognize the top quartile of teachers being able to identify 25 of every 100 teachers as being in the top 25 percent, we show that when imperfections in the measurement system are taken into account, only some portion of the true top 25 percent can be identified with confidence. Further, that portion would be greater in an identical sized district that has better measures and more data.

Although we provide a solution to what may seem to be a narrowly-focused administrative challenge, i.e., funding a teacher recognition program from the state or federal level, the underlying approach we offer has more general uses to which we will turn in the final section of this report.

How state or federal teacher recognition programs can accommodate district evaluation systems of differing quality

A major source of debate about the methods of estimating teacher performance is the statistical reliability of such measures and whether they are sufficiently precise to support attaching consequences to them such as pay for performance and tenure decisions.[7]

Our concern in this report is with the reliability of the evaluation system as a whole. We focus on the information that is necessary to determine the extent to which teacher evaluation systems are likely to result in classification errors (e.g., classifying teachers as highly effective when they are not or failing to classify them as highly effective if they are). For this discussion we will not address the potential problem of systematic bias in which the evaluations for some teachers are systematically too high or too low in comparison to the teachers' true effectiveness. Clearly, a desirable evaluation system will adjust for differences in the classrooms and schools in which teachers work to reduce or eliminate such biases. We focus here primarily on the reliability (or precision) of the estimates.

No evaluation approach will be exact for all teachers and thus designers and those using the evaluations should consider the implications of the imprecision for the decisions they make. If designers were to dismiss any evaluation systems that had error in identification, they would have to dismiss all possible systems and end up with no evaluation at all. Given that error is a fact in evaluation, understanding the implications of this error and how error varies across different approaches to evaluation can be helpful in choosing an effective approach.

In what follows, we describe how policymakers can determine the number of teachers that would accurately and inaccurately qualify to be singled out for special treatment given the power of the system to predict future teacher performance and the level of teacher exceptionality that is the criterion for special treatment. We will describe how to estimate the extent of misclassification, as well as the average difference in later effectiveness between groups identified by the evaluation. We also address how the tolerance that policymakers are willing to permit for misclassification plays a role in the number of teachers that can be accurately identified as exceptional. These subjects go to the heart of the issue of the performance of district-level teacher evaluation systems relative to each other, and provide the basis for a solution to the challenge of building a fair system for distributing state or federal funds to support district-level programs for teacher recognition.

We describe how policymakers can determine the number of teachers that would accurately and inaccurately qualify to be singled out for special treatment given the power of the system to predict future teacher performance and the level of teacher exceptionality that is the criterion for special treatment.

The factors that influence the accuracy of teacher identification systems

Using teacher performance measures to identify teachers for special treatment is, fundamentally, an exercise in prediction. For example, the use of measures of past performance of novice teachers to decide who will be tenured assumes that the better-performing novice teachers will be better teachers after receiving tenure than would the lower-performing teachers had they been given tenure. Likewise, the common district-level practice of selecting a small number of teachers as "master teachers" to serve as role models and supports for beginning or struggling teachers involves the implicit assumption that those teachers are persistently high performers who will continue to be stars in the classroom. Thus the use of teacher evaluation measures to identify different levels of teacher performance in one period as a basis for personnel action nearly always assumes that identification in one period signifies something about how teachers will perform in the future.

Our approach to judging the relative performance of teacher evaluation systems rests on determining their ability to predict future performance. We propose to judge teacher performance measures based on the degree to which they accurately estimate future teacher performance from past years of teaching, i.e., how reliable they are as a predictor of future effectiveness.

A more reliable measure is one that will yield similar answers when it is used to take more than one reading of the same phenomenon. We use the correlation of value-added measures from one period to the next as one component of a gauge of reliability, but the degree to which a performance measure in one period predicts performance in the future will depend on both the degree to which the *measure* is related to true performance and the extent to which *true performance* is stable from one period to the next. Differences between the measurement of performance and true performance are considered measurement error. If there is a significant amount of measurement error such that the performance measure is only loosely related to true performance, we would refer to it as "noisy." If on the other hand true performance changes from one period to the next, even a perfect measure of performance in one period will not accurately predict performance in the next period. There is no reason to think that true teacher performance is completely stable from one period to the next, e.g., teachers who are quite effective one year may encounter problems at home or changed work conditions that lead them to be less effective in a subsequent year, and teachers may become more effective over time as a result of experience (learning on the job) and professional development.

For an evaluation system to be useful, the true performance of teachers must be sufficiently stable over a period of a few years for predictions of future performance from past performance to be worthwhile. This assumption is buttressed, in the case of value-added measures, by the fact that value-added measures from one period predict student achievement in future periods.[8] It is also buttressed by anecdotal evidence that some teachers are simply more effective than other teachers and, as a result, parents work to get their children into these teachers' classrooms. For this discussion, we will assume that true performance, while variable from year to year, is stable enough for there to be meaningful differences in the average effectiveness of teachers over time.

In addition to picking up variation in true performance from year to year, any measure of performance will have error, i.e., will be an imperfect reflection of true teacher effectiveness. However, while all measures have error, some measures are likely to capture enough of the true differences in teacher effectiveness to be useful. Indeed, the same studies that permit an inference that there is stability in the true performance of teachers also demonstrate that the measures used in those studies are sufficiently reliable to capture at least some of those true performance differences.

Because we can neither know the precise degree of error in a given measure of performance nor the actual stability of true teacher ability from one period to the next, the correlation of a school district's measures of teacher performance from one period to the next cannot be judged against an absolute standard. Thus, our approach to judging the quality of evaluation systems must be relative: if we use common yardsticks we can demonstrate that some evaluations systems are more reliable than others and by what degree.

Value-added as the common yardstick

If we are to judge the quality of teacher evaluation systems relative to each other, there must be some common measure across those systems that is sensitive to true differences in teacher performance. Without such a common measure, the quality of teacher evaluations systems cannot be meaningfully compared across districts.

If we are to judge the quality of teacher evaluation systems relative to each other, we have to have a common measure or a set of common measures across those systems that are sensitive to true differences in teacher performance.

The focus of this paper is on using such a common measure to assess the reliability of evaluation systems. However, it is important to keep in mind that reliability is not useful unless a measure also has validity. To produce valid scores a measure must pick up differences in teacher performance that are important to student learning. Thus, while a teacher's height is strongly correlated from one year to the next, can be measured precisely, and is available for every teacher in the country, it would not be a good common measure for our purpose because it does not capture teacher performance. Similarly, suppose district A's evaluation system produced scores for individual teachers based on a weighted average of years of teaching experience, route into teaching, certification status, receipt of advanced degrees, and principal ratings of performance on a pass-fail system. The correlation of these scores for the district's teachers from one year to the next would be high, i.e., they would be very reliable. Suppose that district B deployed a very different evaluation system based on classroom observations, value-added test scores, and student surveys. The year-to-year correlation of evaluation scores would likely be much lower for district B than for district A. However, that would not mean that district B had the weaker evaluation system. In fact, the measures used by district A have been shown empirically to be only weakly related to classroom performance whereas those used by district B have a stronger evidence base. A system for determining whether a district's evaluation system passes muster in terms of recognizing exceptional teacher performance should not be designed to favor year-to-year reliability disconnected from what is being measured.

If we are to judge the quality of teacher evaluation systems relative to each other, we have to have a common measure or a set of common measures across those systems that are sensitive to true differences in teacher performance. Without such common measures, it is difficult to meaningfully compare the quality of teacher evaluations systems. A number of different measures of teacher effectiveness have at least basic face validity for measuring teacher effectiveness, including: direct measures of teaching such as teachers' scores on observational protocols of teaching quality; measures of student learning while in a teacher's classroom such as value-added measures; principals' ratings; and survey-based assessments of teachers by students and parents.

Currently value-added measures are, in most states, the only one of these measures that is available across districts and standardized. As discussed above, value-added scores based on state administered end-of-year or end-of-course assessments are not perfect measures of teaching effectiveness, but they do have some face validity and are widely available. In our analysis below we use value-added as the metric for comparing the quality of evaluation systems; however, we are limited in our goal of comparing systems by the limitations of these measures. As other measures become widely available and as the tests on which the value-added measures are based become better aligned with societal goals, our ability to judge and compare systems of evaluation will improve

Note that we do not recommend that states or the federal government be prescriptive about the components that districts should include in their teacher evaluation systems or how they should be weighted or how the information from those systems should be used for high-stakes personnel decisions. We use growth in student achievement scores on standardized assessments as an outcome measure to judge the relative quality of teacher evaluation systems but should other standardized measures of teacher performance that have clear face validity come to be used in public schools statewide in the future, they could augment or replace the use of standardized achievement test gains without changing the conceptual and computational model we put forward. For example, one could imagine measures of student engagement or assessments of so-called soft skills finding their way into statewide assessments of student outcomes.

An approach based on relative strength of prediction

The goal of this report is to outline an approach for states to use in judging district-developed teacher evaluation systems. Better systems of evaluation will have a number of features—five that we outline here. ***First***, they will differentiate teachers. If there is little differentiation between teachers' evaluations (e.g., if 95 percent of teachers receive the same rating) then the system will not be useful. ***Second***, this differentiation should not be driven by observable characteristics that are unlikely to be strong predictors of effectiveness. For example, if teachers were ranked based on their years of experience, there would be differentiation but that differentiation would not have face-validity as a measure of effectiveness, even though we might imagine reasons (e.g., ease) that could lead to school leaders differentiating teachers based on experience. ***Third***, the evaluations should be predictive of future evaluations of effectiveness. That is, the system should be able to identify teachers who not only performed well in the current year but who continue to perform well in subsequent years. ***Fourth***, the system should employ multiple measures that include not only value-added measures on state-level standardized tests but formal or informal observation measures and/or other measures of student progress. This need for multiple measures arises, at least in part, from the measurement error in value-added measures and the low proportion of teachers for which value-added measures are available. It also arises from the importance of goals for students that are not captured by available tests. ***Finally***, the system should be applicable to all teachers. It is unlikely that all measures used in the evaluation system will apply to all teachers. For example, teachers serving some groups of special education students may not have appropriate value-added from student achievement measures because the tests do not sufficiently capture their contribution. However, even though these teachers do not have all the metrics, a good system will provide enough alternative metrics to reliably assess these teachers. In what follows, we concentrate on the fourth and fifth features of a strong evaluation system.

A system based only on value-added measures would not be helpful for identifying the most effective teachers in a district.

Value-added measures have the benefit of being based on student outcomes. If we believe that a teacher's effectiveness is, by definition, the contribution that he or she makes to students, then a student performance measure is ideal. However, there are clear drawbacks with current value-added measures. In particular, they are usually based on student outcomes over a very narrow set of domains; they have substantial measurement error; and they are usually only available for a small subset of teachers. As a result, a system based only on value-added measures would not be helpful for identifying the most effective teachers in a district.

In order for districts to develop useful systems of evaluation they will need meaningfully to supplement the state test score data. One approach to this supplementation would be to develop other measures of student performance that more fully capture the range of outcomes that the district cares about and provides information on the learning in classrooms of a greater share of the district's teachers. A second approach to this supplementation is to collect data other than student performance data that gives insights into teaching effectiveness. This data could include formal or informal assessments from school leaders or parents or students as examples. In a system that collects these three types of information, some teachers are likely to have data on all three types of measures—state value-added, additional value-added, and non-student-based measures—while other teachers may only have data on a single type of measure.

The measures based on information aside from the state tests have advantages. For example, non test-based measures are likely easier to collect for a wider range of teachers. In addition, they may provide direct information for teachers on how to improve. On the other hand, these measures are not directly linked to student benefits, which is the ultimate goal of teaching. Similarly, student performance-based measures that use information other than performance on the state tests allow a fuller coverage of content areas than do state test-based measures but may be based on invalid or unreliable instruments.

We see a two-pronged approach for how states judge the non state-test measures utilized by a district's evaluation system. In the first prong, states may approve a measure or a type of measure for use as a core part of the evaluation system. For example, the state may approve any value-added measures based on student test performance, even those not using the required state exams. Alternatively, a state may choose a common observational protocol collected in a standard way to be used as a core part of any evaluation system. Yet, even with state allowance of measures other than value-added on the state tests, each district's evaluation system will likely utilize a range of other measures such as principal evaluation. Whether a district's evaluation system passes muster for the purpose of identifying effective teachers will depend in many cases on districts demonstrating the validity of these non state-test measures. We propose an approach that uses measures of teacher effectiveness for teachers that have both core (e.g., value-added on state tests) and non-core (e.g., principal evaluations) elements to validate the full range of measures in a teacher evaluation system for a full range of teachers.

For the rest of this report we will treat value-added measures as core measures for the purpose of assessing the reliability of the district system. However, policymakers may choose other approaches that either do not allow any non state-test measures of value-added or that allow some non state-test value-added measures to serve as the benchmark for assessing the reliability of the other measures used in the evaluation system.

Our model addresses the use of non value-added measures by examining how they perform in the grades and subjects in which both value-added and non value-added evaluation data are collected. We then extrapolate the findings from non value-added measures in the grades in which they are combined with value-added measures to their use in the grades and subjects in which value-added data cannot be calculated. Because the larger proportion of most districts' teachers teach in untested grades and subjects, the performance of the non value-added components of a district's teacher evaluation system has a strong influence on the number of highly-effective teachers who can be accurately identified as exceptional in our model. The importance of measuring teacher effectiveness in non-tested grades and subjects gives school districts a strong incentive to develop reliable and valid teacher evaluation measures in addition to value-added. These measures could be based on student performance on other tests, on other student work, and/or on the assessments of school leaders, peers, parents, or students.

The factors that influence the extent to which performance information about teachers in one period predicts their effectiveness in the future will vary from teacher to teacher and locality to locality. We propose judging the sufficiency of district-developed teacher evaluation systems by identifying just seven parameters, permitting the computation of the proportion of a district's teacher workforce that could be identified accurately for special treatment using the same rules for all districts. This proportion then becomes a metric for assessing the effectiveness of the evaluation system. Specifically, the proportion of highly-effective teachers that will qualify for special treatment will depend on a) two critical values adopted by policymakers (what we call exceptionality and tolerance), b) three correlations calculated using teacher-level data for teachers within each district, and c) a count of the teachers in each district who are subject to the full evaluation system including value-added measures and a count of the number who are subject only to the non value-added components (e.g., because they teach in untested grades and subjects).

The first parameter that we need to identify, the exceptionality parameter, identifies the point in a distribution of teacher evaluation scores that singles out individual teachers as being exceptional and thus subject to some type of special treatment. The more stringent the definition of exceptionality, the fewer teachers who will be identified for special treatment.

- ***Exceptionality*** is the cutoff in a teacher rank distribution that is used for decision-making. For example, to identify the "top 20 percent" or "bottom 5 percent" the exceptionality parameter would be 80 percent and 5 percent respectively.

As we have indicated previously, errors of measurement are endemic in nearly all assessments of complex human performance. Measures of teacher performance will not be nearly as reliable as measures of teacher height, for example. Because the degree of noise in measures of teacher performance can be quantified, policymakers are put in a position of needing to decide how tolerant they wish to be of two offsetting types of errors—over-inclusion and under-inclusion.

Some policymakers may wish to be very certain that the teachers identified as exceptional are truly exceptional. These policymakers would have low tolerance for classifying teachers as exceptional whose true performance level falls below the cutoff for exceptionality in the recognition program. These policymakers are intolerant of over-inclusion. Other policymakers may want to be very certain that errors of measurement do not result in excluding teachers from the exceptional category who truly belong there. They are tolerant of over-inclusion because they want to minimize errors of under-inclusion. Since errors of under-inclusion and over-inclusion are countervailing, policymakers can't have their cake and eat it too. They have to decide whether to equate the probability of the two types of errors or to favor reduction in one type of error at the cost of increasing the other.

Statistically, the value of the tolerance parameter that produces the lowest number of classification errors creates an equal probability of errors of over- and under-inclusion. That value is .5, and we have set that as the default in our subsequent spreadsheet calculator. There might be occasions when policymakers would want to select a different value, perhaps to align the tolerance parameter for assessing the evaluation system itself with the one used to make decisions about individual teachers within a district system. That said, we expect there will be little reason to alter the default value at the state or federal level for the purposes of determining the degree to which districts pass muster. We introduce the tolerance parameter here primarily to expose the mechanics and conceptual model underlying our calculator and to drive home the point that errors of measurement are inescapable in systems that attempt to evaluate complex human performance and require decisions on how they are to be handled.

- ***Tolerance*** represents how willing policymakers are to risk an error of over-inclusion. It is the probability that the average teacher in the group defined by the exceptionality parameter does not actually belong in the exceptional category based on his or her true performance.

The ***Correlation*** parameters are the simple correlations between a) the full evaluation scores for the teachers for whom value-added can be calculated in one year or more years (baseline) and their value-added test scores in the subsequent year (outcome); b) the evaluation scores of these same teachers calculated without the value-added component for the same baseline years(s) and their value-added scores in the outcome year; and c) the value-added scores of these same teachers in the outcome year and the year just preceding.

The first of these correlations, which we will refer to as the ***full evaluation correlation***, indicates the strength of the observed relationship between teachers' scores on a district's full teacher evaluation system and subsequent student achievement gains (value-added). Higher values are more desirable. The components in the complete evaluation system would include value-added as well as non value-added components such as classroom observations and evaluations by administrators. The components would be weighted as they are in the evaluation system that is actually deployed or will be deployed by the district. The number of years of data used as a baseline for evaluating teachers would also be the one deployed by the district. It could be a minimum of one year. The full evaluation correlation indicates the degree to which this total, composite evaluation score from one or more previous years predicts value-added for the same teachers the next year.

Imagine an Excel spreadsheet: Each row represents a teacher in a district for whom full evaluation scores are available for the most recent year for which state-mandated student assessment scores are available and for as many preceding years as the district wishes to employ in generating a baseline evaluation score. Each teacher's aggregate evaluation score for the baseline year(s) is represented in one column and that teacher's aggregate value-added score in the subsequent and most recent year is represented in the second column. The full evaluation correlation is the simple Pearson's *r* between those two columns of data.

The second correlation, which we will refer to as the ***non value-added correlation***, indicates the strength of the observed relationship between teachers' scores on the non value-added components of the district's full teacher evaluation system in the baseline year(s) and value-added in those teachers' classrooms in the outcome year.

Value-added cannot be calculated for teachers in untested grades and subjects. Such teachers represent the majority of the workforce in most districts. How can the evaluation system that is applied to these teachers be evaluated fairly across districts? The non value-added correlation addresses this question by assuming that the relationship that exists between value-added and an alternative teacher performance instrument, for example a classroom observation, for teachers for whom both measures are collected also applies in the case of teachers in subjects where value-added measures are not available. For instance, we would assume that the correlation between observationally-based ratings of teachers and value-added in math would be the same in history, where value-added measures are not available.[‡]

Operationally, the non value-added correlation is collected for the same teachers in a district for whom the full evaluation correlation is calculated. These are the teachers who are subject to the complete evaluation system for whom value-added can be calculated in the outcome year. The correlation coefficient would be calculated between the composite of the non value-added components of the evaluation score for these teachers in the baseline year(s) and the value-added component of those teachers' evaluation score for the next year, i.e., the outcome year. In order to increase the predictive power of the non value-added components, a district might choose to use more years of data on those components than it does for the value-added component. That is acceptable in our model. The only caveat is that the conditions the district sets on the data it uses to generate the correlation for the non value-added components have to be the conditions the district uses for its actual decision making system. The non value-added correlation coefficient estimates the degree to which the non value-added components of the teacher evaluation system, e.g., classroom observations and administrator ratings, predict future value-added.

The third correlation, which we will call the ***value-added correlation***, estimates the reliability of the value-added measure itself. The first two correlations express the observed predictive relationship between evaluation scores and value-added. We are expressing these relationships with simple correlations that can range between r = .00 (no association whatsoever) and r = 1.00 (teachers' value-added scores are perfectly predicted by their evaluation scores). But even if the true underlying relationship between evaluation scores in a baseline period and value-added the next were a perfect r = 1.00, we would not see anything close to that coefficient in the observed correlation unless the value-added measure itself were perfectly reliable. However, we know and have described previously many sources of noise in measures of value-added, e.g., student gain scores on standardized assessments capture imperfectly what students have learned from their teacher; teachers may improve from one year to the next. Thus the maximum values of the first two correlations (the full evaluation correlation and the non value-added correlation) are constrained by the reliability of the value-added measure itself.

[‡] This assumption could be evaluated by occasional testing of students in subjects that are not typically tested year after year.

Imagine early astronomers trying to fix the coordinates of dim stars at particular times of year using a crude telescope. Before deciding how much stars actually change their positions as observed from earth, it would be important to know the reliability of the telescope itself. If objects appear to move in the telescope based on errors of refraction in the glass itself, it would prudent to adjust for those errors before drawing conclusions about true celestial movements.

We use the correlation of value-added between the outcome year and the preceding year to estimate the reliability of the value-added measure. Just as we would want to assess the astronomer's model of annual changes in the coordinates of stars by adjusting for what the telescope can reliably detect, so too do we want to assess a district's evaluation system by asking how much it can account for of what it can detect as the persistent year to year contribution of teachers to student achievement.

We do this by adjusting upward the full evaluation correlation and the non value-added correlation to take into account the reliability in the value-added measure. For example, suppose the full evaluation correlation accounted for 20 percent of variability in the outcome year's value-added scores, whereas value-added scores from the preceding year accounted for only 40 percent of the variability in the next year's value-added scores. In this case, the full evaluation correlation would have accounted for 50 percent (20% / 40%) of the persistent relationship between value-added from one year to the next. The full evaluation correlation would be adjusted upward to reflect this. The technical details of this adjustment are described in the appendix.

Operationally, the value-added correlation is calculated on a district-by-district basis for the same teachers from whom the first two correlations are derived. We use a district-level adjustment rather than a standard adjustment for all districts because districts differ in the reliability of the value-added component of their evaluation systems for a number of reasons, including the statistical model that is used to adjust for student and school background variables, the quality of the state assessments, and the heterogeneity of the teachers for whom value-added has been calculated.

- The ***Number*** parameters are a) the number of teachers in a district's present workforce that is subject to the same full evaluation system that was used to calculate the correlation parameters, and b) the number of teachers in a district's present workforce that is subject only to the non value-added components of the district's evaluation system (the same non value-added components that were used to generate the correlation parameters).

Because the reliability of the teacher evaluation system differs for the teachers who are subject to the full system vs. only parts of the system, we need to know the number in each category to calculate the overall identification rate for exceptional teachers. Districts will typically have teachers in their workforce who do not fall into either the full evaluation or the non value-added evaluation category. For example, many novice teachers will not have had enough time on the job to be subject to the evaluation system, and part-time teachers or teachers serving primarily in administrative roles may not be subject to the system. Such teachers should not be included in calculating the number parameters, which is to say that our model for identifying exceptional teachers applies only to teachers who are subject to the evaluation system for which reliability can be calculated in our model. This means that if there are two identical districts except that one manages to include more teachers in its evaluation system than the other, then the more inclusive district will be able to identify more teachers as exceptional.

A worked example: Application to America's Teacher Corps

We have previously described an earlier report from this task group recommending a program called America's Teacher Corps (ATC). We will use some of the details in that proposal to work through an example of how the model we have proposed could be applied.

The ATC proposal recommended that the federal government provide funds to school districts to augment the salary of teachers who are in the top quartile of the performance distribution and serve in high poverty schools. Thus the exceptionality parameter in our model is set at the 75th percentile for the ATC.

The ATC proposal did not touch on the issue of errors of over- and under-inclusion. For this example we will use the 50 percent tolerance value. As we have indicated previously, this particular tolerance value equalizes the number of over- and under-inclusion errors and minimizes the total number of errors. The 50 percent value also maximizes the average difference in teachers' contribution to student achievement between teachers in each identified group, meaning that the value-added of ATC identified teachers is as large as possible relative to non-ATC teachers.

Having established the exceptionality and tolerance parameters, we can calculate how many teachers would be eligible for the ATC given the three correlation coefficients we have previously described (full evaluation, non value-added, and value-added) and the number of teachers in a district subject to the full evaluation system vs. only the non value-added components of the evaluation system. Districts will typically rely on a suite of performance measures, such as principal or peer observations and value-added. But regardless of the system they use to identify teachers, the proportion of teachers that will be ATC-eligible rises with the correlation between the performance metric in the baseline year(s) and teacher value-added in the next.

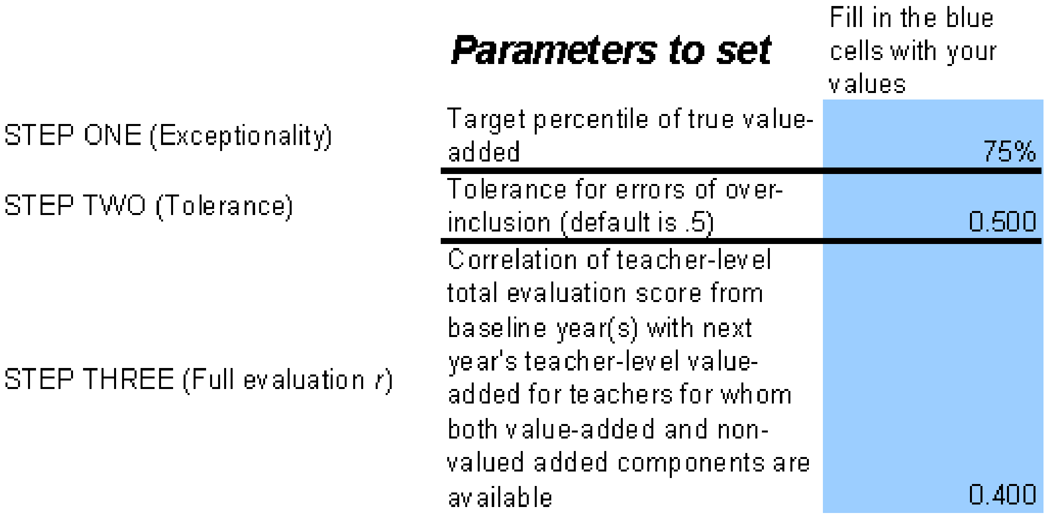
We illustrate what this might look like for actual school systems by drawing values for the required correlations from prior research. Studies of the stability of correlations of teacher evaluation measures across years find different values in different states. For example, Goldhaber and Hansen[9] find much higher correlations in North Carolina for elementary school math and reading value-added estimates than are reported by Harris and Sass[10] using Florida data. To be conservative we will use lower range estimates in our example.

The multi-district Measures of Effective Teaching project… found that teacher value-added in mathematics correlated 0.40 from one year to the next whereas the corresponding correlation in English language arts was 0.20.

The Harris and Sass report found a correlation of 0.40 between evaluation scores from two baseline years and value-added in the subsequent year for Florida elementary school teachers, where the evaluation scores combined value-added and principal evaluations. We will use that value for our full evaluation correlation. The same study found a correlation of 0.18 between principal ratings of teachers and value-added scores in the subsequent year. We will use that as our non value-added parameter. The multi-district Measures of Effective Teaching project being carried out by the Bill and Melinda Gates Foundation found that teacher value-added in mathematics correlated 0.40 from one year to the next whereas the corresponding correlation in English language arts was 0.20.[11] Later we describe why two years of baseline data are preferable to one, including that the correlations tend to be higher. But for our worked example, we will take the average of these two year-to-year correlations, 0.30, as the value for our value-added correlation.

We have developed a spreadsheet based on our model, which can be downloaded here.[§] Users enter values such as those chosen as illustrative above. An example of the spreadsheet using the values described above is presented in the next two figures. The first figure presents the portion of the spreadsheet in which values are entered by the user. The number of teachers subject to the total evaluation system vs. only the non valued-added portions of the district's teacher pool is set at 400 and 600 respectively for this example.

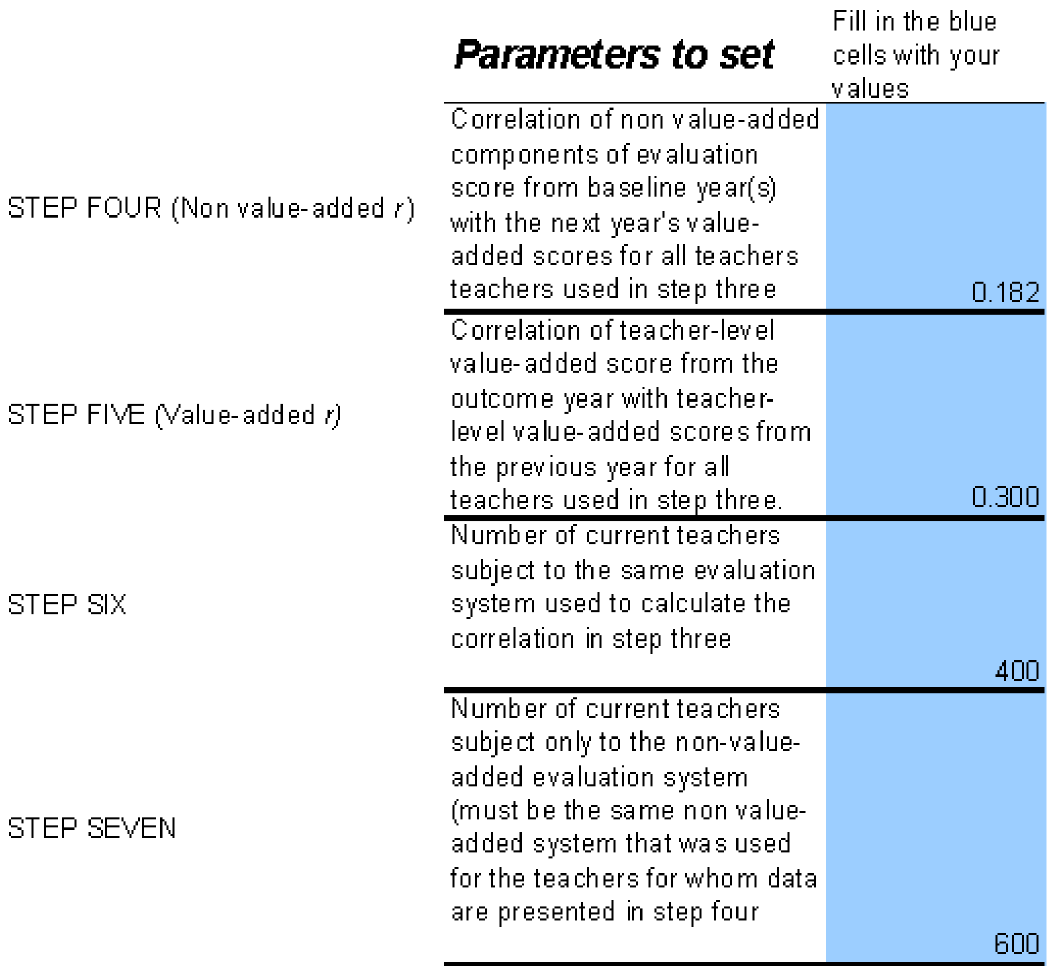
Figure 1



	Parameters to set	Fill in the blue cells with your values
STEP ONE (Exceptionality)	Target percentile of true value-added	75%
STEP TWO (Tolerance)	Tolerance for errors of over-inclusion (default is .5)	0.500
STEP THREE (Full evaluation *r*)	Correlation of teacher-level total evaluation score from baseline year(s) with next year's teacher-level value-added for teachers for whom both value-added and non-valued added components are available	0.400

[§] http://www.brookings.edu/reports/2011/0426_evaluating_teachers.aspx

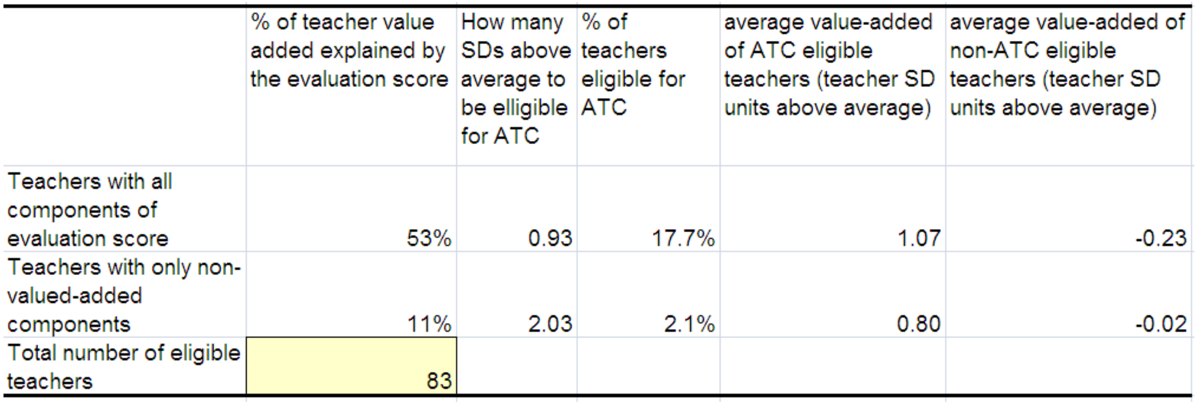
Figure 1 – Continued



	Parameters to set	Fill in the blue cells with your values
STEP FOUR (Non value-added *r*)	Correlation of non value-added components of evaluation score from baseline year(s) with the next year's value-added scores for all teachers teachers used in step three	0.182
STEP FIVE (Value-added *r*)	Correlation of teacher-level value-added score from the outcome year with teacher-level value-added scores from the previous year for all teachers used in step three.	0.300
STEP SIX	Number of current teachers subject to the same evaluation system used to calculate the correlation in step three	400
STEP SEVEN	Number of current teachers subject only to the non-value-added evaluation system (must be the same non value-added system that was used for the teachers for whom data are presented in step four	600

The next figure, Figure 2, presents the results that are calculated based on the values entered and displayed in the Figure 1.

Figure 2



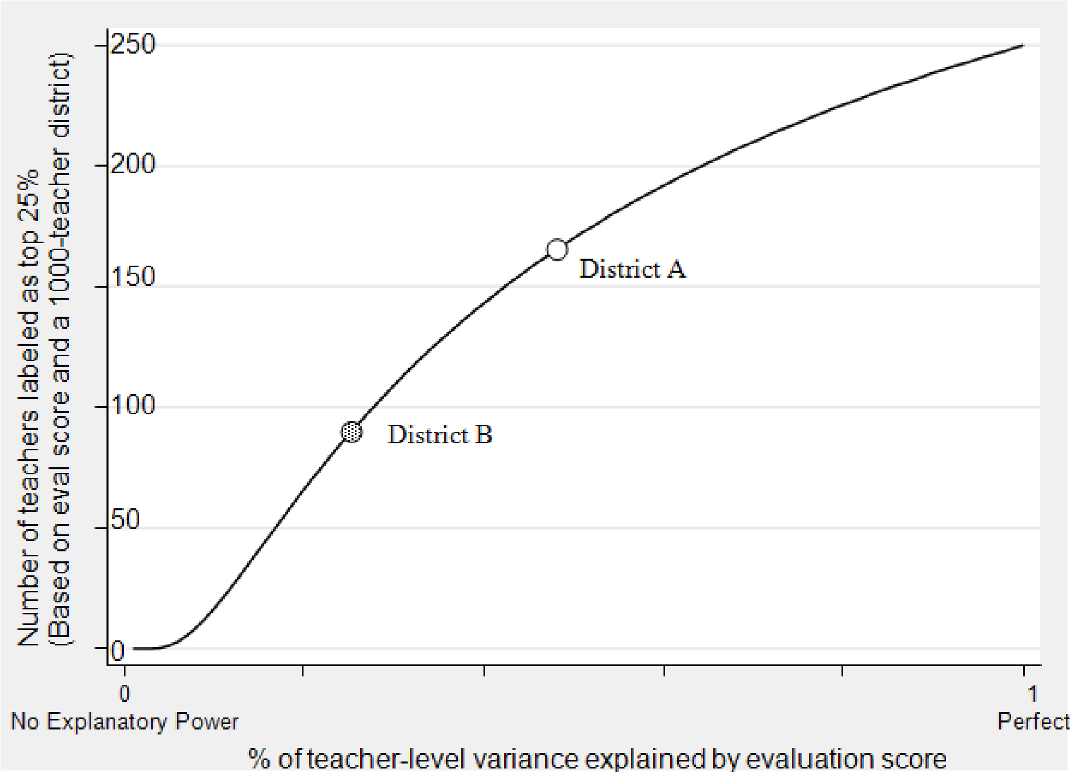
	% of teacher value added explained by the evaluation score	How many SDs above average to be elligible for ATC	% of teachers eligible for ATC	average value-added of ATC eligible teachers (teacher SD units above average)	average value-added of non-ATC eligible teachers (teacher SD units above average)
Teachers with all components of evaluation score	53%	0.93	17.7%	1.07	-0.23
Teachers with only non-valued-added components	11%	2.03	2.1%	0.80	-0.02
Total number of eligible teachers	83				

Under this scenario, this school district is capable of identifying 8.3 percent of its total workforce, i.e., 83 out of the 1000 teachers subject to the evaluation system, as sufficiently exceptional to be eligible for ATC status. This percentage is far less than the 25 percent that is set as the exceptionality parameter by policymakers. The difference is due to the unreliability of the district's evaluation system as a measure of the persistent value-added performance of its teachers. The more reliable the evaluation system (i.e., the stronger the correlation of evaluation scores in one year with the persistent teacher value-added effect), the greater the proportion of teachers who can be identified, up to the theoretical limit of 25 percent that is set by the exceptionality parameter.

This relationship of number of teachers who can be identified and the reliability of the measures is highlighted in Figure 2 by the difference in the identification rate for teachers who are subject to the full evaluation system vs. teachers who are subject only to the non value-added components (in this case just principal ratings). Recall that the full evaluation correlation is 0.40 whereas the non value-added correlation is only 0.18. These differences lead to 17.7 percent of the teachers subject to the full evaluation system being identifiable as exceptional whereas only 2.1 percent of those subject to only the non value-added evaluation are identifiable.

The relationship between measure reliability and the number of teachers who can be identified as exceptional is illustrated for the full range of values of reliability in Figure 3. Here the values on the vertical axis represent the proportion of teachers who can be identified up to the exceptionality parameter used in our ATC example, 25 percent. The horizontal axis represents the first number column in the spreadsheet in Figure 2 and represents the amount of variance in the persistent teacher value-added estimate explained by the evaluation score: the lower the number, the lower the explanatory power—the higher the number, the better the prediction.

Figure 3

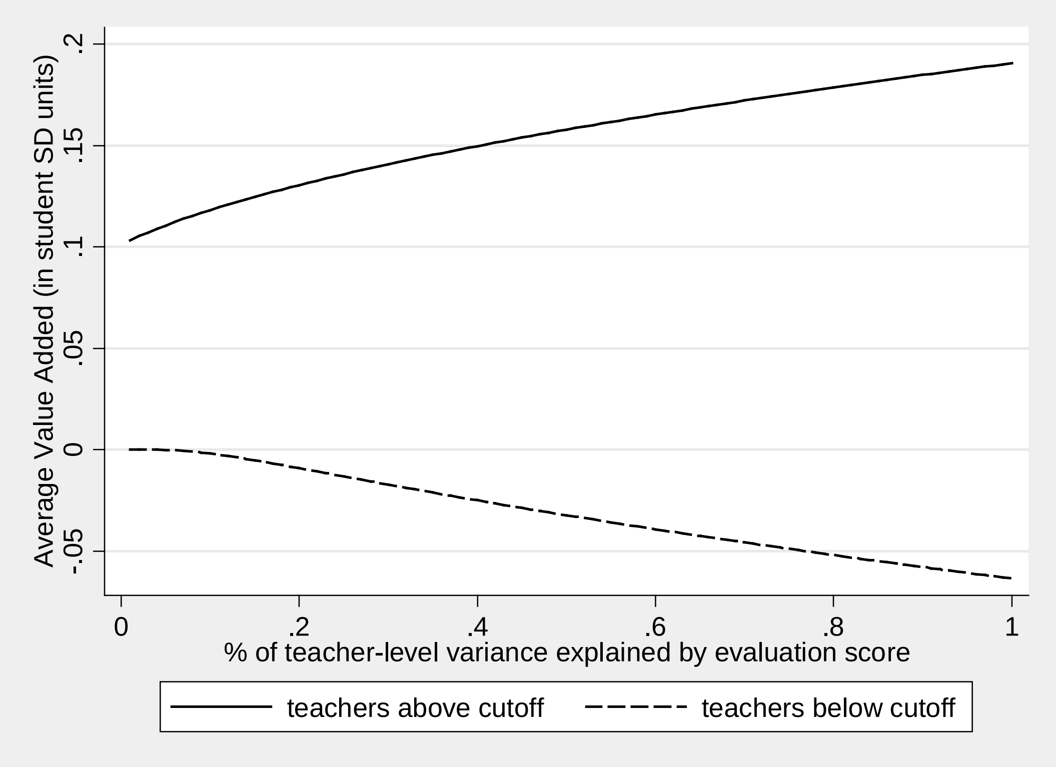


Although there is considerable measurement error in the evaluation system we are illustrating, the decisions it supports are robust under our model because the number of teachers who can be identified is directly determined by the reliability of the measures. For instance, in Figure 3 one can see that district A's evaluation system which explains 50 percent of the teacher-level variance would be able to identify approximately 160 of the district's 1000 teachers as being the top 25 percent; whereas, district B's evaluation system only captures approximately 25 percent of the teacher-level variance and would only allow approximately 90 of the 1000 teachers to be identified.

What does this mean in terms of the actual performance of those who fall into the identified or not identified categories? Note from the spreadsheet in Figure 2 that for our illustrative case the average identified teacher subject to the full evaluation system is 1.3 standard deviations more effective in raising student achievement than the average non-identified teacher (the difference between the last two columns). A standard deviation of difference among teachers in their effectiveness corresponds to about a month of student learning during one year of elementary school.[12] Thus in the scenario we have illustrated, which uses plausible values drawn from the empirical literature, the average progress for children in the classrooms of teachers in the year after their identification as exceptional teachers via the full evaluation system is equivalent to about 5 weeks more schooling than for students in the classrooms of the other teachers in the district.

This relationship is illustrated in Figure 4, which depicts the link between the degree to which the evaluation score predicts subsequent value-added and the difference in student gain produced by identified and non-identified teachers for the full range of possible values for the ATC case. Note that an evaluation system that approached a perfect correlation with the persistent component of teacher value-added would generate about a .25 standard deviation difference in student learning between identified and non-identified teachers in the year after identification. This is a difference of about one-quarter of a year of schooling.[**] This illustrates, again, the importance of developing sophisticated teacher performance evaluation systems, be they for the purpose of implementing the ATC concept or for the myriad high-stakes purposes promised, for instance, in states' Race to the Top applications.

Figure 4



[**] It is important to note that many reports of the strength of teacher effects based on value-added are based on estimates from the same year(s) of data, whereas our model and examples are based on predictions to the next year from data collected in the previous year. Correlations between value-added estimates in one year and performance in subsequent years decay over time, but are still statistically significant 9 years later. The correlations are much higher and the decay in magnitude stabilizes after 5 years if the initial estimates are based on two years rather than one year of data. Goldhaber, D. & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. (Washington, DC: CALDER, The Urban Institute).

A technical presentation of the rationale and calculations for the material in the figures and the associated spreadsheet calculator is found in the appendix of this document.

Questions and Answers

Q1. Should districts identify exceptional teachers in different proportions in the tested vs. untested grades and subjects based on the different identification rates from your model?

A1. Doing so increases the accuracy of prediction but may be undesirable for other reasons.

Districts may well not wish to identify different proportions of teachers from the tested grades and subjects vs. the non-tested grades and subjects as exceptional simply because the prediction correlations are different for these two categories. There is no conceptual reason to believe that greater proportions of elementary school teachers are exceptional than high school teachers, for example. That more elementary school teachers can be identified than high school teachers within the same margin of error is simply an artifact of value-added being available for some elementary school grades but for few high school grades and subjects.

In many districts only about 20 percent of the teacher workforce is deployed in tested grades and subjects. Providing disproportionate access to recognition and reward for teachers who happen to be subject to a district's full evaluation system could have unintended negative consequences such as incentivizing teachers to teach in the tested vs. non-tested grades and subjects or in undermining the sense of fairness in the evaluation process as a whole.

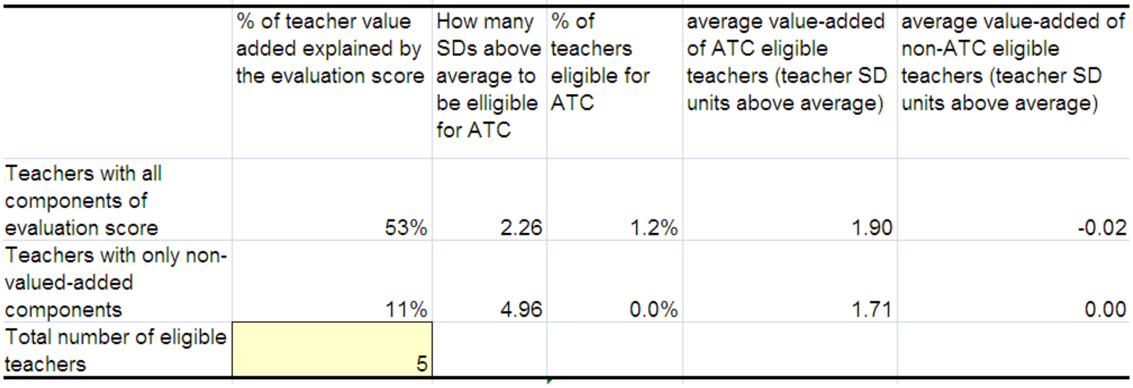
We expect that many districts would want to provide equitable access to incentive and recognition programs across all categories of teaching while recognizing that a higher rate of misclassification will occur for categories of teachers for whom the full suite of evaluation data are not available. The implicit tradeoff between equity of access and accuracy of prediction can be reduced by including a greater proportion of the teacher workforce in the full evaluation system and by including more years of baseline data for predictions for the non value-added components of the evaluation system.

The one requirement of our model is that districts use their evaluation scores within each teacher category to identify exceptional teachers. Thus a district may decide to provide a recognition program for junior teachers that recognizes a higher proportion of junior teachers than does a recognition program for more senior teachers. The reason for doing so might be to increase the retention rate of high performing junior teachers. This could be a worthwhile objective notwithstanding the fact that the full evaluation correlation is greater for senior teachers. Or a district might decide to provide equal probability of recognition to teachers in untested and tested grades and subjects despite the different misclassification rates that result. That is acceptable as we frame policy and administrative uses of our model, as long as teachers are selected in rank order within their category based on their evaluation score.

Q2. Would your model apply to selecting low performing teachers for special treatment?

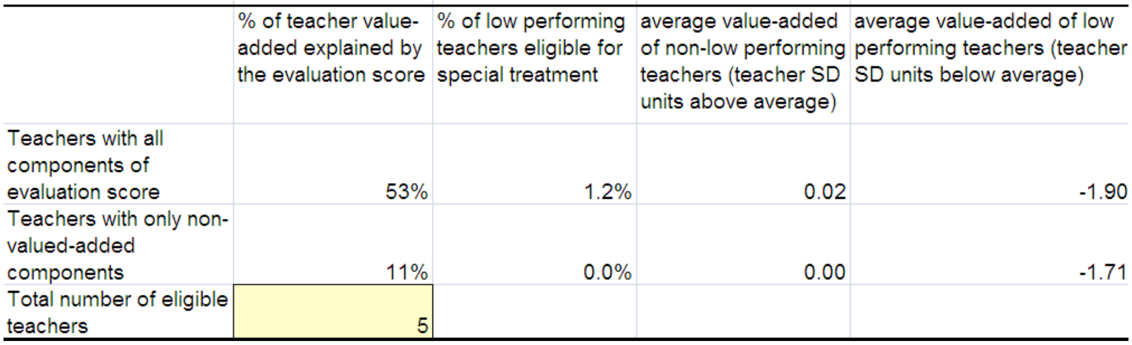
A2. <u>In general, yes, but some specifics would change.</u>

Our model and the associated calculator can be used to estimate the number of teachers who could be identified and their contribution to student value-added for either tail of the distribution and for any value of exceptionality. For example, if a district wanted to identify the lowest 5 percent of its teachers in terms of their true contribution to value-added for intensive remediation, they could use our calculator, enter 95 percent as the exceptionality parameter, and simply change the numerical signs and descriptors for the output in the last two columns, e.g., "SDs above average" to "SDs below average." Using all the other values entered into our previous worked example, but changing exceptionality from .25 to .95, generates the following output:



	% of teacher value added explained by the evaluation score	How many SDs above average to be elligible for ATC	% of teachers eligible for ATC	average value-added of ATC eligible teachers (teacher SD units above average)	average value-added of non-ATC eligible teachers (teacher SD units above average)
Teachers with all components of evaluation score	53%	2.26	1.2%	1.90	-0.02
Teachers with only non-valued-added components	11%	4.96	0.0%	1.71	0.00
Total number of eligible teachers	5				

The necessary changes to the headings and signs to make the output relevant for low performers follow:



	% of teacher value-added explained by the evaluation score	% of low performing teachers eligible for special treatment	average value-added of non-low performing teachers (teacher SD units above average)	average value-added of low performing teachers (teacher SD units below average)
Teachers with all components of evaluation score	53%	1.2%	0.02	-1.90
Teachers with only non-valued-added components	11%	0.0%	0.00	-1.71
Total number of eligible teachers	5			

This output makes clear the difficulty of reliably identifying appreciable numbers of teachers in the extreme tails of the distribution for special treatment based on their evaluation score assuming correlations similar to those we have employed as examples. In this case whereas the target is 5 percent, or 50 teachers in a district of 1000, only 5 can be reliably identified. Here is a case in which a change in the tolerance factor might be required. For example, by increasing the tolerance value from 50 percent to 75 percent, the number of teachers who could be identified rises from 5 to 22, or to roughly half of the theoretical target of 50 teachers. Increasing the tolerance for over-identification while reducing the tolerance for under-identification may be a reasonable policy decision with respect to interventions for very low-performing teachers.

Q3. What is the optimal number of years of baseline evaluation data for making selection decisions and for passing muster?

A3. There is no optimal number of years of baseline data. The advantage of additional years of data or additional data in any form in improving prediction has to be weighed against the costs of collecting such data and the practical needs of decision making.

Our worked example is based on two or three years of data: one or two baseline years of evaluation data used to predict a subsequent year's value-added. Some districts have more years of evaluation data on some teachers; some less. Inclusion of more data points for individual teachers will increase the precision of prediction. Further, using at least two years of data for the baseline period substantially reduces the biases that can occur in the assignment of students to teachers in a single year, e.g., an unusually disruptive student who depresses a teacher's ability to create good conditions for learning for all the students in the classroom. However, the practical realities of personnel decisions may require time frames that are shorter than those that are optimal for the most accurate decisions, e.g., ten years of data would allow much more accurate predictions than two years of data but districts should not wait 10 years to evaluate their teachers. We leave these design details to districts with the understanding that the more data they collect the better their predictions are likely to be, and the better their predictions the greater the proportion of evaluated teachers that can be identified as exceptional under our model.

Q4. Which model for calculating value-added do you recommend?

A4. For the purposes of this report, we do not prefer any particular value-added model.

All teacher value-added models start with the gain scores of individual students on academic assessments that are associated with the time period when those students are the responsibility of a particular teacher. A large number of statistical adjustments are then possible to try to remove influences on student achievement that are not the effect of the teacher. For example, classrooms with more economically disadvantaged students may, on average, show higher or lower rates of growth than classrooms of more advantaged students, so value-added models often adjust for that variable and other student background characteristics that reflect the mix of students in a particular teacher's classrooms. Adjustments can also be made for school characteristics, for district characteristics, for attendance, and other factors. There are typically several ways of calculating the effects of variables regardless of which variables are included in the value-added model. Further, models that are formally very similar can produce different results depending on the number of years of prior data on student achievement that are used as an adjustment for student background. There is as yet no clear consensus among the econometricians and statisticians who construct these models as to which is best. Further, models that may be superior technically can fail as management tools or face political or public challenges because they are difficult for anyone except specialists to understand.[††]

Because of the unsettled nature of the engineering of value-added models and the tension between technical strength and public transparency we do not recommend a particular value-added model. However, we believe there is now a sufficient body of research to recommend that the most important adjustment in a value-added model is the inclusion of student achievement from prior years, e.g., adjusting for how much a 6th grade teacher's students learned in 5th and 4th grade.[13] Although the design of value-added models clearly matters, we believe that for the broad purposes of passing muster, which is intended to require only a light hand from the federal or state level, it is better to leave those design decisions to responsible education authorities.

Q5. Is the passing muster model useful for anything in addition to the task of identifying exceptional teachers?

A5. Yes. There are a number of possible uses for both policy and management decisions.

We have built our narrative around federal and state programs that provide incentives to local education agencies to recognize exceptional teachers. However, the model we have put forward could be used to create conditions for funding broader education programs that are predicated on school districts having meaningful teacher and principal evaluation systems. For example, the Obama administration has identified improved teacher and principal effectiveness as a centerpiece of its plans for reauthorization of the Elementary and Secondary Education Act (ESEA).[14] We can imagine Congress building a requirement into Title I or Title II of ESEA that districts report on the reliability of their evaluation

[††] For an example of the clash between ease of understanding and technical sophistication see: Michael Winerip, "Evaluating New York teachers, perhaps the numbers do lie," *The New York Times*, March 7, 2011, p. A15.

system using our model as a foundation. With those data available for every school district in the nation, we could imagine states taking the next step and determining threshold values of evaluation reliability that districts would be expected to meet. We could also see the federal government incentivizing states or districts to reach threshold values or to demonstrate improvement in the reliability of their evaluation systems.

Technical Appendix: Standard-Setting for Value-added Based Teacher Identification Systems

This appendix provides the technical details underlying our approach to evaluating the reliability of a teacher evaluation system, and the related question of how state or federal teacher recognition programs can accommodate district evaluation systems of differing quality. We approach this question as a Bayesian decision problem: Based on an imprecise teacher evaluation, we must (1) form posterior beliefs about each teacher's effectiveness and then (2) base decisions about recognition on these posterior beliefs in a way that minimizes mistakes. The key insight is that decisions should be made based on posterior beliefs, and these beliefs are more precise when the teacher evaluation system is more reliable allowing us to recognize more teachers while making fewer mistakes.

The appendix is organized into four sections. The first section formalizes our approach using a simple model. The second section shows how posterior beliefs can be formed about teacher effectiveness, and how these relate to the reliability of the evaluation. The third section develops a simple method that can be used to estimate the reliability of the teacher evaluations from district data. The final section shows how this information can be used to select teachers for recognition in a way that minimizes mistakes.

1. A Simple Model for the Evaluation of Teacher Evaluation Systems

Our goal is to identify teachers whose effectiveness, defined as the effect they have on student test scores, is above some threshold. For this example, we will consider a threshold at the 75th percentile of teacher effectiveness, so that only the top 25 percent of teachers would truly exceed this threshold if we had a perfectly reliable evaluation. We assume that the impact of teacher (j) on student test scores is the teacher effect (μj) in the following value-added model:

Equation 1: $A_{ijc} = X_{ijc}\beta + \mu_j + \theta_{jc} + \varepsilon_{ijc}$

Where Xijc represents variables that control for student (and potentially peer) factors such as prior test scores, μj is the teacher effect, θjc is an idiosyncratic effect on student scores in a given classroom (c) of students, and εijc is a student-level idiosyncratic residual. We will assume that the teacher effect is normally distributed with mean 0 and variance σ_μ^2.

Districts cannot observe the teacher effect (μj) directly, but instead must evaluate teachers based on a variety of imperfect measures such as value-added estimates, classroom observations, etc. Suppose that a district uses whatever information they have available to create an overall evaluation score (Sj) for each teacher. This score could be formed in any way, such as based on informal reviews by principals, or formal composite measures based on multiple measures of teacher performance. We will assume that the evaluation score is continuous and normally distributed, but if this were not the case the ideas here would generalize (although some of the specifics would differ). For simplicity, we will also assume that the evaluation of all teaches is based on the same information—e.g., all teachers are evaluated based on classroom observations and value-added estimates.

2. Posterior Beliefs and the Reliability of Teacher Evaluation

We can summarize our posterior beliefs about μj after observing Sj as:

Equation 2:

$$\mu_j = S_j\alpha + \nu_j$$

Where the posterior mean is $S_j\alpha$, and the posterior variance is $Var(\nu_j) = \sigma_\nu^2$. Thus, α represents the increase in average teacher value-added associated with a one point increase in the overall evaluation score, while σ_ν^2 represents the amount of remaining variation in the teacher effect among teachers with a given score. With an assumption that teacher effects are normally distributed, the mean and variance summarize the posterior distribution (again, the ideas here would generalize without normality). Note that we are assuming the same posterior variance for everyone, i.e., that the evaluation score (Sj) provides the same amount of information for all teachers.

A simple statistic summarizes the quality of Sj in terms of how much it improves our ability to predict the teacher effect:

Equation 3:

$$R^2 = \frac{\sigma_\mu^2 - \sigma_\nu^2}{\sigma_\mu^2} = \frac{Var(S\alpha)}{\sigma_\mu^2}$$

This statistic is analogous to the teacher-level R-squared statistic, i.e., the percentage of the variance in the teacher effect on student test scores that was explained by the evaluation score (Sj). The square root of this statistic is the correlation between the evaluation score (Sj) and the teacher effect on test scores (μj):

Equation 4:

$$Correlation(S,\mu)=\sqrt{R^2}=\sqrt{\frac{Var(S\alpha)}{\sigma_\mu^2}}$$

The more strongly correlated the score is with the teacher effect, the more reliable is the evaluation. We discuss how to estimate this statistic from a district's data below.

The R-squared statistic in Equation 3 is directly related to the posterior uncertainty about a teacher's effectiveness. One can rearrange Equation 3 to write the posterior standard deviation as $\sigma_v=\sigma_\mu\sqrt{1-R^2}$. Thus, the closer the R-squared is to one, the more precise our posterior beliefs are relative to the total variation across teachers.

3. Estimating the Reliability of Teacher Evaluations from District Data

There are a number of ways to use district data to estimate the correlation between the teacher evaluation score and teacher impacts on test scores. Plugging Equation 2 into Equation 1 yields:

Equation 5:

$$A_{ijc}=X_{ijc}\beta+S_j\alpha+v_j+\theta_{jc}+\varepsilon_{ijc}$$

Note that this is just a value-added model that includes Sj as an additional regressor. In principal, one could estimate the posterior mean ($S_j\alpha$) and variance (σ_v^2) by estimating Equation 5 using hierarchical linear models (HLM), and then use these to calculate the correlation in Equation 4 (noting that $\sigma_\mu^2=Var(S_j\alpha)+\sigma_v^2$). This HLM approach could be difficult to implement in practice, since it requires student-level data, multiple classrooms per teacher (to separately identify the classroom-level effect from the teacher effect), and a teacher evaluation score that was constructed in a different year (so that it is not correlated with the errors in equation 5).

A simpler method is the following. Suppose that a district can provide 3 pieces of information for each teacher:

1. The evaluation score ($S_{j,t}$) from a given year (t).
2. A value-added estimate from the same year ($VA_{j,t}$). The value-added estimate for each teacher is the average residual over all of their students, using the residual from a value-added regression as specified in Equation 1:

$$VA_{j,t} = \frac{1}{n}\sum_{i=1}^{n}\left(A_{ijc} - X_{ijc}\beta\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mu_j + \theta_{jc} + \varepsilon_{ijc}\right) = \mu_j + error_{j,t}$$

So that the value-added estimate is equal to the teachers true impact on test scores (μ_j) plus some estimation error.

3. A similar value-added estimate from a different year ($VA_{j,s}$), where this could be the year after.

The simple correlation between the evaluation score and the teacher's value-added in the same year will be biased because both the evaluation score and teacher's value-added will be influenced by idiosyncratic factors that occurred during that year (i.e., because the evaluation score depends on this year's performance, it will be correlated with the idiosyncratic student and classroom errors in Equation 1. To avoid this bias, we use the correlation between the evaluation this year and value-added in a different year: $Corr\left(VA_{j,s}, S_{j,t}\right) = Corr\left(\mu_j + error_{j,s}, S_{j,t}\right)$. However, this correlation is the correlation between the evaluation score and the noisy estimate of the teacher effect on test scores, not the correlation with the true teacher effect on test scores as defined in Equation 4. Because of the noise in the value-added estimate, this correlation will understate the true correlation.

It is straightforward to show that to correct for this bias, we must adjust the correlation for the proportion of the variation in the value-added measure that is due to the true teacher effect, as opposed to estimation error (the reliability of the value-added measure). One simple estimate of the reliability of the value-added measure is the correlation in the value-added measure from one year to the next ($Corr\left(VA_{j,s}, VA_{j,t}\right)$). Using this estimate for the reliability of the value-added measure, we can construct an adjusted estimate of the correlation between the evaluation score and the true teacher effect on test scores using:

Equation 6:

$$Corr\left(\mu_j, S_{j,t}\right) = \frac{Corr\left(VA_{j,s}, S_{j,t}\right)}{\sqrt{Corr\left(VA_{j,s}, VA_{j,t}\right)}}$$

Thus, we can estimate the correlation between a district's evaluation measure and teacher effects on test scores based simply on information about the correlation of the evaluation measure with value-added in a different year, and the correlation of value-added from one year to the next.

4. Selecting Teachers to Minimize Mistakes

For any given evaluation system, we can calculate the posterior distribution for each teacher based on the model just described and methods that will be described below. Given the posterior distribution, how should we decide whether any teachers (and if so, how many) should be eligible for top-25 percent recognition?

The general answer to this question from a technical perspective is that eligibility should just depend on the likelihood that a teacher exceeds a specified threshold based on each teacher's posterior distribution. As discussed above, the posterior distribution for each teacher's effect (μj) is a normal distribution with mean $S_j\alpha$ and variance $\sigma_v = \sigma_\mu\sqrt{1-R^2}$. Therefore, we can use the posterior distribution to calculate the probability that each teacher is below the threshold for recognition. All that needs to be determined is the threshold above which teachers are recognized for exceptional performance (the exceptionality parameter) and the minimum probability of being below the threshold that one is willing to tolerate when identifying any teacher as exceptional (the tolerance parameter), i.e., the willingness to classify a teacher as exceptional who does not actually belong in the exceptional category based on his or her true performance.

The exceptionality parameter is a policy choice. For ATC, we have chosen to focus on recognizing teachers who are in the top quartile (top 25 percent) of all teachers in terms of raising student test scores. Therefore, we set the threshold at the 75th percentile of the distribution of teacher effects, which is equal to $0.67\times\sigma_\mu$ (based on the normal distribution).

The tolerance parameter is more of a technical choice, driven by the relative costs of making errors of omission (not rewarding a teacher who is truly in the top 25 percent) versus errors of commission (rewarding a teacher who is not actually in the top 25 percent). We chose a tolerance of 50 percent, meaning that to be selected for ATC the teacher must have at least a 50 percent chance of being in the top quartile (or, equivalently, every ATC teacher has no more than a 50 percent chance of lying below the 75th percentile of teacher effectiveness). A tolerance of 50 percent minimizes the total number of errors—the total number of teachers who are misclassified due to errors of omission or commission. If the cost of incorrectly rewarding a non-exceptional teacher were higher than the cost of incorrectly overlooking an exceptional teacher, one would want to set a lower tolerance. For example, a tolerance of 10 percent would ensure that any teacher selected for ATC had at least a 90 percent chance of having true performance above the 75th percentile. While this standard would reduce the chance of rewarding a teacher who is not actually in the top quartile, it would reward many fewer teachers and thereby increase errors of omission—omitting many more teachers whose performance was actually in the top quartile.

A tolerance of 50 percent results in a very simple selection rule: select any teacher whose posterior median (which is the posterior mean, $S_j\alpha$, for a normal distribution) exceeds the threshold for recognition. The posterior mean will not exceed the threshold for many teachers when the evaluation (S_j) is an unreliable predictor of teacher performance. In the extreme, if the evaluation score was uncorrelated with teacher impacts on test scores ($\alpha = 0$), then no teacher would qualify. At the other extreme, when the evaluation score is perfectly correlated with teacher performance (the R-squared in Equation 3 is equal to one), the posterior mean will be a teacher's true impact on test scores and we will correctly identify all 25 percent of teachers in the top quartile for ATC.

More generally, for any given level of tolerance and exceptionality, there is a direct relationship between the reliability of the evaluation score, as summarized by the R-squared in Equation 3 (or, equivalently, the correlation in Equation 4), and the proportion of teachers who will be eligible for ATC. As the correlation between the evaluation scores and the teacher effects on test scores increases from zero to one, the proportion of teachers eligible for ATC increases from 0 to 25 percent. Figure 3 in the main body of the paper plots this relationship, and the lookup tables provide the calculation of what percent of teachers should be eligible for any level of tolerance, exceptionality, and reliability of the teacher evaluation.

One nice feature of this approach is that it is not an "all or nothing" approach that requires every district's evaluation system to exceed a certain level of reliability. Instead, all districts can participate in ATC. Those with better evaluation systems will have more teachers eligible for ATC than those with worse evaluation systems.

While all of the discussion to this point has assumed that value-added estimates are available for all teachers, the system can be accommodated to include teachers evaluated by other methods (e.g., classroom observations, student evaluations) in untested grades and subjects. For these teachers, we would suggest using results from tested grades and subjects as a guide. In particular, for teachers in tested grades and subjects, the district could construct an evaluation score using the more limited information used in untested grades and subjects (e.g., only using classroom observations & student evaluations, but excluding value-added). Then, the correlation between this limited evaluation score and teacher impacts on test scores could be estimated as we have discussed, and this correlation would determine the proportion of teachers eligible for non-tested grades and subjects. Because evaluations would most likely be less reliable without value-added information, fewer teachers would be eligible in non-tested grades and subjects if decisions were based entirely on the reliability of prediction. However, we describe in the narrative of our report why districts might find it desirable to create equal opportunities for recognition among different categories of teachers even though this would create differences in the rate of classification errors in the different categories.

About the Brown Center on Education Policy

Established in 1992, the Brown Center on Education Policy conducts research and provides policy recommendations on topics in American education. The Brown Center is part of The Brookings Institution, 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.

Brown Center on Education Policy

The Brookings Institution
1775 Massachusetts Ave. NW
Washington DC, 20036
202.797.6090
202.797.6144 (f)
http://www.brookings.edu/brown.aspx

Support for this publication was generously provided by the *Bill & Melinda Gates Foundation*.

Governance Studies
The Brookings Institution
1775 Massachusetts Ave., NW
Washington, DC 20036
Tel: 202.797.6090
Fax: 202.797.6144
www.brookings.edu/governance.aspx

Editor
Christine Jacobs

Production & Layout
John S Seo

Email your comments to gscomments@brookings.edu

This paper is distributed in the expectation that it may elicit useful comments and is subject to subsequent revision. The views expressed in this piece are those of the authors and should not be attributed to the staff, officers or trustees of the Brookings Institution.

Endnotes

[1] Weisberg, W., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York: The New Teacher Project.

[2] Gabriel, T. (2010, September 2). "A celebratory road trip by education secretary," *New York Times*, p. A24.

[3] Hillsborough County Public Schools, New evaluation form for teachers, (June 8, 2010), available at http://communication.sdhc.k12.fl.us/empoweringteachers/?p=568.

[4] District of Columbia Public Schools (2010). *IMPACT: The District of Columbia Public Schools effectiveness assessment system for school-based personnel 2010-2011. Group 1 general education teachers with individual value-added student achievement data.* Washington, DC: DCPS.

[5] Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G.J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

[6] Glazerman, S., Goldhaber, D., Loeb, S., Staiger, D., & Whitehurst, G.J. (2010). *America's Teacher Corps*. Washington, DC: The Brookings Institution.

[7] Clotfelter, C., Glennie, E., Ladd, H., & Vigdor, J. (2006). *Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina*. NBER Working Paper 12285. Cambridge, MA: National Bureau of Economic Research.

Harris, D.N. (2009). "Teacher value-added: Don't end the search before it starts," *Journal of Policy Analysis and Management*, 28(4), pp. 693-699.

Hill, H.C. (2009). "Evaluating value-added models: A validity argument approach," *Journal of Policy Analysis and Management*, 28(4), pp. 700-709.

Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers*. Briefing Paper 278. Washington, DC: Economic Policy Institute.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G.J. (2010). *Evaluating teachers: The important role of value-added*. Washington, DC: The Brookings Institution.

MET Project (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

Rothstein, J. (2011). Review of "Learning about teaching: Initial findings from the measures of effective teaching project." Boulder, CO: National Education Policy Center.

Goldhaber, D. & Hansen, M. (2010a). "Is it just a bad class? Assessing the stability of measured teacher performance." CEDR Working Paper #2010-3. Seattle, WA: University of Washington.

Kane, T.J. & Staiger, D.O. (2002). "The promise and pitfalls of using imprecise school accountability measures," *The Journal of Economic Perspectives,* 16, 91-114.

Schochet, P.Z., & Chiang, H.S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education.

[8] Kane, T.J. & Staiger, D.O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* NBER Working Paper W14607. Cambridge, MA: National Bureau of Economic Research.

[9] Goldhaber, D. & Hansen, M. (2010). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: CALDER, The Urban Institute.

[10] Harris, D.N. & Sass, T.R. (2009). *What makes for a good teacher and who can tell?* Washington, DC: CALDER, The Urban Institute.

[11] MET Project (2010). *Learning about teaching: Initial findings from the measures of effective teaching project*. Seattle, WA: Bill & Melinda Gates Foundation.

[12] Schochet, P.Z. & Chiang, H.S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

[13] Kane, T.J. 2008. *Estimating teacher impacts on student achievement: An experimental evaluation*. Working Paper 14607. Cambridge, MA: National Bureau of Economic Research.

[14] U.S. Department of Education (2010, March 13). "A blueprint for reform," available at http://www2.ed.gov/policy/elsec/leg/blueprint/publicationtoc.html