Center for American Progress

ISTOCKPHOTO/NEILSTOCKIMAGES

When the Stakes Are High, Can We Rely on Value-Added?

Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions

Dan Goldhaber December 2010

Center for American Progress

When the Stakes Are High, Can We Rely on Value-Added?

Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions

Dan Goldhaber December 2010

Contents

Introduction and summary

The formula is simple: Highly effective teachers equal student academic success. Yet, the physics of American education is anything but. Thus, the question facing education reformers is how can teacher effectiveness be accurately measured in order to improve the teacher workforce?

There is a growing body of quantitative research showing teaching ability to be the most important school-based factor influencing student performance. The evidence that effective teachers significantly influence student achievement is clear. Unfortunately, improving the effectiveness of the teacher workforce is not a straightforward proposition; while research shows teacher effectiveness to be a highly variable commodity, it also shows that it is not well explained by factors such as experience, degrees, and credentials that are typically used to determine teacher employment eligibility and compensation.

When faced with high-stakes personnel decisions such as laying off teachers, granting tenure, or even paying out bonuses, many school districts, several states, and even the federal government are increasingly pushing for the use of measures of teacher effectiveness. From the Department of Education's Race to the Top initiative that urges states and districts to use teacher performance to inform personnel decisions, to the District of Columbia's IMPACT system that led both to significant bonuses for high-performing teachers and the dismissal of low-performing teachers, educational policy makers and administrators increasingly need transparent and accurate methods to quantify teacher performance.

The importance placed on identifying good teachers and bad teachers stands in stark contrast to the teacher evaluation system. Recent research suggests that teacher evaluation is a broken system. Drive-by classroom visits and binary ratings systems are insensitive to teaching assignments and typically assign unsatisfactory ratings to less than 1 percent of teachers. This "Lake Wobegon effect," where the great majority of a group is characterized as above average, fails to acknowledge

and represent the variation in teacher quality we know exists in the teaching workforce.[1] It is nearly impossible to use many existing evaluation methods for high-stakes personnel decisions such as: When all teachers are above average, how do you decide which teachers to lay off? Which teachers should receive tenure? Which teachers have earned bonuses in a performance-based system?

Given the demand for objective, quantitative measures of teacher performance and the shortcomings of many existing evaluation systems, it is not surprising that a number of districts and states have begun to utilize so called value-added models, or VAMs, to evaluate teachers. Based on the notion that gains in student test scores can be attributed to their teachers, VAMs are designed to measure the impact of individual teachers on student achievement, isolating their contribution to student learning from other factors (such as family background or class size in the early grades) that also influence student achievement.

The use of VAMs is highly controversial and tends to center, at least rhetorically around the notion that VAM measures of teachers will lead to perverse incentives or the misclassification of teachers. I would argue, however, that at least some of the policy debate on this issue masks the more fundamental issue of whether *any* system ought to differentiate teachers and act upon differences.

Today most teacher-evaluation systems rely on observational protocols (by principals or other trained professionals) and generally provide little real information about teacher effectiveness. Part of the reason is that teacher ratings are often on a binary scale where teachers are judged to be either "satisfactory" or "unsatisfactory." Even when the scale used allows for more nuanced judgments, most teachers receive a top-tier rating that fails to differentiate among teachers to any significant degree.

There are various ways teacher performance measures might be utilized were they to provide more information about the variation in teacher effectiveness. These range from low-stakes uses such as determining professional development, mentoring, or other means of remediating teachers deemed to be underperforming, to high-stakes uses such as compensation, promotion, or lay-off decisions.

When it comes to VAM estimates of performance, we actually know quite a bit. Researchers find that the year-to-year correlations of teacher value-added job performance estimates are in the range of 0.3 to 0.5. These correlations are gener-

ally characterized as modest, but are also comparable to those found in fields like insurance sales or professional baseball where performance is certainly used for high-stakes personnel decisions. Part of the reason that the correlations are only modest is that VAM estimates of effectiveness include measurement error, both because standardized tests are imprecise measures of what students know and because there are random elements such as classroom interaction that influence the performance of a group of students in a classroom.

The fact that measurement error exists may suggest that VAM effect estimates are too unstable to be used for high-stakes purposes because they will lead to teachers being misclassified into the wrong effectiveness categories. This is certainly a valid point to consider, but it is also essential to ground debates over changes to teacher evaluation in what is best for students. Classification error will occur with any evaluation system, but an exclusive focus on the potential downside for teachers ignores the fact that misclassification that allows ineffectiveness in the teacher workforce to go unaddressed is harmful to students. Ultimately, one has to make a judgment call about the risks of misclassification, but it is important to stress here that VAMs should be compared to the human capital systems currently in place and not to a nirvana that does not exist.

The argument for using VAMs is not merely based on the notion that its estimates provide important information about teacher effectiveness, as there is little doubt that they do. Rather it is an argument rooted in the idea that using VAMs is fundamentally important given the evidence that school systems, facing cultural or political constraints, have generally been institutionally incapable of differentiating among teachers. VAMs can be an honest broker when it comes to teacher-performance evaluation, ensuring any performance evaluation system recognizes that teachers are not widgets when it comes to helping students learn. Given this, it should come as no surprise that I believe we ought to experiment with the use of VAM teacher-effectiveness estimates to inform teacher policy.

Concerns about using VAMs are legitimate, but they overlook the fact that any type of teacher-performance evaluation with high-stakes consequences for teachers would be controversial. This controversy, however, rarely arises today because the performance evaluations that are currently being used typically are not high-stakes for teachers, either because they are not designed to be or because the evaluation itself is so inexact that the issue is rarely relevant *for teachers*. But the issue is very relevant *for students*. The misclassifications under the evaluations

governing the teacher workforce today come almost entirely in the form of false positives. I would hazard to say that few would disagree that there is at least some (possibly small) share of the teacher workforce in classrooms who should not be in the classroom despite the fact that they have the credentials and evaluations required to practice.

Unfortunately, much of the policy debate about VAM performance estimates is framed around the potential consequences for teachers rather than focusing on the consequences for students. It is entirely possible that the interests of teachers are not entirely congruent with the interests of students when it comes to teacher evaluation and classification. Certainly imperfect evaluation systems (the only types that exist), for example, that are connected to high-stakes policies, will lead to some incorrect teacher dismissals or rewards. The question however, should not be whether this is good or bad for teachers, but whether the number of incorrect classifications is acceptable given the impact on student learning.

My judgment is that current teacher policies lean too far in the direction of protecting teachers from the downsides of misclassification at the expense of the overall quality of the teacher workforce. It is for this reason that I advocate experimenting with teacher-evaluation system reforms (VAM-based and otherwise) that allow policy to better reflect the variation in performance that we know exists in the teacher workforce.

Given the high-stakes issues of student classroom achievement and teacher outcomes even up to dismissal, it is imperative that teacher evaluation methods provide spot-on performance assessments. The key then is having a system like VAM that truly differentiates among teachers while avoiding the pitfalls of misclassification. Still, regardless of the method used to evaluate teacher performance at the very least it must:

- Be rigorous and substantive while allowing for nuance.
- Provide meaningful teacher feedback.
- Be directly linked to consequences and outcomes.
- Be seen as trustworthy.
- Ultimately result in improved learning and achievement for students.

VAMs can be the honest broker when it comes to teacher-performance evaluation, ensuring that any performance evaluation system recognizes that teachers aren't widgets when it comes to helping students learn. Yet, having said that, VAM is often treated as if it is the magical elixir for all that ails the teacher workforce. There are good reasons to believe this is not the case. Thus, I also recommend that school systems implement a performance evaluation infrastructure that builds confidence in performance measures and provides teachers with timely feedback.

Teachers as widgets: The failure to recognize effectiveness

There are good reasons to critically examine teacher-evaluation systems. The weakness of today's evaluation systems effectively means that the differences in teacher effectiveness that we know exist appear not to influence teacher workforce decisions.[2] We know from value-added measures that teachers vary considerably from one another in effectiveness. The differences among teachers however, are not well explained by the kinds of teacher characteristics and credentials such as licensure, experience, and degree level that are used to determine employment eligibility and compensation.[3]

When determining compensation, most school districts rely on the single-salary schedule, meaning that teacher pay is determined almost exclusively on degree and experience level. And while there is clear evidence that teachers become more effective in their first few years in the classroom, the research also shows that this relationship levels out beyond the first three to six years of experience.[4] While experience does predict teacher effectiveness, there is still a significant share of, for instance, novice teachers who are more effective than third-year teachers and vice versa.[5] Further, teacher degree level appears to be completely unrelated to effectiveness outside of a few narrow exceptions.[6]

The fact that measured teacher performance varies so significantly, and that input measures do not accurately define teacher effectiveness, puts a premium on the evaluation of teachers. Unfortunately, much of the existing school system infrastructure for in-service evaluation appears to be weak. In their report "Rush to Judgment: Teacher Evaluation in Public Education", Tom Toch, who is currently executive director of the Association of Independent Schools of Greater Washington, and Robert Rothman, a senior fellow with the Alliance for Excellent Education, provide a comprehensive review of teacher-evaluation systems. They suggest that teacher evaluation is, on the whole, a broken system. Evaluations often constitute educational "drive-bys," quick classroom visits by principals or other school administrators that produce no substantive feedback for teachers on classroom practices. Teacher rating systems themselves, often reduced to a

simple "satisfactory" or "unsatisfactory" summary judgment, usually are inconsiderate of differences in the contexts (subjects, types of students, grade level, etc.) in which teachers work.

Given the state of affairs with teacher-evaluation systems, it should come as no great surprise that teacher evaluations appear to be far less than rigorous. It is clear that there is little variation in teacher evaluations, even when performance-rating systems allow for it. In 2007, for instance, an assessment of teacher evaluations in Chicago Public Schools by the New Teacher Project found that 93 percent of the system's teachers received ratings of "excellent" (61 percent) or "superior" (32 percent), and less than .5 percent of teachers received an "unsatisfactory" rating.

Why are teachers treated like widgets?

The generally dismal teacher-evaluation system can be summarized by two words: cost and consequences. First, rigorous evaluation—at least one that would be seen as credible—is likely to be costly. The typical drive-by evaluation system requires little beyond one visit to a teacher's classroom. Yet this brief encounter complies with most state requirements for teacher evaluation.[7] But as Robin Chait and Raegen Miller, both with the Center for American Progress, describe in "Treating Different Teachers Differently," rigorous evaluation likely entails frequent classroom visits (possibly by multiple evaluators) and assessing different types of evidence of teaching practices and student outcomes. Such an evaluation system requires an infrastructure that is probably far more costly than what is currently in place in many states and school districts.

Much the same could have been said for VAM several years back. Today, however, the extensive testing and accountability systems introduced under No Child Left Behind guarantees there are state assessments in third- through eighth-grade in math and reading. This does not necessarily mean the assessments are well-designed to facilitate VAM or that using VAMs in smart ways does not require ancillary investments. It certainly does, however, imply that in most states and localities the costs of utilizing a value-added system are far lower than they would have been a decade ago. Surely, most school systems could better take advantage of the student achievement information that could be derived from the vast amounts of data already being collected.

Second, typically there are no direct consequences associated with the evaluation that teachers receive. In the private sector much of the managerial control over employee quality is influenced by compensation. Such control is absent in a system that relies on the single-salary schedule. This means teacher evaluations default to the much higher stakes of whether or not teachers are allowed to remain in the profession. In theory, teacher tenure decisions or dismissals (in the case of layoffs) might be based on performance evaluations, but in practice this has been rare (though as I discuss later, the situation may be changing). Beyond compensation, principals often lack any direct managerial control over hiring, firing, or teacher placement. Given that there are often few consequences associated with a performance evaluation, evaluators are responding to the incentives they face when making judgments about teachers. Why would principals take evaluation seriously if it has no bearing on teacher performance? Moreover, in teacher culture, anything less than a satisfactory evaluation is contrary to the norm and thus likely to raise the ire of those receiving an unsatisfactory rating. Indeed there is some empirical evidence that school administrators are reluctant to give teachers a negative evaluation for fear of hurting the morale of the teaching faculty.[8] Again, why create waves if those waves are likely to be meaningless?

The bottom line is teacher input and evaluation policies fail to recognize the differences in performance we know exist among teachers. This is perhaps best exemplified by a report by Daniel Weisburg, vice president of policy and general counsel with the New Teacher Project, and colleagues in their New Teacher Project report, "The Widget Effect," which true to its title argues that the education system treats teachers like widgets, despite the substantial evidence that teachers are anything but.

Teacher performance measures and the move toward value-added

Teacher-evaluation systems differ from one another, but most rely on evaluation of observational protocols (by principals or other trained professionals); a far smaller proportion rely on examinations of portfolios that include teacher lesson plans and demonstrated commitment to practice and profession (or a combination of the two).[9] A 1996 study of evaluation in 68 of the 100 largest U.S. school districts found that about 95 percent use direct teacher observations as part of the evaluation system. Less than half (46 percent) use self-evaluations, and even less portfolio (24 percent), peer (16 percent), or student (9 percent) based evaluations.

Regardless of the method of teacher evaluation, it appears that evaluations generally provide little information about teacher performance, as the vast majority of teachers are labeled as some form of satisfactory. This statement should not necessarily be interpreted as a critique of the method of evaluation. Rather it reflects the fact that most evaluation systems suggest that all teachers are at the top of whatever scale is used to judge their performance. This point is exemplified in a study by Pamela Tucker, an associate professor in the Department of Leadership, Foundations, and Policy at the University of Virginia, in her analysis titled "Lake Wobegon: Where All Teachers Are Competent (Or, Have We Come to Terms with the Problem of Incompetent Teachers?)." Tucker's study focuses on principals and teachers in Virginia and finds that far fewer teachers are classified as incompetent than are believed to be so by either their principals or other teachers. Specifically, principals reported teacher incompetence rates averaging 5 percent, but only about half of those were formally identified as being incompetent. And, interestingly, the rates of formal identification were actually slightly higher for tenured teachers (about 1.5 percent) than untenured teachers (about 1.1 percent), possibly because formal documentation is far more important for the removal of a tenured than an untenured teacher. Tucker's results were corroborated a decade later in "The Widget Effect."

Part of the problem is that binary (satisfactory or unsatisfactory) measures of teachers provide little room for nuance in any assessment. The authors of "The Widget Effect" report that more than 99 percent of teachers in districts with a

binary evaluation system are rated as satisfactory. But even when a multitiered system is utilized, there still appears to be little teacher differentiation. Again from "The Widget Effect": More than 94 percent of teachers in districts with broader rating options receive one of the top two ratings, and less than 1 percent of teachers receive an unsatisfactory rating.

It is conceivable that existing teacher-evaluation tools could be used to distinguish among teachers more productively, but unless there really are only about 1 percent of teachers who are not meeting society's expectations, it is also quite clear that the culture or political constraints in schools make honest assessments of performance difficult. I believe this is leading the march toward the use of value-added. A key advantage this methodology has compared to many existing methods of evaluation is that, while imperfect, it brings greater objectivity and thus has the potential to surmount cultural and political obstacles.

Observational-based performance systems and student achievement

Research by Brian Jacob, a professor of education policy with the Gerald R. Ford School of Public Policy at the University of Michigan and Lars Lefgren, associate professor of economics at Brigham Young University, finds evidence that confidential (so not necessarily reflecting documented performance) principal evaluations of teachers are significantly correlated with teacher VAM effect estimates, implying that these subjective evaluations can serve as predictors of teacher contributions to student learning. Interestingly, they also find evidence that principal evaluations of teachers predict student achievement in statistical models that include estimates of the prior VAM measures, implying that these subjective assessments provide performance information about teachers that goes beyond that provided by VAM measures.[10]

In 2004 the *Peabody Journal of Education* devoted a special issue to exploring the link between observational-based evaluations and student achievement.[11] Many of the papers in this special issue sought to validate the use of these evaluations by assessing the relationship with the estimated value-added of teachers. Three observational studies in this issue, covering three different sites—Ohio, California and Nevada—show that non-VAM measures hold promise. Specifically, the studies all find statistically significant positive relationships between evaluation scores using the Framework for Teaching—an evaluation method that combines results

of standards-based principal evaluations with nonclassroom observations, reflection forms, unit plans, logs of professional activities, and parent contacts, along with measures of value-added teacher performance.[12]

Two recent studies further explore the correlation between non-VAM evaluation systems and quantitative measures of teacher performance. First, there is a 2010 study that estimates the association between evaluation scores using Cincinnati's Teacher Evaluation System, based on the Framework for Teaching evaluation and student achievement in Cincinnati. The study found that having a teacher with a one point higher rating, equivalent to moving up one step in the rating categories of—unsatisfactory, basic, proficient, and distinguished—corresponds to a large improvement in student achievement; roughly 15 percent of a standard deviation improvement in math achievement and 20 percent of a standard deviation improvement in reading achievement.

Second, a 2010 study that evaluated middle school language arts teachers using a combination of the Classroom Assessment Scoring System (widely know as "CLASS"), and a new observation system they developed specifically for secondary English/language arts instruction, found that teachers in the highest quartile of value-added scored better than teachers in the second quartile on all 16 elements of instruction they measured.[13]

Portfolio and other evaluation systems

Portfolio-based evaluations are one alternative, or a complement, to the use of observational ratings for teacher evaluation. Probably the best known assessment that relies heavily on a portfolio is that designed by the National Board for Professional Teaching Standards (NBPTS). The NBPTS-credential has been widely adopted and teachers who are certified by the National Board often receive additional compensation.[14] Thus, the validity of the credential has attracted significant research attention.

Some empirical evidence shows a statistically significant positive relationship between teachers being NBPTS-certified and student achievement, but the findings are not universal.[15] And even though my previous research does show that this credential is predictive of student performance, I also illustrate that the variation in teacher performance due to being credentialed is swamped by other

differences within the groups of credentialed and noncredentialed teachers. In the case of my research, about 60 percent of teachers who become certified are more effective than their noncertified counterparts when judging them by students' achievement in mathematics tests and about 55 percent when judging them based on students' achievement on reading tests.[16]

Outside of the research on NBPTS, there is relatively little quantitative evidence tying teacher ratings based on portfolio evaluations to student achievement. Likewise, the research base informing what we know about other forms of teacher evaluation—peer, self, and student ratings of instruction—is also sparse.[17] This significant gap in the literature will likely be addressed in the near future as private and federal efforts are being made to identify ways to evaluate teachers. A number of philanthropic organizations are funding research on new teacher evaluation techniques. The Bill and Melinda Gates Foundation's Measures of Effective Teaching study, for example, is focused on assessing the relationship between various methods of evaluating teachers—including video-taped observations, student surveys, and tests of teacher pedagogical knowledge and student achievement.[18] Importantly, all the teachers voluntarily participating in the study agreed to be randomly assigned to their classrooms in one year, allowing the study to address one of the central critiques of performance-based evaluation systems, which is that the system in general may be unable to accurately distinguish teacher contributions to students from the effects of having students with particular backgrounds grouped together in a classroom. Teacher evaluation is also a central theme in the federal government's Race to the Top competition.

Finally, there is growing attention to the teacher policies in other countries. This is due to evidence that U.S. students score in the middle of the pack on international assessments and some international competitors to the United States are perceived to be very successful in recruiting, training, and evaluating teachers.[19]

Growing interest in using value-added measures of teachers

Over the last decade, and increasingly now, a number of districts and states have begun utilizing VAM methods to evaluate teachers based on the notion that gains in student test scores can be attributed to their teachers. VAMs have been used for teacher evaluation as far back as the 1990s in Tennessee and Dallas. More recently many proposals for the Race to the Top federal grant program, including a winning entry from Delaware that makes the bold statement that "satisfactory student

growth is the minimum requirement for any educator to be rated effective," propose using value-added measures to inform teacher evaluations. Moreover, some states and districts are considering using VAM estimates as the primary criterion for awarding teacher tenure,[20] the potential consequences of which I recently explored in a paper with Michael Hansen, a researcher with the Education Policy Center at the Urban Institute.[21]

The growing interest in using VAMs to evaluate individual teachers likely arises because it is now possible, at least in theory, to use VAMs in most states since students in third-grade through eighth-grade are tested annually, so one can obtain value-added estimates.[22] But perhaps more important in driving the current policy interest is the finding mentioned above, for myriad reasons, that few teachers ever receive anything but a satisfactory evaluation, regardless of the rating system utilized (and regardless of whether there are consequences attached to the rating). This point is well-documented and cannot be emphasized enough.

Policymakers likely see VAMs as a way to address the "Lake Wobegon effect"—specifically, the political and cultural constraints leading most evaluation systems to conclude that all teachers are excellent. And while VAMs have the potential to do this, using student achievement to judge teacher performance is limited. First, VAM measures currently could only cover roughly 25 to 35 percent of the teacher workforce; the majority of teachers are in classrooms or grades not covered by state assessments, such as music, art, or first and second grades.

Second, for many, the idea of using student achievement on tests to make judgments about teachers (or schools) is problematic given that tests are limited in terms of what they measure. Specifically, tests can only be used to assess a subset of the myriad objectives of schooling. It is unlikely, for example, that tests will cover socialization behavior that is taught in schools. Even when it comes to academic competence, tests are necessarily limited in terms of their ability to identify the extent to which students have learned important topics covered in the classroom.[23]

And third, focusing accountability narrowly on tests may result in perverse incentives, leading teachers to try to "game" the system by drilling in on test-taking skills that are not generally beneficial to students, or outright cheating to make the performance of their students appear to be better than it really is.

It is worth noting that one must weigh these limits against other factors. The state assessments often used for value-added are typically constructed with input from

numerous stakeholder groups that help to drive educational policies (such as school accountability), and thus presumably reflect educational goals. Student achievement on these tests has been shown to be related to both the later life outcomes of students and aggregate country growth rates, and there is very little evidence supporting the idea that perverse incentives associated with test-based accountability has been detrimental to student achievement.[24]

Beyond the question of whether tests are a good way of measuring the important lessons that students learn in schools is the issue of whether student assessments can be used to assess teacher performance. I do not delve too deeply into the complex statistical issues here, but it is worth noting that the models used to predict the factors that influence student learning rely on a number of strong assumptions about both the student learning process and the psychometric properties of the tests themselves.[25]

One of the most important issues that arise is the amount of true information about teacher effectiveness that can be gleaned from value-added and whether this information can be used to classify teachers. It would, for instance, be questionable to attach great meaning to changes in student test performance if a teacher only has a small handful of students. One reason is that standardized tests are imprecise measures of what students know. Time and resources limit the amount of content that they cover, and random occurrences, such as the amount of sleep a student gets prior to taking a test, will influence individual performance. This issue is exacerbated at the classroom level because of the potential that other random occurrences—the dynamics between students in a class in a particular year, or the oft-cited distracting dog barking outside the classroom on testing day—influence the performance of the entire class of students. In a small class the achievement of one or two students can greatly affect the estimate of teacher performance.

Further, teachers whose measures of effectiveness are based on only a few student test scores are far more likely to be influenced by these random factors—"noise" or "measurement error" in statistical parlance. Given that VAM estimates are, to some degree, going to be noisy estimates of true teacher performance, one must worry about using them for high-stakes purposes. It is possible that some teachers, who are for instance, classified by VAMs as highly effective or very ineffective, actually fall into the other category.

I want to emphasize that the problem of misclassification exists whether or not one uses a VAM approach to judge teacher performance. Any evaluation/perfor-

mance assessment system is subject to classification error. Observers of teacher behavior or those who judge portfolios, for instance, may make mistakes that lead to teachers being misclassified. The single-salary schedule that determines teacher pay clearly leads to classification errors (if the desire is to classify according to effectiveness). VAMs simply make the potential for classification errors more transparent and explicit because they are less subjective and can be easily evaluated given that they are statistically-based measures. Consequently, it is far easier to determine the likelihood that they lead to misclassification, even if they perform better than other methods of evaluation.

How could teacher performance measures be utilized?

Low-stakes use of VAM

There are various personnel decisions that might be informed by teacher performance measures. Low-stakes uses of these measures might be to make decisions about professional development, mentoring, or other means of remediating teachers deemed to be underperforming. But it is the atypical school or school system that actively uses evaluations to help determine *individual* teacher needs. When it comes to professional development for example, schools usually employ one of two strategies; either they bring the professional development to the school in the form of a workshop, or they leave it up to individual teachers to determine their professional development needs (often from a menu of approved offerings).[26] In the case of the former, there is no differentiation in the professional development since all teachers in the school are receiving the same training.

The latter model offers greater opportunity for differentiation, but it is left up to teachers to decide on training and, given that performance evaluations tend to be undifferentiated, they may not have information on what areas to improve when choosing. Moreover, one might argue that even if teachers are aware of their areas of weakness, they do not have terribly strong incentives to choose professional development that addresses those weaknesses. The loose connection between performance evaluation and professional development implies that evaluation is not typically being used in an effective way, even for low-stakes purposes.

Medium-stakes use of VAM

The default in thinking about how VAM estimates are used is usually one of the poles: high- or low-stakes, but there are actually at least two of what I would consider to be medium-stakes ways in which VAM estimates might be utilized. One is to gauge the effectiveness of teachers graduating from different training institutions and use these measures to inform accountability policies at the institution

level.[27] This obviously would be a high-stakes use of VAM to institutions, but have no immediate direct impact on individual teachers. A second would be to help determine which teachers might qualify for differentiated roles in schools. There is increased interest in creating career path options for teachers that allow them to stay in the classroom, but help other teachers with their craft, or reach students in other ways.[28] Of course these differentiated roles may also lead to greater compensation, which would put them in the realm of high-stakes.

High-stakes use of VAM

High-stakes uses of performance evaluations would include using them to help determine compensation, renewal of teacher contracts (for untenured teachers), teacher tenure, and/or dismissal of chronically ineffective teachers. Using performance evaluations as a factor in compensation—the lowest-stakes of these high-stakes options—is precluded in most school systems based on their use of the single-salary schedule. I argue that teacher compensation is the lowest-stakes use of performance evaluations because the other uses are all or nothing outcomes that involve determining whether or not teachers will have a teaching job.

Tenure decisions or dismissals are rarely based on performance evaluations. The issue of using performance as a basis for teacher dismissals has recently been in the public eye, most notably in the recent dismissal of 241 teachers by Michelle Rhee, Washington, D.C. Public Schools Chancellor. Of that number, 165 of these dismissals were the result of poor performance on DC's new teacher-evaluation system known as IMPACT, which uses value-added estimates along with a rigorous rubric-based observation system to evaluate all teachers in the district.[29] The Chicago Public School District also recently approved a policy that allows administrators to dismiss tenured teachers who were rated "ineffective" before following the "last hired, first fired" procedure spelled out in the district's collective bargaining agreement. These policies have garnered considerable public attention and have ignited a heated debate about the use of teacher performance over seniority in teacher layoffs.

Despite these controversies, a number of studies show that the teacher dismissal rate is typically in the neighborhood of 1 to 2 percent per year.[30] That dismissals are rare is not surprising, particularly dismissals of tenured teachers. As CAP's Robin Chait reports in "Removing Chronically Ineffective Teachers," teacher dismissals are quite costly; the dismissal of a tenured teacher can cost in the hun-

dreds of thousands of dollars.[31] This is consistent with data from the Schools and Staffing Survey showing tenure to be the "barrier to dismissal of poor or incompetent teachers" most likely to be identified by school principals.[32]

There is, however, an important caveat to the 1 percent figure cited above; the surveys on teacher dismissals document only formal dismissals, so it may be that many untenured teachers are counseled out of schools without formal documentation so that what is tantamount to a dismissal flies under the radar. The downside to this informal approach is that there is no track record of performance so teachers can easily obtain a job in another school or district, even if they may not be well suited to the profession. In fact, in a forthcoming paper with Betheny Gross and Daniel Player, researchers at the University of Washington's Center on Reinventing Public Education and Mathematica Policy Research respectively, "Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping Their Best?" we offer empirical evidence supporting this notion.[33] We investigate the mobility patterns of teachers who fall into different effectiveness categories (judged based on value-added models) and find that the least effective teachers are more likely than more effective teachers to leave teaching (at least in the North Carolina public school system), but they are also more likely to "churn" through the public school system moving from school to school.

These findings suggest that performance evaluation could be used more effectively to identify the lowest performing teachers for dismissal and/or create a track record of performance that would follow teachers so that evaluations from a prior teaching position would contain real information about performance that could inform decisions about teachers who move from school-to-school or district-to-district.

The use, or nonuse, of performance to inform tenure may be on the cusp of change. The federal Race to the Top grant program specifically encourages states to use teacher performance to inform "key decisions in such areas as evaluation and development, compensation and advancement, tenure and removal," and each of the winning applications in the first round pledged to use teacher performance to inform tenure decisions.[34] Tennessee's application has a set of policy changes that includes "denying tenure to teachers who are deemed ineffective as gauged partly by student growth,"[35] while Delaware will grant tenure to teachers only "if they demonstrate satisfactory student growth for two or more years, and have no more than one year of 'ineffective' teaching."[36]

Moreover, even states that have not received Race to the Top grants have been able to use the program to give them "the political cover they need to push through reforms unpopular with unions."[37] One of those states is Colorado, which recently overhauled its teacher tenure and evaluation rules so students' academic progress now counts for half of a teacher's evaluation rating, and teachers are now required to receive three consecutive years of positive evaluations to earn tenure.

Lastly, the issue of teacher layoffs has arisen as a consequence of the economic downturn. The evidence on the importance of teacher quality clearly implies that when layoffs are necessary it matters a great deal to students which teachers are laid off. Not surprisingly, layoffs are typically addressed in collective bargaining agreements, and in the overwhelming majority of these agreements, seniority is the primary criterion for determining which teachers are laid off. An analyses of the National Council on Teacher Quality TR^3 database shows that "last hired, first fired" seniority provisions exist in all of the 75 largest school districts in the nation, and seniority is the *sole* factor that determines the order of layoffs in over 70 percent of these districts.

The implications of using seniority rather than effectiveness as the determinant of layoffs is explored in a new policy brief published by the National Center for the Analysis of Longitudinal Data in Education Research. In the brief, the authors simulate seniority-based versus performance-based (as measured by value-added) layoffs in New York City and find that there would be significant differences in the average performance of teachers who are laid off. Specifically, they estimate that teachers laid off based on value-added judgments of teacher performance would be about a 25 percent of a standard deviation (in student achievement) *less* effective than those laid off under a seniority-based system.[38]

Throughout this report I have suggested that performance evaluations could be used to inform both low- and high-stakes teacher workforce decisions, but this presumes the performance evaluations are in fact predictive of teacher quality. I will next explore this question in more detail, describing new findings on the extent to which a particular type of teacher performance measure—value-added estimates of teacher effectiveness—actually predict *future* teacher performance, an important quality if they are used to determine employment eligibility.

Are VAM measures good predictors of future teacher performance?

There are good reasons to be concerned about whether teacher performance measures in one time period are related to those measures in other time periods. The "intertemporal stability" (how correlated the measure is in one period of time to the next) of the measure provides an indication of whether the measure is providing true teacher performance information. The notion here is that underlying teacher quality is unlikely to change radically from one period to another. Therefore, radical change in the measure of teacher performance suggests that it may be a poor proxy for true performance. Moreover, measures with really low intertemporal correlations are likely to be poor predictors of future job performance. This does not mean that a measure having very high intertemporal stability is necessarily going to mean it is a good predictor of job performance. As noted in the introduction, the degrees held by a teacher serve as very poor predictors of his or her effectiveness, even though that factor is quite stable; once a teacher gets a master's degree, they always have it.

As I noted earlier, outside of a few exceptions, there is relatively little quantitative evidence linking non-VAM measures of teacher performance to student achievement, hence it is no surprise that we know little about whether these types of evaluations predict future teacher performance. There exists far more information about the intertemporal properties of VAM effect estimates.

Intertemporal properties of VAM teacher performance measures

Researchers generally find that the year-to-year correlations of teacher value-added job performance estimates are in the range of 0.3 to 0.5, with correlations of teacher value-added in math generally found to be higher than value-added in reading/English language arts.[39] These correlations have been characterized as "modest," as would be expected given that a significant proportion of the intertemporal variation in these effects appears to be due to "statistical noise."[40]

Unfortunately, it is impossible to know precisely how much of the intertemporal instability of these VAM measures is due to measurement error versus true year-to-year changes in teacher performance. But, we do know from a statistical perspective that many estimated teacher effects will not be statistically distinguishable from the effect of the mean teacher on student achievement.[41] And the more fine-grained we try to be, the less likely teachers will be truly distinguishable in a statistical sense.[42] Finally, we know that teacher rankings based on these estimates will change from year to year and some of this is due to chance.[43]

A simple way to improve the stability of VAM performance estimates is to use more information about student achievement to inform teacher effect estimates. Calculating multiyear estimates of teacher performance, for example, are likely to be less noisy because a lucky draw in one year is offset by an unlucky draw the next.[44] Recent simulation work by Peter Schochet and Hanley S. Chiang, researchers with Mathematica Policy Research, examines this issue by using VAM estimates to identify teachers whose performance differs in a statistically significant way from an average teacher. They conclude that using three years of data for estimating teacher effectiveness, the probability of identifying an average teacher as being "exceptional" (Type I error)—defined by them as teachers who are roughly one standard deviation of teacher performance above (or below) the mean—is about 25 percent.[45] Conversely, the probability that a truly exceptional teacher is not identified for special treatment (Type II error) is also roughly 25 percent.

The work by Schochet and Chiang is useful in framing the potential for misclassification that exists when using VAMs to inform policy, but it is worth noting that they picked cutoffs such that the Type I and Type II error rates were equivalent. In practice, districts might pick very different cutoff scores to minimize the likelihood of sanctioning a teacher who ought not to be sanctioned (a Type I error) or failing to reward a teacher who deserves a reward (a Type II error). Even so, the numbers presented by Schochet and Chiang may suggest to some that VAM effect estimates are too noisy, and therefore too unstable, to use for high-stakes purposes.

Ultimately this is a judgment call, but there are two points worth considering. The first is that the estimates are strikingly similar to estimates we see in sectors of the economy that consider them for job retention and pay determination.[46] And second, as I mention above, the teaching profession is currently utilizing credentials and evaluations for high-stakes purposes (or in some sense not utilizing them for in-service teachers since there is rarely any variation in evaluations) that are themselves noisy, but less transparently so. Advocating for more stability of VAM

(or any other) measures may be setting a higher bar for change (possibly a much higher bar) than the status quo.

Moreover, as Mike Hansen, a researcher with the Education Policy Center at the Urban Institute and I show in a recent report, "Is it Just a Bad Class? Assessing the Stability of Measured Teacher Performance," there are reasons to believe that focusing on the transition of teachers between different effectiveness categories (e.g. quintiles) may overstate the degree of instability of performance estimates because the baseline assumption is that the transitions are due to measurement error. Specifically, we present evidence (using a simulation approach) that some of the year-to-year transitions between effectiveness categories result from *true* changes in teacher performance.[47] If so, VAMs may be somewhat more accurate than they have been portrayed to be in the literature.

Using VAM estimates for tenure determination

The ultimate test of any teacher job performance estimate used for employment determination is whether the estimate predicts teacher performance in the future. If it does not, or does so poorly, one probably would not want to use the measure for this type of high-stakes purpose.[48]

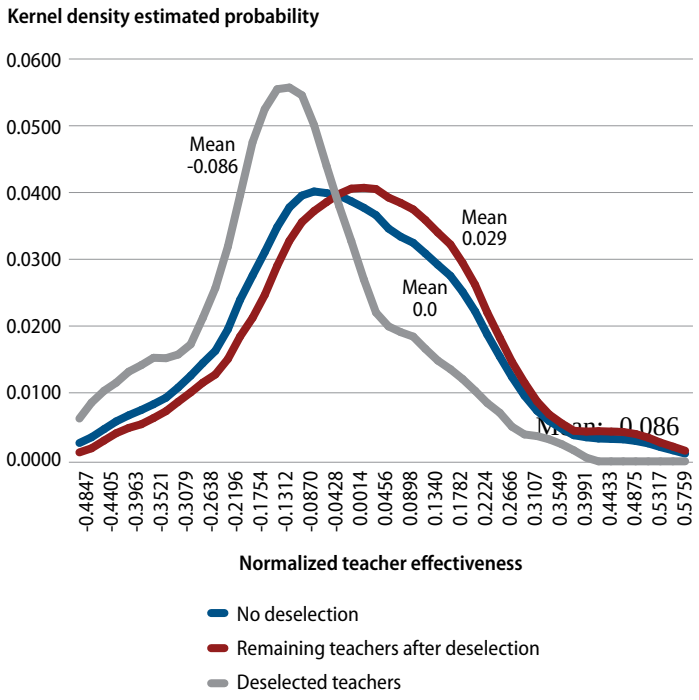
I investigate the question again with Mike Hansen in "Using Performance on the Job to Inform Teacher Tenure Decisions." We estimate teacher performance for a group of elementary school teachers in North Carolina using their first two years of classroom performance to inform VAM math estimates. We then compare how these early career performance estimates perform relative to a host of observable teacher characteristics and credentials (licensure test scores, experience, and degree levels, etc.) by using each to predict student achievement out of a sample and then comparing each of these predictions to *actual* student achievement.[49] Given the well-established literature showing that observable teacher credentials are at best weak predictors of student achievement, it is not terribly surprising that early career VAM estimates are considered far superior to teacher characteristics and credentials at predicting student achievement in the future (in this case after a three-year gap between the estimated VAMs and the student achievement predictions).

What would using VAM effect estimates to inform tenure determination mean for workforce quality? We assess this by investigating a hypothetical case where school systems opted not to tenure teachers who fall into the lowest quartile of the pre-tenure math performance distribution (based on their first two years in

the classroom). Twenty-five percent of the workforce may seem like a very high proportion of the workforce to select out (or "de-select") from the teacher workforce, but it is not as large as it may sound at first blush. The reason is that early career teacher attrition is relatively high and many of those teachers who leave tend to be ineffective.[50] In our dataset, the natural attrition that takes place implies that districts would need to de-select an additional 14 percent of the teachers eligible for tenure in order for the weakest 25 percent to be weeded out through a combination of purposeful de-selection and attrition. Figure 1 shows the estimated post-tenure (fifth year) VAM effects (in student achievement terms) for those teachers who would be de-selected (the lowest 25 percent), the remaining teachers who are selected for tenure (the upper 75 percent), and the pooled distribution of all teachers (imposing no selection rule).

VAM calculations

The effect of teacher de-selection on the math effectiveness of the teacher wokforce



Source: D. Goldhaber and M. Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review: Papers and Proceedings,* 100 (2)(2010): 250-255.

The difference in effectiveness between the de-selected and selected distributions is educationally significant: the de-selected teachers are estimated to have math impacts that are more than 11 percent of a standard deviation of student achievement lower than selected teachers. This is certainly a large effect for students who would be educated by selected versus deselected teachers. To put this in context, the study by Schochet and Chiang suggests that 20 percent of a standard deviation of student achievement is roughly equivalent to three months of typical learning growth, so our findings suggest the average difference between selected and de-selected teachers is about 1.5 of months of typical learning growth. And, the estimated impact on the quality of the teacher workforce overall is also large: the difference between the selected distribution and a distribution with no de-selection is almost 3 percent of a standard deviation of student achievement. Taking the thought-experiment a step further and assuming that de-selected teachers are replaced with teachers who have effectiveness estimates that are equal to the average effectiveness of teachers in their first and second years, the average effect for post-tenure teachers under the VAM-based screen is still about 2.5 percent of a standard deviation larger than if no screen had been applied.

Conclusion

The notion that we ought to use student performance to help guide low-stakes decisions, like the types of professional development that are offered, seems pretty benign. Who would argue that we ought to disregard all the information we are collecting about student achievement? Similarly, the idea that teacher performance ought to be used to inform high-stakes personnel decisions is likely to be uncontroversial; again, who would argue that we should not care about how effective a teacher is when deciding, for instance, whether he or she ought to be allowed to be in a classroom? It is only when these two ideas are grouped together using student performance to guide high-stakes personnel policies that sparks fly.

I would argue that, at a more fundamental level, the controversy over using VAM estimates is about whether school systems actually act on the differences that exist between teachers. If true, VAM is a convenient foil for the larger debate, and if policymakers were implementing non-VAM teacher evaluation reforms that led to the same degree of differentiation of teachers, as does VAM, the policy conversation would look very similar. Thus, the argument for VAMs goes beyond whether it provides useful information (the evidence is that it does) to whether it is necessary to use a system that yields objective assessments of teachers to help sidestep the political and/or cultural impediments to rigorous evaluation and teacher differentiation. The key then is having a system that really does differentiate among teachers. And regardless of the methods used to evaluate teachers, this likely requires that school systems have performance ratings that allow for greater nuance than the binary satisfactory-unsatisfactory option.

The policy conversation about using VAMs for high-stakes purposes is mostly about the possibility that VAM measures of teachers will lead to perverse incentives or the misclassification of teachers. Most of the academic evidence about VAMs explores these issues. But while it makes sense to try to limit dysfunctional teacher behavior and classification errors, particularly those that may cost teachers their jobs, it is also essential to ground debates over changes to teacher evaluation in what is best for students. Any evaluation system can create incentives for teach-

ers and is subject to classification errors. I would argue that today much of the debate about VAMs tends to focus on the downsides—the fact that some teachers are likely to cheat—rather than the limitations of the status quo.

There is no doubt that under VAMs or any other system that classifies teachers as ineffective, there exists the potential that some good teachers will be deemed to be ineffective. Focusing on this potential harm to teachers is appropriate, but one should not ignore the fact that misclassification in the other direction is harmful to students. No one wants misclassification, but the socially optimal number of misclassifications is certainly not zero.

The research suggests that given the limitations of any one way of evaluating teachers, there are benefits to using multiple methods or indicators. As CAP's Raegen Miller notes in a recent report, "Adding Value to Discussion About Value-Added", "…the more serious a decision is, the more important it is that multiple indicators of effectiveness inform the decision [and] … the more serious a decision is, the more important it is that indicators of effectiveness be trust-worthy." In fact, multiple indicators are likely to both limit classification errors and increase confidence in the ultimate judgments about effectiveness. In practice, for instance, one might use VAM methods to identify a pool of teachers for special consideration and then alternative methods might be used to better understand the strengths and weaknesses of those teachers.

VAMs can be the honest broker when it comes to teacher-performance evaluation, ensuring that any performance evaluation system recognizes that teachers are not widgets when it comes to helping students learn. This is fundamentally important given the evidence that school systems, under cultural or political constraints, have generally been institutionally incapable of differentiating among teachers.

Having systems that do differentiate among teachers and act on those differences is a good first step towards improving the effectiveness of the teacher workforce, but it is not likely to be a panacea or even generate much short-term benefit. This conclusion was recently illustrated by a 2010 study of a pay-for-performance incentive experiment in Nashville where teachers were being judged and rewarded for student achievement, but the evidence suggests little systemic benefit from the incentive system. There are myriad potential reasons why the incentive did not make a difference in Nashville, but what I want to mention here is that a single reform is unlikely to be a magic bullet. To really harness the power of good evaluation and incentives, school systems need a performance evaluation

infrastructure that instills confidence in the evaluation system, and, importantly, provides teachers with timely feedback about their performance so that they have the opportunity to act on any assessment. But, more than that, hoping for a large immediate student achievement effect is probably shortsighted. One of the powerful ways that evaluation and incentives are likely to play out is through changing the mix of people who opt to teach or, over time, the performance of those who act on their evaluations.[51] These types of outcomes clearly take time to occur.

There are, of course, myriad potential effects associated with any fundamental change of the consequences of teacher-performance evaluations, VAM-based or otherwise. Teaching is a low-risk occupation when it comes to compensation and job security, so introducing greater risk in the form of consequential performance evaluations might reduce the number of people aspiring to become teachers in the absence of other complementary changes. For this reason it might be advisable for policymakers considering VAM-based reforms to think about building resources into any plan that could be used to pay teachers a "risk premium."

Obviously, some of the prior discussion is based on theoretical speculation; it is, of course, not possible to really know the general equilibrium effect of implementing a VAM-based evaluation system in the absence of actual policy variation. Thus, it is ultimately a judgment call as to whether the potential benefits of VAM-based reforms offset the potential risks. And it is important that the risks be juxtaposed against the Lake Wobegon status quo where nearly all teachers are deemed to be equally excellent. The true risk of the status quo is not just that it fails to recognize and reward effective teachers, but more importantly it results in too many students being educated by ineffective teachers.

Clearly, we ought to be experimenting with reforms of the teacher-evaluation system (VAM-based and otherwise) that focus on teacher effectiveness and better reflect the variation in teacher performance to inform policy that will improve the teacher workforce.

References

Aaronson, D., L. Barrow, and W. Sander. 2007. "Teachers and Student Achievement in the Chicago Public High Schools." *Journal of Labor Economics* 25 (1) (2007): 95–135.

Ballou, D. 2005. "Value-Added Assessment: Controlling for Context with Misspecified Models." Washington: Urban Institute Longitudinal Data Conference.

Ballou, D., W.L. Sanders, and P.S. Wright. "Controlling for Student Background in Value-Added Assessment of Teachers." *Journal of Education and Behavioral Statistics* 29 (1) (2004): 37–65.

Bill and Melinda Gates Foundation. 2010. "Working With Teachers to Develop Fair and Reliable Measures of Effective Teaching." Seattle.

Borko, H. "Professional Development and Teacher Learning: Mapping the Terrain." *Educational Researcher* 33 (8) (2004): 3–15.

Boyd, D. and others. 2010. "Teacher Layoffs: An Empirical Illustration of Seniority v. Measures of Effectiveness." CALDER working paper.

Cantrell, S. and others. 2008. "National Board Certification and Teacher Effectiveness: Evidence From a Random Assignment Experiment." Cambridge: Harvard Graduate School of Education.

Cavalluzzo, L. C. "Is National Board Certification an Effective signal of Teacher Quality?" Alexandria: CNA Corporation.

Centra, J. A. "Colleagues as Raters of Classroom Instruction." *The Journal of Higher Education* 46 (3) (1975): 327–337.

Chait, R. 2010. "Removing Chronically Ineffective Teachers." Washington: Center for American Progress.

Chait, R. and R. Miller. 2010. "Treating Different Teachers Differently: How State Policy Should Act on Differences in Teacher Performance to Improve Teacher Performance and Equity." Washington: Center for American Progress.

Clayson, D. E. "Student Evaluations of Teaching: Are They Related to What Students Learn?" *Journal of Marketing Education* 31 (1) (2009):16–30.

Clotfelter, C. T., H.F. Ladd, and J.L. Vigdor. "Teacher-Student Matching and the Assessment of Teacher Effectiveness." *Journal of Human Resources* 41 (4) (2006): 778–820.

Cohen, P. A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research* 51 (3) (1981): 281–309.

Cohen, P. A. and W.J. Mckeachie. "The Role of Colleagues in the Evaluation of College Teaching." *Improving College and University Teaching* 28 (4) (1980): 147–154.

Dickinson, D. J. "The Relationship Between Ratings of Teacher Performance and Student Learning." *Contemporary Educational Psychology* 15 (1) (1990): 142–151.

Feldman, K. A. "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers." *Research in Higher Education* 30 (2) (1989): 137–194.

Ferguson, R. 2008. "Raising Achievement and Closing Gaps in Whole School Systems: Recent Advances in Research and Practice." Cambridge: Harvard University, Annual Conference of the Achievement Gap Initiative.

Follman, J. "Secondary School Students' Ratings of Teacher Effectiveness." *The High School Journal* 75 (3) (1992): 168–178.

Gallagher, H. A. "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education* 79 (4) (2004): 79–107.

Goldhaber, D. "The Mystery of Good Teaching." *Education Next* 2 (1) (2002): 50–55.

_____ "National Board Teachers Are More Effective, But Are They In the Classrooms Where They're Needed the Most?" *Education Finance and Policy* 1 (3) (2006): 372–382.

_____ 2006. "Teacher Pay Reforms: The Political Implications of Recent Research." Washington: Center for American Progress.

_____ "Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?" *Journal of Human Resources* 42 (4) (2007): 765–794.

_____ 2009. "Lessons from Abroad: Exploring Cross-Country Differences in Teacher Development Systems and What They Mean for U.S. Policy." In D. Goldhaber and J. Hannaway, eds., *Creating a New Teaching Profession*. Washington: Urban Institute Press.

Goldhaber, D. and E. Anthony. "Can Teacher Quality be Effectively Assessed?" National Board Certification as a Signal of Effective Teaching." *Review of Economics and Statistics* 89 (1) (2007): 134–150.

Goldhaber, D., D. Brewer, and D. Anderson. "A Three-Way Error Components Analysis of Educational Productivity." *Education Economics* 7 (3) (1999)**:** 199–208.

Goldhaber, D., B. Gross, and D. Player. "Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping their Best?" *Journal of Public Policy and Management* (Forthcoming).

Goldhaber D. and M. Hansen. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review: Papers and Proceedings* 100 (2) (2010): 250–255.

_____ 2010. "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance." Seattle: Center on Reinventing Public Education.

Goldhaber, D., and D. J. Brewer, Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources* 32 (3) (1997): 505–523.

Goldrick, L. 2002. "Improving Teacher Evaluation to Improve Teacher Quality." Washington: NGA Center for Best Practices.

Gordon, R. J., T.J. Kane, and D.O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job." Washington: Brookings Institution.

Grogger, J. and E. Eide. "Changes in College Skills and the Rise in the College Wage Premium." *Journal of Human Resources* 30 (2) (1995): 280–310.

Grossman, P. L. and others. 2010. "Measure for Measure: The Relationship Between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores." CALDER Working Paper.

Hamilton, L. S. and others. 2007. "Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States." Santa Monica: RAND Corporation.

Hanushek, E. and others. "Education and Economic Growth." *Education Next* 8 (2) (2008): 62–70.

_____ "The Economics of Schooling - Production and Efficiency in Public-Schools." *Journal of Economic Literature* 24 (3) (1986): 1141–1177.

_____ "The Trade-Off Between Child Quantity and Quality." *Journal of Political Economy* 100 (1) (1992): 84-117.

_____ "Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 1 9 (2) (1997): 141–164.

Harris, D. and T. Sass. 2008. "The Effect of NBPTS-Certified Teachers on Student Achievement." Washington: The Urban Institute.

Hassel, E.A. and B. Hassel. 2009. "3X for all: Extending the Reach of Education's Best." Public Impact White Paper.

Hess, Frederick M. 2009. "The Human Capital Challenge: Toward a 21st-Century Teaching Profession." In D. Goldhaber and J. Hannaway, eds., *Creating a New Teaching Profession*. Washington:Urban Institute Press.

Hoffman, D. A., R. Jacobs, and J.E. Baratta. "Dynamic Criteria and the Measurement of Change." *Journal of Applied Psychology* 78 (2) (1993):194–204.

Hoffman, D. A., R. Jacobs, and S.J. Gerras. "Mapping Individual Performance Over Time." *American Psychological Association* 77 (2) (1992):185–195.

Jacob, B. A. and S.D. Levitt. (2003). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economic* 118 (3) (2003): 843–877.

Jacob, B. and L. Lefgren. "Principals as Agents: Subjective Performance Assessment in Education." *Journal of Labor Economics* 26 (1) (2007): 101–136.

Jacob, B., L. Lefgren, and D. Sims. "The Persistence of Teacher-Induced Learning Gains." *Journal of Human Resources*." (Forthcoming)

Kane, T. J., J.E. Rockoff, and D.O. Staiger. "Photo Finish: Teacher Certification Doesn't Guarantee a Winner." *Education Next* 7 (1) (2007)**:** 60–67.

_____ 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Cambridge: NBER.

Keigher, A. and F. Cross. 2010. "Teacher Attrition and Mobility: Results from the 2008-09 Teacher Follow-Up Survey." Washington: National Center for Education Statistics, available at http://nces.ed.gov/pubsearch.

Kimball, S. M., B. White,and A.T. Milanowski. "Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of Education* 79 (4) (2004): 54–78.

Koedel, C. and J.R. Betts. 2007. "Re-examining the Role of Teacher Quality in the Educational Production Function." San Diego: University of Missouri.

Koedel, C. and J.R. Betts. "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation." *Education Finance and Policy* 5 (1) (2010): 54–81.

Koretz, D. "A Measured Approach: Maximizing the Promise, and Minimizing the Pitfalls, of Value-Added Models." *American Educator* 39 (Fall) (2008): 18–27.

Ladd, H. "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes." *Economics of Education Review* 18 (1) (1999): 1–16.

Ladd, H. "Teacher Labor Markets in Developed Countries." *The Future of Children* 17 (1) (2007): 201–217.

Long, M. 2007. "Teacher Dismissals: A review of the Literature and Thoughts for the Future." Washington: National Center for Education Statistics. Symposium on Data Issues in Teacher Supply and Demand.

Loup, K. S. and others. "Ten Years Later: Findings From a Replication of a Study of Teacher Evaluation Practices in Our 100 Largest School Districts." *Journal of Personnel Evaluation in Education* 10 (1) (1996): 203–226.

Mccaffrey, D. F. and others. "Models for Value-Added Modeling of Teacher Effects." *Journal of Educational and Behavioral Statistics,* 29 (1) (2004): 67–101.

Mccaffrey, D. F. and Lockwood, J. R. 2008. "Value-Added Models: Analytic Issues." Washington: National Research Council and the National Academy of Education, Board on Testing and Accountability, Workshop on Value-Added Modeling.

Mccaffrey, D. F. and others. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4 (4) (2009) 572–606.

Mccarthy, S. J. and K.D. Peterson. "Peer Review of Materials in Public School Teacher Evaluation." *Journal of Personnel Evaluation in Education* 1 (1) (1988): 259–267.

Mcguinn, P. 2010. "Ringing the Bell for K-12 Teacher Tenure Reform." Washington: Center for American Progress.

Medina, Jennifer. "Bloomberg Says Test Scores Will Factor Into Teacher Tenure." *The New York Times*, November 25, 2009.

Medley, D. M. and H. Coker. "How Valid Are Principals' Judgments of Teacher Effectiveness?" *Phi Delta Kappan* 69 (2) (1987): 138–140.

Milanowski, A. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79 (4) (2004): 33–53.

Miller, Raegen. "Adding Value to Discussions About Value Added." Washington: Center for American Progress.

Murnane, R. J. and D.K. Cohen. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive." *Harvard Educational Review* 56 (1) (1986): 1–17.

Murnane, R. J., J.B.Willett, and F. Levy. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77 (2) (1995): 251– 266.

Murnane, R. and J. Steele. "What is the Problem? The Challenge of Providing Effective Teachers for all Children." *The Future of Children,* 17 (1) (2007): 15–44.

National Center on Teacher Quality. 2009. "Teacher Rules, Roles, and Rights." Washington.

New Teacher Project. 2007. "Hiring, Assignment, and Transfer in Chicago Public Schools." Brooklyn.

Painter, S. "Principals' Perceptions of Barriers to Teacher Dismissal." *Journal of Personnel Evaluation in Education* 14 (3) (2000): 254–264.

Painter, S. "Barriers to Evaluation: Beliefs of Elementary and Middle School Principals." *Planning and Changing* 22 (2001): 58–70.

Paulson, Amanda. "How Race to the Top is Recasting Education Reform in America." *The Christian Science Monitor* Tuesday, June 1, 2010.

Peterson, K. D., A. Driscoll, and D. Stevens. "Primary Grade Student Reports for Teacher Evaluation." *Journal of Personnel Evaluation in Education* 4 (1) (1990): 165–173.

Podgursky, M. J. and M.G. Springer. "Teacher Performance Pay: A Review." *Journal of Policy Analysis and Management* 26 (4) (2007): 909–949.

Rivkin, S. G., E.A. Hanushek, and J.F. Kain. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2) (2005): 417–458.

Rockoff, J. E. "The Impact of Individual Teachers on Students' Achievement: Evidence From Panel Data. *American Economic Review* 94 (2) (2004): 247–252.

Rockoff, J. E. and others. "Can You Recognize An Affective Teacher When You Recruit One?" *Education Finance and Policy* Forthcoming.

Rodin, M. and B. Rodin. "Student Evaluations of Teachers." *The Journal of Economic Education* 5 (1) (1973): 5–9.

Rothstein, J. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy* 4 (4) (2009): 537–571.

_____ "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics* 125 (1) (2010): 175–214.

Sanders, W. L. and J.C. Rivers. 1996. "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement." Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Schochet, P. Z. and H.S. Chiang. 2010. "Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains." Washington: National Center for Education Evaluation and Regional Assistance.

Springer, M.G. and others. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Nashville: National Center on Performance Incentives at Vanderbilt University.

Sturman, Michael C., Robin A. Cheramie, and Luke H. Cashen. "The Impact of Job Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test-Retest Reliability of Employee Job Performance Ratings." *Journal of Applied Psychology* 90 (2005): 269–283.

Toch, T. and R. Rothman. 2008. "Rush to Judgment: Teacher Evaluation in Public Education." Washington: Education Sector.

Todd, P. E. and K.I. Wolpin. "On the Specification and Estimation of the Production Function for Cognitive Achievement." *Economic Journal* 113 (485) (2003): F3–F33.

Tucker, P. D. "Lake Wobegon: Where All Teachers Are Competent (Or Have We Come to Terms with the Problem of Incompetent Teachers?)." *Journal of Personnel Evaluation in Education* 11 (1) (1997): 103–126.

Tucker, P. D. and others. "The Efficacy of Portfolios for Teacher Evaluation and Professional Development: Do they make a difference?" *Educational Administration Quarterly* 39 (5) (2003): 572–602.

Tyler, J. H. and others. "Using Student Performance Data to Identify Effective Classroom Practices." *American Economic Review* 100 (2) (2010): 256–260.

Weisburg, D. and others. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act of Differences in Teacher Effectiveness." Brooklyn: New Teacher Project.

Weiss, Joanna. "Education's 'Race to the Top' Begins." *Education Week*, July 23, 2009.

Endnotes

1 The "Lake Wobegon effect" is used to describe the characterization of the great majority of a group as above average.

2 Differences in teacher effectiveness ("effectiveness" and "quality" are used interchangeably here), for instance, swamp the impact of other educational investments such as reductions in class size (See, for instance, D. Aaronson, L. Barrow, and W. Sander, "Teachers and student achievement in the Chicago public high schools," *Journal of Labor Economics* 25 (1) (2007): 95–135; D. Goldhaber, D. Brewer, and D. Anderson, "A Three-Way Error Components Analysis of Educational Productivity," *Education Economics* 7 (3) (1999)**:** 199–208; S.G. Rivkin, E.A. Hanushek, and J.F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73 (2) (2005): 417–458.).

3 This finding may seem counterintuitive to those not steeped in education research, but it is now a widely replicated finding, see D. Goldhaber, D. Brewer, and D. Anderson, "A Three-Way Error Components Analysis of Educational Productivity," *Education Economics* 7 (3) (1999)**:** 199–208 or Hanushek and Riven, 2010.

4 See, for instance, research by Charles Clotfelter, H.F. Ladd, and J.L. Vigdor, "Teacher-Student Matching and the Assessment of Teacher Effectiveness," *Journal of Human Resources* 41 (4) (2006): 778–820; and S.G. Rivkin, E.A. Hanushek, and J.F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica* 73 (2) (2005): 417–458.

5 See Figure 4 in R.J. Gordon, T.J. Kane, and D.O. Staiger, "Identifying Effective Teachers Using Performance on the Job" (Washington: Brookings Institution, 2006).

6 For a more comprehensive discussion, see work I did with Dominic Brewer (1997) or reviews by Eric Hanushek, "The Economics of Schooling - Production and Efficiency in Public-Schools," *Journal of Economic Literature* 24 (3) (1986): 1141–1177; and Eric Hanushek, "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis* 1 9 (2) (1997): 141–164.

7 For example, only 15 states require new teachers to be observed more than once a year. For more information on state policies on teacher evaluations, see the National Center on Teacher Quality's TR^3 database: http://www.nctq.org/tr3.

8 S. Painter, "Barriers to Evaluation: Beliefs of Elementary and Middle School Principals," *Planning and Changing* 22 (2001): 58–70.

9 See T. Toch, and R. Rothman, "Rush to Judgment: Teacher Evaluation in Public Education" (Washington: Education Sector, 2008) for a more comprehensive discussion.

10 The use of principal evaluations for high-stakes decisions, however, does not have a good track record of success (in terms of the longevity of a program). In particular, reforms from the 1980s that tied principal evaluations to high-stakes decisions, like merit pay, tended not to last long either because teachers saw them as subjective and arbitrary or because so many teachers were awarded a bonus that the programs became prohibitively expensive (R.J. Murnane and D.K. Cohen, "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive," *Harvard Educational Review 56* (1) (1986): 1–17.)

11 For more background, see D.M. Medley and H. Coker, "How Valid Are Principals' Judgments of Teacher Effectiveness?" *Phi Delta Kappan* 69 (2) (1987) 138–140, who cite a long history of studies that find little to no relationship between the ratings assigned by principals and student learning.

12 In S.M. Kimball, B. White,and A.T. Milanowski, "Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County, " *Peabody Journal of Education* 79 (4) (2004): 54–78, the authors find that a 1-point increase in teacher evaluation scores is associated with a 5.41 point increase in student performance. In H.A. Gallagher, "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education* 79 (4) (2004): 79–107, the author finds correlations from .18 to .50 between VAM scores in different subjects and evaluation scores. In A Milanowski, "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79 (4) (2004): 33–53, the author finds similar correlations between .27 and .43 in different subjects.

13 These differences are only statistically significant for two elements of instruction: "explicit strategy instruction" and "guided practice."

14 For more detail on the states and localities where NBPTS teacher receive additional compensation, see http://www.nbpts.org/resources/state_local_information.

15 See, for instance, studies by: S. Cantrell and others, "National Board Certification and Teacher Effectiveness: Evidence From a Random Assignment Experiment" (Cambridge: Harvard Graduate School of Education, 2008); L.C. Cavalluzzo, "Is National Board Certification an Effective signal of Teacher Quality?" (Alexandria: CNA Corporation, 2004); D. Goldhaber and E. Anthony, "Can Teacher Quality be Effectively Assessed?" National Board Certification as a Signal of Effective Teaching," *Review of Economics and Statistics* 89 (1) (2007): 134–150; and D. Harris and T. Sass, "The Effect of NBPTS-Certified Teachers on Student Achievement" (Washington: The Urban Institute, 2007).

16 For more detail on these calculations, see D. Goldhaber, "National Board Teachers Are More Effective, But Are They in the Classrooms Where They're Needed the Most?" *Education Finance and Policy* 1 (3) (2006): 372–382.

17 Most of the research on peer and self-review of instruction is more concentrated at the postsecondary level with little to no work connecting the evaluation to K-12 student achievement (J.A. Centra, "Colleagues as Raters of Classroom Instruction," *The Journal of Higher Education* 46 (3) (1975): 327–337; P.A. Cohen and W.J. Mckeachie, "The Role of Colleagues in the Evaluation of College Teaching," *Improving College and University Teaching* 28 (4) (1980): 147–154; K.A. Feldman, "Instructional Effectiveness of College Teachers as Judged by Teachers Themselves, Current and Former Students, Colleagues, Administrators, and External (Neutral) Observers," *Research in Higher Education* 30 (2) (1989): 137–194). An example of one study exploring the use of peer evaluations at the elementary and secondary level is S.J. Mccarthy and K.D. Peterson, "Peer Review of Materials in Public School Teacher Evaluation," *Journal of Personnel Evaluation in Education* 1 (1) (1988) 259–267.They focus on a career ladder program in the Salt Lake City School District and compare peer and principal approval

of a teacher's submitted materials, the ratings aligned in 39 of the 50 cases with principals more likely to approve of the materials than other teachers. For an example of educational research that uses the results of student surveys, see R. Ferguson, "Raising Achievement and Closing Gaps in Whole School Systems: Recent Advances in Research and Practice" (Cambridge: Harvard University, Annual Conference of the Achievement Gap Initiative, 2008).

18 For more information on this, see Bill and Melinda Gates Foundation, "Working With Teachers to Develop Fair and Reliable Measures of Effective Teaching," (2010).

19 For more on how U.S. students compare with students from other countries in math and science performance, see the results of the 2007 Trends in International Math and Science Study (TIMSS): http://nces.ed.gov/timss/; and for a more in-depth discussion of teacher policies in other countries see D. Goldhaber, "Lessons from Abroad: Exploring Cross-Country Differences in Teacher Development Systems and What They Mean for U.S. Policy." In D. Goldhaber and J. Hannaway, eds., *Creating a New Teaching Profession (*Washington: Urban Institute Press, 2009); H. Ladd, "Teacher Labor Markets in Developed Countries," *The Future of Children,* 17 (1) (2007): 201–217; and R. Murnane and J. Steele. "What is the Problem? The Challenge of Providing Effective Teachers for all Children." *The Future of Children,* 17 (1) (2007): 15–44.

20 See, for instance, P. Mcguinn, "Ringing the Bell for K-12 Teacher Tenure Reform" (Washington: Center for American Progress, 2010); and Jennifer Medina, "Bloomberg Says Test Scores Will Factor Into Teacher Tenure," *The New York Times*, November 25, 2009.

21 D. Goldhaber and M. Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions," *American Economic Review: Papers and Proceedings* 100 (2) (2010): 250–255.

22 Credible value-added estimates require statistical evaluation of changes in test scores or covariate adjustments that include a control for a base year of achievement. For more on this, see J. Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4 (4) (2009): 537–571; P.E Todd and K.I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* 113 (485) (2003): F3–F33.

23 For more on the limitations of standardized tests, see D. Koretz, "A Measured Approach: Maximizing the Promise, and Minimizing the Pitfalls, of Value-Added Models," *American Educator* 39 (Fall) (2008): 18–27.

24 See, for instance, J. Grogger and E. Eide, "Changes in College Skills and the Rise in the College Wage Premium," *Journal of Human Resources* 30 (2) (1995): 280–310, and R.J. Murnane, J.B.Willett, and F. Levy, "The Growing Importance of Cognitive Skills in Wage Determination," *Review of Economics and Statistics* 77 (2) (1995): 251– 266 on the link between individual test achievement and post-secondary success; and Eric Hanushek and others, "Education and Economic Growth," *Education Next* 8 (2) (2008): 62–70 on the relationship between student achievement and aggregate country growth rates. See L.S. Hamilton and others, "Standards-Based Accountability Under No Child Left Behind: Experiences of Teachers and Administrators in Three States" (Santa Monica: RAND Corporation, 2007) for evidence on teachers focusing on test-taking skills,and Jacob, B. A. and S.D. Levitt. (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economic* 118 (3) (2003): 843–877 for evidence on teacher cheating.

25 For instance, see P.E. Todd and K.I. Wolpin, "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal* 113 (485) (2003): F3–F33 on the assumptions about the nature of student learning over time; see D. Ballou, W.L. Sanders, and P.S. Wright, "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Education and Behavioral Statistics* 29 (1) (2004): 37–65 and D.F. Mccaffrey and others, "Models for Value-Added Modeling of Teacher Effects," *Journal of Educational and Behavioral Statistics,* 29 (1) (2004): 67–101, on the implications of using student background covariate adjustments; and see T. J. Kane, J.E. Rockoff, and D.O. Staiger, "Estimating Teacher Impacts on Student Achievement: An experimental evaluation," (Cambridge: NBER, 2008) and J. Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4 (4) (2009): 537–571 on the potential that VAM teacher effect estimates may be biased by the nonrandom sorting of teachers and students in schools.

26 For a more thorough discussion, see H. Borko, "Professional Development and Teacher Learning: Mapping the Terrain," *Educational Researcher* 33 (8) (2004): 3–15 or T. Toch and R. Rothman, "Rush to Judgment: Teacher Evaluation in Public Education." (Washington: Education Sector, 2008).

27 Louisiana is one of the few states able to do this due to the Louisiana Education Accountability Data System (LEADS).

28 This idea is discussed by Frederick M. Hess, "The Human Capital Challenge: Toward a 21st-Century Teaching Profession." In D. Goldhaber and J. Hannaway, eds., *Creating a New Teaching Profession* (Washington:Urban Institute Press, 2009) and E.A. Hassel and B. Hassel, "3X for all: Extending the Reach of Education's Best." Public Impact White Paper (2009).

29 Specifically, for math and reading teachers in fourth through eighth grades, 50 percent of the evaluation score is based on individual teacher value-added and 40 percent is based on observational ratings, while for teachers in other subjects 80 percent of the evaluation score is based on observational ratings with another 10 percent based on results of a teacher-selected assessment. All teachers are also graded on "core professionalism" such as timeliness and absences, as well as on overall school performance. About 16 percent of D.C.'s teaching corps received the top rating this year, and are therefore eligible to receive performance bonuses.

30 The 2007-08 SASS survey (Keigher and Cross 2010) indicates that school districts dismiss about 2.1 percent of teachers for poor performance. See M. Long, "Teacher Dismissals: A review of the Literature and Thoughts for the Future" (Washington: National Center for Education Statistics. Symposium on Data Issues in Teacher Supply and Demand, 2007) for a review of the evidence on teacher dismissals.

31 Chait reports that Illinois school districts that hired outside legal help spent an average of $219,504.21 in legal fees for dismissal cases and related litigation. This figure includes pending cases, so likely understated the actual per-case cost.

32 M. Long, "Teacher Dismissals: A review of the Literature and Thoughts for the Future."

33 D. Goldhaber, B. Gross, and D. Player, "Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping their Best?" *Journal of Public Policy and Management* (Forthcoming).

34 Joanna Weiss, "Education's 'Race to the Top' Begins," *Education Week*, July 23, 2009.

35 Tennessee's Race to the Top application mandates that at least 50 percent of a teacher or principal's evaluation must be based on student achievement data—35 percent on teacher value-added, and 15 percent on other measures of student achievement. The act also recommends local school boards only grant tenure to teachers who are deemed "effective" on the new evaluation system.

36 Delaware will modify its long-standing teacher evaluation system—based on the Charlotte Danielson method—to mandate that teachers cannot be rated as effective or better unless they have demonstrated "satisfactory levels of student growth."

37 Amanda Paulson, "How Race to the Top is Recasting Education Reform in America," *The Christian Science Monitor* Tuesday, June 1, 2010.

38 There is, of course, measurement error in these estimates, but when they account for this by simulating layoffs in 2007 and exploring how that would affect student achievement in 2009, the difference is still found to be 12 percent of a standard deviation of student achievement.

39 See, for instance, Aaronson, Barrow, and Sander, "Teachers and Student Achievement in the Chicago Public High Schools"; D. Ballou, "Value-Added Assessment: Controlling for Context with Misspecified Models" (Washington: Urban Institute Longitudinal Data Conference, 2005); Goldhaber and Hansen, "Using Performance on the Job to Inform Teacher Tenure Decisions"; C. Koedel and J.R. Betts, "Re-examining the Role of Teacher Quality in the Educational Production Function" (San Diego: University of Missouri, 2007).; D.F Mccaffrey and others, "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4 (4) (2009) 572–606; Rivkin, Hanushek, and Kain, "Teachers, Schools, and Academic Achievement."

40 Work I have completed with Michael Hansen (D. Goldhaber and M. Hansen, "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance" (Seattle: Center on Reinventing Public Education, 2010) shows that the correlations corrected for measurement error are .60 for reading and .69 for math.

41 Research in Ballou, "Value-Added Assessment: Controlling for Context with Misspecified Models," for instance, finds that—using a single year's worth of data to inform the teacher-effect estimate—only 2.5 percent of primary and 7.6 percent of middle school teacher effects were different from the average effect in reading; in math the corresponding figures are 17 percent and 30 percent for primary and middle school teachers respectively.

42 For instance, teachers falling on either side of a cut-point (e.g. teachers at the 74th and 75th percentiles, in the case where teachers are rewarded financially for being in the top quartile of the performance distribution).

43 Research in D.F Mccaffrey and others, "The Intertemporal Variability of Teacher Effect Estimates," calculating teacher effects based on a single year of student-teacher data (as opposed to, for instance, using VAM effect estimates that are based on several years of student achievement to inform teacher effect estimates) finds that 28 to 39 percent of teachers who are ranked in the top quintile of performance in one year remain in the top quintile in the next, and 7–15 percent of these teachers move from the top quintile to the lowest quintile.

44 See Goldhaber and Hansen, "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance" and C. Koedel and J.R. Betts, "Value-Added to What? How a Ceiling in the Testing Instrument Influences Value-Added Estimation, " *Education Finance and Policy* 5 (1) (2010): 54–81. Also, Koedel and Betts argue that multi-year estimates have the additional benefit of having a lower probability of being biased.

45 One standard deviation above the mean puts teachers at the 83rd percentile of performance and one standard deviation below the mean puts teachers at the 17th percentile. And, this one standard deviation change in teacher performance is roughly equivalent to 0.2 standard deviations of student achievement or three months of typical learning growth.

46 For instance, the productivity of realtors, insurance salespeople, and professional baseball players. For more on this, see D.A. Hoffman, R. Jacobs, and S.J. Gerras, "Mapping Individual Performance Over Time," *American Psychological Association* 77 (2) (1992):185–195; D.A. Hoffman, R. Jacobs, and J.E. Baratta, "Dynamic Criteria and the Measurement of Change," *Journal of Applied Psychology* 78 (2) (1993):194–204; Goldhaber and Hansen, "Is It Just a Bad Class? Assessing the Stability of Measured Teacher Performance"; Mccaffrey and others, "The Intertemporal Variability of Teacher Effect Estimates"; and Michael C. Sturman, Robin A. Cheramie, and Luke H. Cashen, "The Impact of Job Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test-Retest Reliability of Employee Job Performance Ratings," *Journal of Applied Psychology* 90 (2005): 269–283.

47 For instance, teachers may become more effective because they have a professional experience that enhances their classroom abilities, or less effective because they lose the motivation to teach.

48 One might still argue, however, for its use in an incentive program.

49 Teachers in North Carolina are tenured after their forth year of consecutive service in the same school district.

50 Indeed some proportion are likely to have been counseled out of the profession (Goldhaber, Gross, and Player, "Teacher Career Paths, Teacher Quality, and Persistence in the Classroom: Are Public Schools Keeping their Best?")

51 The Nashville study was primarily concerned with whether performance incentives increased the effort level of those teachers currently in the workforce.

About the author

Dan Goldhaber is the director of the Center for Education Data & Research and a Professor in Interdisciplinary Arts and Sciences at the University of Washington-Bothell. He is also an affiliated scholar at the Urban Institute, and the co-editor of *Education Finance and Policy*. Goldhaber previously served as an elected member of the Alexandria City School Board from 1997-2002, and as an associate editor of *Economics of Education Review.*

Goldhaber's work focuses on issues of educational productivity and reform at the K-12 level, with a current focus on the broad array of human capital policies that influence the composition, distribution, and quality of teachers in the workforce. Topics of published work in this area include studies of the stability of value-added measures of teachers, the effects of teacher qualifications and quality on student achievement, and the impact of teacher pay structure and licensure on the teacher labor market. Previous work has covered topics such as the relative efficiency of public and private schools, and the effects of accountability systems and market competition on K-12 schooling.

Acknowledgements

The Center for American Progress thanks the Bill & Melinda Gates Foundation for generously providing support for this paper. The author is indebted to Roddy Theobald and Steven Dieterle who provided excellent research assistance. The author would also like to acknowledge Robin Chait, Cindy Brown, Bryan Hassel, and Jennifer Steele, for their thoughtful review and feedback. All opinions expressed in this paper represent those of the author and do not necessarily reflect those of the Center for Education Data & Research or the University of Washington. All errors in this paper are solely the author's responsibility.

The Center for American Progress is a nonpartisan research and educational institute dedicated to promoting a strong, just and free America that ensures opportunity for all. We believe that Americans are bound together by a common commitment to these values and we aspire to ensure that our national policies reflect these values. We work to find progressive and pragmatic solutions to significant domestic and international problems and develop policy proposals that foster a government that is "of the people, by the people, and for the people."

Center for American Progress

1333 H STREET, NW, 10TH FLOOR, WASHINGTON, DC 20005 • TEL: 202-682-1611 • FAX: 202-682-1867 • WWW.AMERICANPROGRESS.ORG