



Designing High Quality Evaluation Systems for High School Teachers

Challenges and Potential Solutions

John H. Tyler November 2011

Center for American Progress



Designing High Quality Evaluation Systems for High School Teachers

Challenges and Potential Solutions

John H. Tyler November 2011

Contents

1 Introduction and summary

5 Practice-based teacher evaluation

8 Practice-based evaluation in high school

13 Using student performance measures to evaluate teachers

17 Evaluating high school teachers based on student performance:
Challenges and solutions

23 Other student performance measures

27 Conclusion

31 About the author and acknowledgements

32 Endnotes

35 References

Introduction and summary

A central part of education reform today is the wide-ranging and unprecedented effort to either revamp existing teacher evaluation systems or develop and implement entirely new systems. In the past three years, for example, 32 states and the District of Columbia have made some change to their state teacher evaluation policy, and 23 states currently require that teacher evaluations include objective evidence of student learning, up from only four states in 2009.¹ The success of this work will in large part be judged by the extent to which the resulting systems can evaluate teachers with rigor, objectivity, and in ways that differentiate teachers' abilities to promote student learning.

Meeting this high bar in our nation's high schools poses especially difficult challenges, and yet the stakes for doing so are enormous, a point brought home by the extant research. One particular strand of research focuses attention to the importance of identifying and addressing teacher effectiveness *within schools*, where the bulk of the variation in teacher effectiveness resides.² At the same time, research indicates a clear and urgent need to accomplish this task in our nation's high schools.

The argument for focusing attention at the high school level is three-pronged. First, the performance of high school students lags behind that of demographically similar students in the elementary and middle grades, which suggests that, relative to the earlier years, there is a heightened need for improving the quality of instruction in high school.³ Second, dropout decisions are made by students in their high school years, which means improving average teacher quality in high school is one potential avenue for addressing the stubbornly persistent dropout rate. The research-based linkage is that student engagement is related to dropping out and teachers' behaviors and practices are, in turn, related to student engagement.⁴

Third, high school is our last line of defense for preparing students for college and the world of work—and teachers are an obviously critical component of the quality of that preparation. Students entering college lacking a solid high school educa-

tion often have to spend time in remedial college courses, a sidetrack associated with fewer earned credits and a lower likelihood of graduating with a degree.⁵ In terms of labor market consequences, young people who enter today's labor force without basic academic skills, the ability to think critically and creatively, and who are deficient in so-called "noncognitive" skills are at a competitive disadvantage in the global, information-age economy.⁶

High-quality teacher evaluation systems are seen as one lever for improving the teacher workforce and hence the outcomes of students, including high school students. The current degree of consensus around efforts to improve teacher evaluation is striking for the world of kindergarten-through-12th-grade education. From traditionally conservative education observers and activists to teachers' union leaders, from professional education organizations to philanthropic foundations, and from the U.S. Department of Education to local education agencies, a wide array of individuals, groups, and organizations are involved and often cooperating in efforts to support the design, testing, and implementation of the next generation of teacher evaluation systems.⁷

On the public side the U.S. Department of Education made teacher evaluation an integral part of the Obama administration's \$4.3 billion Race to the Top competitive grant initiative designed to encourage and reward states that are creating the conditions for education innovation and reform. Meanwhile, in the nonprofit arena some of the nation's most prominent foundations are reallocating grant money toward teacher evaluation initiatives, one example of which is the Bill & Melinda Gates Foundation investment of \$290 million in four "intensive partnership" sites to support teacher effectiveness initiatives that include teacher evaluation and another \$45 million in the Measures of Effective Teaching, or MET, project, a two-year effort to develop methods and tools for identifying and developing good teaching.

As the various teacher evaluation initiatives move forward over the coming years, how they play out will likely be shaped by a simple but important contextual reality. While there is wide agreement about the need for new and better ways to evaluate teachers, different stakeholders place different emphases on what they ultimately want from evaluation systems. Some see teacher evaluation as a way to identify and remove low-performing teachers. Others view teacher evaluation as the cornerstone of new performance-based teacher compensation systems. Still others think that the emphasis should be on evaluation as a mechanism for improving teaching practice, a way to help teachers get better. At the end of the

day, however, the extent to which any of this can happen rests on evaluation that can consistently determine who are the more and less effective teachers in our classrooms.

Information to accomplish this comes from two sources. First, we can use teacher-related *inputs* to the education process, such as classroom teaching observations or classroom artifacts such as lesson plans and teacher-designed student assessments. From these practice-based measures, we make inferences about a teacher's ability to promote student learning. Second, we can measure *outputs* from the teaching-learning process—actual student performance—and, based on these measures, make inferences about the teacher's contribution to that output. In each case doing this well for high school teachers is a challenge.

In terms of practice-based measures of effectiveness, the many content areas that are covered in a typical comprehensive high school make it impossible for all teachers to be observed and evaluated by individuals who have training in the teacher's content area. This not only potentially compromises the validity and reliability of the evaluation results; it decreases the likelihood that teachers will buy into and support the evaluation system.

Likewise, the difficulties in using student performance data to evaluate high school teachers begins with the fact that these teachers rarely teach in grades or subjects where students have had comparable pre- and post-tests that can be used to construct prototypical value-added measures for the teacher. Another issue in using value-added measures at the high school level is that, unlike the case for elementary students, we have to worry about the fact that students in, say, an 11th-grade English course took different paths to get to that course.

If these different paths affect their outcomes, then value-added models that do not account for this “path dependence” may not accurately estimate the teacher or record's contribution to student learning. A similar problem is present if teachers affect learning across courses in a given year. Failure to account for this kind of “cross-fertilization” would again call into question value-added measures of teacher effectiveness.

Thus, there are clear challenges to effectively evaluating high school teachers. Nevertheless, states and school districts across the nation are confronting these challenges and in the process solutions are emerging. A preview of the potential solutions that the analysis in this paper suggests may be employed in building optimum evaluation systems for high school teachers includes:

- Developing new and enhancing existing assessments that test high school teachers' content-based pedagogical knowledge
- Exploring, developing, and testing the increased use of technology such as classroom video recording as a means for generating efficiency and productivity gains in practice-based evaluation
- Conducting more research on the properties and use of Student Learning Objectives, or SLOs, as a measure of effective teaching based on student performance
- Continuing investigations into how value-added measures can be effectively used at the high school level
- Finding the best ways to incorporate all available information from both practice-based measures and student performance data into the ultimate evaluation of teachers

This paper examines the challenges and potential solutions to evaluating high school teachers, looking first at practice-based evaluation and then turning to student performance as the basis for evaluation. In each case the stage is first set with a brief discussion of the overarching, across-grade issues that accompany each method.

In reviewing the issues at hand, it is important to keep in mind that these two models of evaluation, practice-based and student-performance-based evaluation, make inferences based on different points in the education process—input versus output, and they rely on different kinds of data—qualitative and more subjective versus quantitative and objective. And they are at different stages of developmental evolution—well-established for many years (though evolving) for practice-based evaluation versus rapid developments over the last 10 years in using student performance data for evaluation. Nevertheless, the early evidence is that most new evaluation systems will be characterized by some combination of both of these methods to evaluate teachers, including high school teachers.

Practice-based teacher evaluation

Practice-based teacher evaluation refers to the use of information from various aspects of a teacher's practice in evaluating teacher effectiveness. Practice-based information on teacher effectiveness can come from:

- Classroom observations
- Teacher-produced artifacts such as lesson plans, teacher-designed student assessments, and portfolios
- Evidence of how a teacher works with colleagues and administrators
- Communications with parents
- Professional development activities
- Other evidence of a teacher's professional activities

Of these elements, classroom observations of the teacher at work are arguably the most direct evidence of a teacher's ability to affect student learning. This would suggest that evaluation that uses some or all of the above components should give classroom observations the most weight in computing a final summative score for a teacher. Unfortunately, there is little information regarding how districts currently weight the different measures when they combine them into a summative evaluation score for the teacher.⁸

An important feature of practice-based evaluation to bear in mind is that it is based on input measures into the teaching-learning process rather than output measures directly associated with student learning and achievement. The argument for input-based evaluation is that teachers are evaluated on practices that education experts believe to be related to student learning, an argument that has less resonance in today's environment of education accountability than it might once have had. There are three practice-based evaluation protocols that have been linked to student achievement growth but the great majority of districts use evaluation protocols that have not been validated against student achievement gains.⁹

Until very recently, practice-based measures were the only available means in most districts for evaluating teachers, and while a small handful of districts are beginning to incorporate output measures into evaluation systems, using input measures of teachers' practice is still the way that virtually all districts currently judge teacher effectiveness.¹⁰ In most districts the responsibilities associated with carrying out practice-based evaluation lie with building principals, assistant principals, curriculum heads, and other school-level administrators.¹¹ A major part of their evaluation responsibilities is conducting the classroom observations of teachers, a time-intensive process. The instruments used to guide the evaluation process—the assessment forms, performance criteria, and evaluation procedures—are largely created by the districts, a process that leads, not surprisingly, to considerable variation across districts in the structure, use, and likely quality of the evaluation.¹² Some of the problems with principal-led evaluations, in particular the tendency to rate teachers higher on the evaluations than is likely warranted, so-called “leniency bias,” are detailed in a 1997 paper by Pamela D. Tucker, where principals informally identified 5 percent of their instructional staff as being “incompetent” but gave only 1 percent formal ratings this low.¹³

Leniency bias on the part of principals is not a problem unique to evaluations of teachers that come from principal-led evaluations.¹⁴ “The Widget Effect,” a 2009 study by The New Teacher Project of 12 districts that are considered to have some of the most well-developed evaluation systems in the nation, including systems that utilize trained evaluators for classroom observations and other activities, replicated Tucker's results of a decade earlier. Among the 12 districts that utilized a binary rating system for teachers, 99 percent of the teachers received the highest rating. Meanwhile, in those districts that utilized a four-point rating scale, 94 percent of the teachers received one of the top two ratings, and less than 1 percent were rated at the lowest level.¹⁵ Of course, the problem with this is that when virtually all of the teachers receive the highest scores, an evaluation system foregoes the opportunity to meaningfully differentiate teacher effectiveness in ways that could provide valuable information to the district and to teachers in need of assistance.

In addition to the failure to use existing evaluation to differentiate among the variation in teacher effectiveness within schools that research consistently verifies, another issue is that high-quality practice-based evaluation can be relatively costly. The Cincinnati school district, for example, which is widely recognized as having one of the best evaluation systems in the country, allocated between \$1.8 million and \$2.1 million per year for teacher evaluation between the 2004-05 and 2009-10 school years.¹⁶ This translates into about \$7,500 per teacher for each of

the approximately 250 teachers who underwent a comprehensive evaluation each year.¹⁷ Approximately 90 percent of this cost is attributable to the salaries of the 12-15 master teachers who leave the classroom for three years to serve as trained evaluators in the Cincinnati system.¹⁸

While issues that potentially compromise or complicate the use of practice-based evaluation get manifested at the high school level, so too do the advantages of using practice-based evaluation. In particular, when done well, practice-based evaluation can provide information to teachers and administrators regarding what good teachers are doing well, and therefore could be shared with colleagues, and what struggling teachers are failing to do well and need assistance to improve. Also, advocates of practice-based evaluation suggest that because the rubrics against which a teacher's practice will be judged are known and available in public documents, discussion around these explicit teaching practices can foster dialogue in a school and a district around what good teaching looks like. In this case the evaluation system can change the discussion around teaching. It is also the case that teachers and teachers' unions, including high school teachers, tend to be more supportive of practice-based evaluations relative to evaluations based on student test scores.

Finally, there is emerging evidence from a Cincinnati-based study by Eric S. Taylor and John H. Tyler that suggests well-designed practice-based evaluation can help a teacher get better.¹⁹ Linking 10 years of data from the Cincinnati practice-based system to student test score data in that district, Taylor and Tyler find that teachers who go through a comprehensive evaluation are more effective at promoting student achievement growth than they were in the years prior to being evaluated, controlling for the effects of experience on teacher effectiveness. The research shows that not only are teachers more effective in the year of the evaluation, but the same teachers are even more effective at promoting student learning in the years following evaluation. This research suggests that there are substantial human capital gains associated with going through a rigorous and high-quality evaluation process.

What's more, the size of the estimated effect is substantial. A student taught by a teacher after that teacher participates in the Cincinnati evaluation program will score about 10 percent higher in math than a similar student taught by the same teacher before that teacher was evaluated. If those two students began their respective years at the 50th percentile of math achievement, the student who was taught after the teacher went through evaluation would score at about the 55th percentile at the end of the year while the other student would remain at the 50th. The study is unable, however, to identify the mechanisms of the evaluation process that lead to the gains

in teacher quality, but a prime candidate is the feedback that teachers receive after each of the four classroom observations during the year.

Most discussions around evaluation today focus on their use to identify low- and high-performing teachers so that the poor performers can be removed from the classroom or provided with extra help and, in some few instances, the high-performing teachers can be rewarded via a merit-based compensation system. The possibility raised by the Taylor and Tyler study that rigorous practice-based evaluation can help make teachers better, effectively serving as professional development, adds an important dimension to teacher evaluation reform.²⁰

Practice-based evaluation in high school

Challenges and solutions

The challenge in designing high-quality, practice-based evaluation systems for high school teachers begins with the content specialization embodied in high school education. Unlike elementary school, students in high school move from class to class, from teacher to teacher, and each teacher presumably (though not always) possesses specialized content knowledge. This structure poses real issues for practice-based evaluation of teachers since in the ideal it requires evaluators who are also content specialists. Thus, the high school principals and assistant principals tasked with evaluation in most districts would ideally be well-versed in the content being taught by each teacher they are evaluating as well as in the best practices for teaching that content. Given the number of different courses and content areas taught in a typical comprehensive high school, this ideal is not realistic.

In districts that supplement or supplant principal-led evaluations with master-teacher evaluators, as is done for example in Cincinnati and Washington, D.C., it is more likely that content expertise between the evaluator and the teacher being evaluated can be matched. Even in this situation, however, it is not feasible to have an evaluator with content specialization for every subject covered in high school. A typical comprehensive high school, for example, can have up to 50 different courses that are offered and taught, and when career and technical education courses are included, the number of different courses across content areas can be 100 or more.²¹

Some of the concern regarding content specialization can be mitigated by the nature of the evaluation protocol that is used to evaluate and score teachers' practice. The Danielson Framework for Teaching, for example, is designed around the concept that good teaching looks the same across grades and subjects and thus the rubrics in the framework are appropriate for all teachers, including high school teachers. The Classroom Assessment Scoring System-Secondary, or CLASS-S, tool, the newest protocol from the CLASS-family out of the University of Virginia, is designed for use across subjects in high school. Thus, theoretically, an evaluator lacking expertise in a given content area should be able to evaluate high school teachers across subjects using these tools.²²

Using content-neutral evaluation tools does not, however, solve all of the problems, the most important of which may be teacher buy-in. Consider the case of a calculus teacher who might be evaluated by someone with the closest available expertise—perhaps a individual with a math background and math certification—but who has never taught calculus. Even if the evaluation rubric guides the evaluator on general instructional behaviors that should be seen and scored in a calculus classroom, one can imagine the teacher asking “How can I be fairly evaluated by someone who has never before taught calculus?” This situation suggests a less-than-ideal dynamic between the teacher and evaluator, a dynamic that in the best of situations may inhibit how the teacher values any constructive feedback from the evaluator and in the worst of cases may lead to a union-backed grievance if the evaluation comes in low and there are negative consequences for the teachers.

There are, however, ways to at least partially address the overarching problem. When the principal is not the primary evaluator in the system, districts can attempt to spread their evaluation expertise across content areas when selecting and training evaluators, knowing there will still be some subjects where there is content mismatch between the evaluator and the teacher. When the system relies on building administrators to conduct the evaluation, the model could be adapted so that, for example, at least one classroom observation is carried out by someone in the district with content expertise that matches that of the teacher under evaluation. This will increase the logistical burden of the evaluation system and potentially the cost, but the payoff in terms of teacher and union buy-in could be worth it in the long run. Even so, given the number of different content areas in a comprehensive high school, it is not realistic to think that districts will ever be able to provide content-specific evaluation for all teachers under evaluation.

One possible way to address this problem is for districts to use direct measures of content pedagogy—assessments given to teachers that would measure how well a teacher can teach their specific content. While a wide range of such assessments is not currently available, pioneering work by the Learning Mathematics for Teaching, or LMT, project at the University of Michigan has led to the development of such assessments for math in grades K–8. One element of the aforementioned Gates Foundation’s MET project is to develop content knowledge assessments for teaching math and English language arts. Similar work on science topics is being conducted by the Assessing Teacher Learning About Science Teaching collaborative project between Horizon Research, Inc., and the American Association of the Advancement of Science.²³ Both the LMT and MET projects are working to validate their assessments by establishing the links among content knowledge, observed instructional practice, and student achievement.

Given the K–8 focus of the LMT work and the fact that the only high school teachers targeted by the MET project are Algebra I, ninth-grade English, and high school biology teachers, a collection of content knowledge assessments for use at the high school level is currently not available to districts. That being the case, developing content knowledge assessments for the many content areas and courses taught in the typical high school is a challenge to be met. If content-specific pedagogical assessments could be developed, however, they could potentially be used in concert with content-neutral classroom observation tools to evaluate high school teachers. Such a blended approach could measure a teacher’s broad and general abilities, her content-specific abilities, and at the same time potentially increase teacher buy-in.

The use of digital video technology to capture teaching episodes is another area of promise for addressing high school teacher evaluation issues. As a part of the MET project, a technological innovation being explored is the use of a digital video camera that is set up and operated by the teacher and that allows a simultaneous 360-degree panoramic recording of everything happening in the classroom as well as fixed-position recording of everything that is shared on the classroom blackboard.²⁴ In the MET project the videos are uploaded and coded by trained evaluators hired to work with the project. Project plans call for several uses for the resulting video library, including comparing results across different scoring rubrics and raters, and correlating the evaluator-based scores with student achievement gains.

If video recording of teaching catches on in the field, the results could be far reaching in terms of addressing the high school content-specialization problem. Districts would be able to share secure versions of the digitized teaching episodes with content-specialist

evaluators who could be located anywhere. A district could then have its own core of evaluators who score video files in a central location or even in their homes. One could also imagine districts sharing the costs and efforts of evaluators. One district might not be able to keep, for example, an evaluator of dance instruction busy, but evaluating dance instruction might be a full-time job for a trained evaluator employed across several districts. This concept naturally leads to the possibility of the development of markets for trained evaluators across content areas.

There are other advantages to high-quality classroom teaching videos. First, there are likely substantial cost savings. The cameras used in the MET project currently cost about \$4,500 each and the software license can cost an additional \$150 per teacher evaluated.²⁵ Using Cincinnati as an example, if 12 full-time evaluators in the Cincinnati system were replaced with 12 video cameras and the 250 teachers were evaluated via a new video system, the first-year costs for the camera and licensing would be about \$370 per evaluated teacher. There would, of course, be additional costs associated with paying evaluators to view and score the videos, along with other support such as technical support for the operation and some teacher training on how to set up and use the equipment. Given that Cincinnati currently spends \$7,500 per evaluated teacher, there is a lot of per teacher money left for these additional costs after the initial \$370 video investment. Also, the largest part of the \$370 per evaluated teacher investment, the \$4,500 purchase cost per camera, is not a cost that recurs every year.

Another advantage to a video approach to classroom observations is that since the evaluator has the option to rerun any portion of the video, there is the opportunity to review parts of the episode as might be needed, leading to greater scoring accuracy. The opportunity for multiple evaluators to score the same teaching episode would also allow for increased scoring accuracy and greater objectivity relative to “one-shot” in-person classroom observations.²⁶ Moreover, self-reflective teachers would have the opportunity to watch their own teaching episodes and learn from them. Finally, over time, a district could build a library of exemplary teaching episodes across all content areas for use in professional development activities, particularly with new teachers.

What is not yet clear is whether teachers who currently resist the validity of being evaluated by out-of-content evaluators will have a different attitude toward being evaluated on video by out-of-district evaluators who would potentially have little or no contextual information to accompany the videotaped teaching episode. Nevertheless, given the direct personnel costs and inevitable content mismatch

associated with the current practice-based methods of evaluating high school teachers, the increased use of technology in teacher evaluation holds substantial promise, particularly when it comes to addressing the challenge of effectively evaluating high school teachers across the curriculum.²⁷

While evaluating high school teachers using performance-based measures poses challenges to school districts, the challenges are not insurmountable. The real question in the years ahead will not be whether we have the tools and methods for evaluating high school teachers in valid and reliable ways, but rather whether we exercise the will to evaluate as many teachers as possible to the best of our ability and then use that information to differentiate high school teachers in ways that allow for meaningful personnel and instructional decisions.

Another form of teacher evaluation gaining acceptance is based on student performance measures. Yet, like with practice-based evaluations, there are pluses and minuses.

Using student performance measures to evaluate teachers

Teacher evaluation using student performance measures is a rapidly emerging area that is gaining a foothold in states and districts across the nation. Several factors have come together to make this possible. The testing requirements of NCLB that have emerged in the last decade have provided the necessary measures of student performance; the rapid spread, increased quality, and declining costs of computerization and electronic data warehousing have eased the burden on data storage and manipulation. In addition, advancements in statistical modeling techniques have provided the analytical tools necessary for using student test scores to evaluate teachers. All of these factors have come together to make it possible to evaluate a portion of the teachers in most districts based on their ability to promote student performance on standardized tests, so-called “value-added” measures of teacher effectiveness.

These quasi-experimental statistical models yield estimates of the contribution of teachers to student achievement, controlling for nonschool sources of student achievement growth. Another way to characterize value-added measures is that they are the difference between the actual achievement of a teacher’s group of students and the predicted achievement of these students given their prior test scores, demographic characteristics, and other measures in the model. Simply put, the objective of a value-added growth model for teacher evaluation is to eliminate factors contributing to both student achievement levels *and student growth* over which a teacher who is being evaluated has no control.

From the district standpoint the advantages of using value-added to measure student growth and evaluate teachers begins with the fact that this method is much cheaper than practice-based teacher evaluation. Given that districts already have the data in hand for calculating teacher value-added measures, the primary costs to develop value-added systems are likely to be related to technical assistance consultation that may be needed to help with data system construction and model development. These costs should fall as value-added models and the data structures required for them become more standardized and established over time.

A second advantage of value-added measures of teacher effectiveness is that they evaluate teachers on observed output—student learning. One could argue that the tests on which value-added measures are based do not adequately measure the types of student learning that we want to promote in schools. This, however, is a criticism of the tests currently in use rather than a criticism of the value-added methods that rely on these tests.²⁸

It is also the case that value-added measures are more objective than performance-based measures that ultimately rely on human judgment and that these statistical measures of teacher effectiveness generate differentiation among teachers by construction. Given that districts do not currently use performance-based teacher evaluation in ways that produce meaningful differentiation among teachers, the variation inherent in value-added scores is potentially a significant benefit.²⁹

Not surprisingly, there are cautions that accompany the use of value-added as a method for evaluating teachers, beginning with the fact that value-added measures provide no information that could help a teacher become more effective. With value-added estimates, a teacher only knows where he or she ranks in the value-added distribution relative to other teachers. There is no additional information in a teacher's value-added score that would inform the teacher as to why he or she is ranked low or high relative to others.

There is also concern that if value-added-based evaluations are used for high-stakes decisions, then teachers will have the incentive to “teach to the test” or cheat. In this case “teaching to the test” does not refer to explicitly teaching material that a given test covers, something we might want teachers to do for well-constructed tests. Rather it refers to spending valuable class time on test-taking techniques or focusing on responses to specific, expected questions—it is teaching that does not promote real and lasting learning gains. When it comes to cheating, there is evidence that when test results are tied to high-stakes decisions, some teachers will resort to cheating as a way of artificially inflating test scores.³⁰

Another concern is that, at their best, value-added measures only capture a teacher's contribution to his or her students' learning as measured by standardized test scores. The concern here is that these tests only measure a subset of what we want students to learn and by focusing teacher evaluation on this sphere, we risk distorting teaching and learning that encompasses the breadth of what we want students to learn. One counterargument to this is that standardized tests cover

material that had been developed by relevant stakeholders that is deemed to be important and that these tests tend to be predictive of later life outcomes that we care about.³¹ Another counterargument is that for effective teachers, preparing students to do well on these tests is a byproduct of their normal teaching across the breadth of the curriculum.

Two related issues—measurement error and intertemporal stability—also pose some cause for concern when it comes to value-added estimates. Performance on any assessment will vary from one administration to the next for random reasons—a natural fluctuation known as measurement error. For value-added measures this means that any given teacher’s ranking in a distribution of teacher value-added estimates is an approximation that contains a degree of uncertainty. What this means practically is that identification of teachers in the upper or lower reaches of the value-added distribution can be done with reasonable confidence. The value-added models, however, tend to be less useful for differentiating teachers who are more proximate to each other in the distribution.

Measurement error also comes into play if instead of considering the value-added measures of two different teachers in the same year, one were interested in the value-added measures of the same teacher in different years. The concern here is about the degree of intertemporal stability of the value-added measures: the extent to which value-added scores for any given teacher are stable from one year to the next. While teachers’ effectiveness may change somewhat from year to year, one would not expect radical year-to-year changes in true, underlying ability. Thus, if value-added scores are good estimates of a teacher’s ability, then they should be relatively stable across time.³² Statistically speaking, this means we would expect a high year-to-year correlation in value-added scores. If one year’s value-added score perfectly predicted the next year’s score, then the year-to-year correlation would be one. And if there is no year-to-year relationship, then the correlation would be zero. Results to this point across numerous studies suggest that year-to-year value-added correlations are usually 0.2–0.5 for elementary school teachers, suggesting that a teacher’s value-added score in one year is not necessarily a good prediction of his or her value-added score in the subsequent year—a troubling proposition to some. It is the case, however, that with more information this problem is substantially abated. Using data from two prior years, for example, instead of one, substantially improves the ability to predict future performance.³³

Regarding the statistical properties of value-added estimates, it is worth reiterating two points made in a study by Dan Goldhaber.³⁴ The first is that there are errors in

measurement in any endeavor to capture complex human behavior. This includes using practice-based measures of teaching effectiveness, even though measurement error associated with, say, classroom observation scores is rarely considered or discussed. Second, value-added estimates are on par with performance measures in other fields in terms of their predictive validity. This includes the most statistical of all professional sports, Major League Baseball, where statistical measures of output are regularly used to reward players and construct employment contracts, even though the year-to-year correlation of many output measures are similar to what we see for value-added scores.³⁵

The purpose of controlling for prior student test scores and other student- and class-related factors in value-added measures is to try to isolate the effect of the current teacher on student achievement gains. Jesse Rothstein has conducted research that has brought into question how well typical value-added models actually accomplish this, bringing attention to potential biases in value-added estimates. These biases arise when students are sorted into classrooms on the basis of “dynamic” factors such as student home issues that are time varying, unobservable to the researcher, and are related to student outcomes.³⁶

Similar to the just-discussed instability issues, a response to the concern lies in having additional years of data to correct for this type of potential bias. Using rich data from the San Diego school district, Cory Koedel and Julian Betts were able to replicate the biases revealed in Rothstein’s work. Additionally, they show that a sufficiently complex value-added model that evaluates teachers over multiple years can reduce the sorting bias to statistical insignificance.³⁷ One implication of this line of work is that while one year of value-added data might be used to make low-stakes decisions such as deciding on professional development allocation, a district would want to rely on multiple years of data for making high-stakes decisions such as teacher tenure or job termination.

A last concern with value-added measures of teacher effectiveness is that only 15 percent to 35 percent of the teachers in any given district teach in grades and subjects where students have both a standardized end-of-year test and a suitable pretest.³⁸ Primarily because of the state standardized tests arising from the NCLB testing regime, these “tested” grades and subjects tend to be math and reading in grades 3–8. This is obviously a major challenge for using value-added to evaluate high school teachers but some of the top value-added researchers in the country are working on this problem as will be discussed shortly.

Evaluating high school teachers based on student performance

Inherent challenges

Basing a substantial portion of a teacher's evaluation on student performance measures was not only required to win Race to the Top funds; it is something that many state departments of education and school districts across the nation are exploring even without supporting federal funds. After all, teachers teach in order to impact student academic performance, and many argue that it makes sense to evaluate teachers on how well they accomplish that goal. While value-added measures are currently the most researched and talked about method for doing this, there are other methods under consideration that will be discussed in this section. First, let's look at the special issues that high school teachers pose for value-added models.

As a roadmap to this discussion, the following issues will be taken in turn:

- The availability of suitable tests at the high school level
- The path dependency of students in a given class
- The potential spillover effects of teachers in different subjects
- How to weight the different subjects a teacher might teach in an overall value-added score
- The potential for perverse incentives unique to high school
- The logistics of student attribution
- Having small numbers of teachers in the value-added distribution

Availability of suitable tests

Unlike grades 3–8, there are not regular state tests across grades and core subjects mandated by NCLB at the secondary level.³⁹ As a result, most states have not developed state tests for all high school grades, posing a central challenge at the high school level for using standard value-added models that require both a pre- and post-test.⁴⁰ Additionally, unlike the elementary grades where we can think of math and reading as being “core subjects” in the elementary and even middle school grades, the notion of core subjects has much less traction in high school where students begin to branch out and explore different topics and content.

Some states and districts can partially address the outcome-test problem via “end-of-course,” or EOC, exams or semester exams that have been developed to assess student mastery of course-specific knowledge. These exams are tied to state or district curricula in a specific course, are administered as students complete a course, and usually have high stakes for the student such as whether the course is passed for credit. Currently, there are 23 states that have EOCs for at least some courses, but even in these states, many high school teachers still teach in grades or subjects that are not covered by EOCs.⁴¹ Georgia is an example of a state that has developed statewide EOCs that all students in given courses have to take and pass. There are, however, only eight content areas across the four core subject areas (math, ELA, science, and social studies) that have EOCs in Georgia. Thus, while students taking biology, for example, would have an EOC in Georgia, students taking physics would not.

The one district in the nation that is working on developing a comprehensive arsenal of course-specific semester examinations is Hillsborough County Public Schools, Tampa, Florida. Hillsborough County Public Schools has developed over 500 semester examinations for courses ranging from 9th-grade English to sculpture to trigonometry to welding. Hillsborough also worked with the Florida Department of Education to make their semester examinations available to other districts across the state via the Florida End of Course Exam Clearinghouse. The scope of the Hillsborough effort, over 500 exams, illustrates the magnitude of the effort required to develop course specific examinations that could cover every high school teacher.

Nevertheless, as districts and states develop semester or EOC exams and share them through state sponsored exam clearinghouses or other venues, an increasing number of high school teachers will at least be teaching in courses that will have an associated post-test for use in value-added calculations.

In grappling with the high school test availability issue, researchers often conceptualize the value-added model less in terms of measuring student growth from one year to the next, as is the general conceptual model at the elementary level, and more in terms of exceeding (or falling short of) *predicted achievement*.⁴³ This conceptualization helps clarify the expansion of value-added models to noncore-subject areas and grades, like high school, where growth may not be easily defined or measured. In this realm there may or may not be pre-tests available for use as direct predictors of the outcome and, as a result, researchers will often have to rely on what could be termed related predictors of the outcome.⁴⁴

Taking math as an example, as a result of NCLB mandates, students in grades 4–8 will have both an outcome measure (for example, fourth-grade state math exam) and a direct predictor in math (the previous year’s third-grade state math exam).⁴⁵ Contrast this with the high school situation where even when there is an available EOC exam that can be used as an outcome measure (an EOC exam in geometry), there may seldom be a suitable prior test that could be used as a direct predictor. As a result, value-added models at the high school level will often have to rely on related predictors of the outcome measure.

At this point, the value-added research community is in the early stages of developing and testing models that attempt to address the lack of direct predictors (pre-tests) in high school. In this work, some models use scores from contemporaneous tests in other subjects as related predictors.⁴⁶ That is, if the outcome is an EOC for 10th-grade geometry, then scores from 10th-grade English and science exams, if available, might be used in the predictive model. Other models under consideration and testing include not only contemporaneous, other-subject tests but also an EOC or summative exam from the previous year in the same content area. Still other models being studied would incorporate the content area eighth-grade state exam as a secondary predictor. At this point researchers are engaged in studies to develop the best possible predictive value-added models for use in evaluating high school teachers. Success on this front may only start the decision process for districts as it is unlikely that there will be black and white answers as to what is the “right” value-added model a district should use to evaluate high school teachers.

Since different models will embody different assumptions and rely on different variables to control for outside-of-class influences on student performance, districts will likely have to make policy decisions regarding which models they will use for high school evaluation. Nevertheless, the work on this front is impressive with great strides having been made with many districts and states currently in the process of considering how, not whether, to use value-added as one factor in evaluating their high school teachers.

Path dependency

In elementary school students generally follow the same path through the grades with little deviation in terms of the subjects they study and in what year they study them. Thus, elementary-level value-added estimates rest on a fairly solid assumption that, in general, students arrive at the outcome measure via the same

academic pathway. Students taking the fifth-grade state math exam, for example, get there by passing first through fourth grades and taking the math curricula in each of these grades. Where there are deviations, say, for being held back a year or being in a gifted and talented program, these deviations can be accounted for in the model as long as this information is in the administrative data.

On the other hand, a common path to a given course is rarely the case in high school given that students in the same cohort can take different courses each year as well as different sequencing of the same courses as they progress through high school. If the courses taken prior to or contemporaneous with some outcome measure have the potential to impact that measure, and if the impact is different depending on when the earlier courses were taken relative to the outcome of interest, then it is important to account for such path dependency in high school value-added models. Given that there are many possible “paths” to, say, the EOC exam in 11th-grade physics, accounting for path dependency is not a simple matter.

One way to study this potential problem is to use existing data to study how sensitive value-added measures are to path dependency. Existing data in a given district, for example, could be used to determine the most common two, three, or four paths students take to, say, 11th-grade physics. Value-added estimates that then controlled and didn’t control for these paths could be compared. If it turns out that in most cases path dependency is not a strong predictor of the outcome, given other predictors in the model, then there is less concern with trying to account for path dependency. If this is not the case, then the task will lie in developing tractable methods that control for the different paths students take in arriving at the outcome of interest.

Teacher spillover effects

Again unlike elementary school, where students tend to be taught all of the subjects by one teacher, high school students have up to five or six different teachers every semester. If one’s performance is influenced by all of their teachers, regardless of the outcome measure, then value-added estimates for the teacher of record that did not account for the spillover effects of other teachers could be biased. If most of a given teacher’s students, for example, tended to have very high-performing teachers in their other subjects and there were spillover effects that affected the students’ scores in the outcome subject, then failure to account for this would lead to upward bias in the value-added estimates of that teacher: He or she would

receive credit for the systematically good teachers that his or her students had in their other subjects. This too is a potential thorny issue because of all of the potential combinations of “other teachers.” Again, a potential way to study this issue is to use existing data and model some of the most common combinations of other teachers observed in the data for given outcomes of interest and determine how important it is to account for students’ other teachers.⁴⁷

Teacher weighting across different courses

The issue here is that some, if not many, high school teachers will teach more than one course. The typical high school math teacher, for example, might teach an Algebra I course along with some Algebra II courses and some geometry courses in the same year and/or semester. Or, given the breadth of courses offered in the typical comprehensive high school, a single teacher might be responsible for teaching, say, history, psychology, and sociology. In developing a measure of how effective a teacher is, how should the value-added-related performance in each of the various subjects be weighted? This is a policy decision rather than a modeling decision that districts will have to make, and there is no clear answer as to what is the “right” decision. Since the decision will most likely affect teachers’ final evaluation score, however, it is a serious issue in evaluation system design that districts will have to consider.

Perverse incentives

There are certain perverse incentives that can arise at any grade level when teacher evaluation is based on student test scores. High school teachers whose evaluations are dependent upon their students’ test scores are certainly not immune from the temptation to “teach to the test” or to cheat in ways that could artificially inflate test scores. In addition to these grade-neutral perverse incentives, however, teacher evaluation based on value-added models at the high school level must guard against perverse incentives specific to high school. In particular, we would want to guard against practices that incentivized teachers to encourage their lowest-performing students to either drop out of school or drop the course before the end-of-year exam. High school evaluation systems also need to assure that all students, including those repeating a course, which is most common in the ninth grade, are included in the testing regime so that teachers will have the incentive to teach all students to the best of their ability.

Logistics of attribution

Assigning the correct students to the correct teachers is an important issue for all value-added models, and the evidence is that correctly capturing a teacher's roster of students based on administrative data is a far-from-trivial task. In a study by the nonprofit, school-improvement organization Battelle for Kids that covered 57 districts and 730 schools in two states, 26 percent of all student assignments were moved, added, or deleted from the district-reported data through Battelle's more carefully executed teacher-linking process.⁴⁸ Correct student attribution is an even more difficult process at the high school level than at the elementary level if for no other reason than each teacher has multiple classes of students, each of which require accurate attribution. Roster verification, especially in high school, is a major challenge that districts contemplating the use of value-added measures have to take seriously.

Given both the challenge and the importance of correctly linking students and teachers, several organizations across the nation are devoting resources and time to help states and districts develop data systems and infrastructure that will allow for the kind of linkage required by high-quality value-added systems. Among these organizations are Battelle for Kids, the Data Quality Campaign, and the Teacher-Student Data Link Project of the Center for Educational Leadership and Technology.⁴⁹

Teacher group size

All value-added estimates place teachers in a distribution of other teachers. In the case of value-added at the elementary level, the teacher is usually compared to other teachers in the district in the same grade and year. Thus a fifth-grade teacher in a midsized school district might be in a distribution with 75 to 100 other fifth-grade teachers. A potential issue with high school value-added is that in addition to the grade by year comparison group, teachers need to be compared to other teachers teaching the same subject. This can be potentially problematic for mid- to small-size districts as there may not be enough, say, physics teachers in the district for the comparison to be meaningful. In that type of scenario, what does it mean to be at the bottom (or top) of the value-added distribution of physics teachers when there are four physics teachers total in the district? Value-added with small numbers of teachers in the comparison group probably makes little sense and in these cases other measures of teacher effectiveness will have to be used or the district in conjunction with the state department of education would need to determine whether it makes sense to compare teachers statewide rather than districtwide.

Potential opportunities

There are also potential opportunities for using value-added measures in high school teacher evaluation not present in the elementary school setting. Since the end-of-year state exams used at the elementary level are taken by all students in a state, they are largely designed to be “curriculum free.” In contrast, the EOC tests and other course-related tests being considered for high school value-added are designed to be in alignment with the curriculum being taught. As a result, teachers have more power to impact student performance on tests aligned with the taught curriculum than they do on performance on curriculum-free state exams. It is also likely that incentives are better aligned, which in turn yields greater teacher buy-in when the test over which a teacher will be held accountable is sensitive to the taught curriculum.

A second potential advantage for high school value-added is that in most cases there will be substantially more students on a high school teacher’s roster than an elementary teacher’s since high school teachers as a rule teach several classes. These additional students help reduce the measurement error discussed earlier, meaning that a teacher’s value-added score is more precisely estimated.

A final opportunity worth noting is that the series of challenges a district faces in trying to develop good value-added measures for their high school teachers is leading districts to contemplate the entire evaluation, testing, and data enterprise of the district in a more coherent way than has generally been the case to this point. Districts on the leading edge of this work understand that using value-added to evaluate high school teachers requires thinking about testing and data issues in a coherent and systematic way. When done well this holistic approach can have positive spillover effects that can benefit more than just teacher evaluation in the district.

Other student performance measures

As previously noted, many states and districts are developing student performance measures that are not based on value-added estimates. In large part these efforts are in response to some of the difficulties discussed above in constructing value-added measures for teachers in nontested grades and subjects. One result of this work is the development of Student Learning Objectives to measure teacher effectiveness. SLOs are data-based targets of student growth that teachers set for

the students at the start of the semester or school year. Teachers are then evaluated at the end of the semester or school year on the extent to which their students met the objectives that they set for their students. SLOs are currently in use in Denver, Colorado; Austin, Texas; and Charlotte-Mecklenburg, North Carolina. Many new systems, including the state system in Rhode Island, are incorporating SLOs into new evaluation systems.⁵⁰ In some instances SLOs are used only in nontested grades and subjects, while in other cases a district is incorporating SLOs into the evaluation of teachers in addition to, instead of in place of, value-added measures.

The typical steps a teacher goes through in setting SLOs are:⁵¹

- At the beginning of the semester or year, review available data on the students in the class, including prior-year test performance and any course pre-tests that have been administered.
- Based on the data, set a designated number of objectives, usually two—classroomwide and student- or subgroup-specific. An example of a class-level objective might be: “Increase the Algebra I end-of-course pass rate by 5 percentage points over last year’s 85 percent pass rate.”
- Identify appropriate measures against which objective attainment will be judged. Following the above example, the measure for judging objective attainment would be the class pass rate on the Algebra I end-of-course exam.

In many instances teachers must review and discuss their SLOs with their principal (and sometimes central office staff) as a part of the SLO-setting process. The principal (and central office staff when appropriate) must then approve and sign off on each teacher’s SLOs. While there is little research to date on SLOs, principal and/or central office approval is likely a key part of the SLO process. The incentive with no oversight is for teachers to set easily achievable SLOs that their students can readily meet.

Most of what we know about SLOs at this point comes from Denver and Austin. A 2004 pay-for-performance pilot program based on 17 Denver schools, including two middle schools and two high schools, found that 89 percent to 93 percent of the teachers in the pilot met their objectives over the four years of the pilot (1999–2003). Further, the quality of the SLOs set by teachers went up over the four years.⁵² In studying the relationship between the quality of the SLOs set by middle school and high school teachers and their students’ performance, there

was consistent evidence of a positive relationship, though this evidence was often not statistically significant. Positive and statistically significant correlations between the number of objectives met by teachers and the achievement of their students were found in the two high schools.⁵³

In Austin, SLOs are part of the REACH strategic compensation program.⁵⁴ There, SLOs are based on the state curriculum—the Texas Essential Knowledge and Skills. In 2009-10 teachers in REACH schools received stipends of between \$1,000 and \$1,500 for each met SLO. The SLO process in Austin begins with teachers examining their students' performance data and identifying two areas of greatest needs. Pre-assessments are then administered to the students in the selected area of needs. Based on the results of the pre-tests, teachers then set SLOs that must be reviewed and approved by the campus principal and district central office. Each SLO must indicate performance targets that students will meet by the end of the school year and how performance will be assessed. Student results from the end-of-year post-assessment are then used to determine whether or not a teacher met his or her SLOs. In an analysis based on 2009-10 state test data, students of REACH high school math and science teachers who met at least one of their SLOs demonstrated greater net achievement growth than did the students of REACH teachers who did not meet SLOs.⁵⁵

The nonexperimental nature of both the Denver and the Austin studies relating SLOs to student achievement leave some doubt regarding that relationship and indicate the need for additional studies on the topic. Future studies should also look at the relationship between SLOs and student achievement gains, not just levels. Another issue with which SLO-based evaluation will need to grapple in the future is comparability across classrooms when SLOs are based on the individual choices of teachers. Finally, it is not clear at this point how effective the SLO evaluation process can be at differentiating among teachers who are differentially effective. The early evidence from the Denver pilot project looks disturbingly similar to the results from the “The Widget Effect.”

Another way in which several districts and at least one state (Delaware) are developing nonvalue-added measures of student achievement growth is through “common growth measures.” To develop a set of common growth measures, a committee of teachers and other educators at either the state or district level reviews available measures of student growth and then makes recommendations to the district or state, which then approves a list of measures for each subject and grade under consideration. In Delaware more than 300 educators have

been enlisted to help develop student growth measures in 30 content areas.⁵⁶ In addition to Delaware five upstate school districts in New York are developing common growth measures to use in evaluation systems. This is a promising and much-needed development in that the common growth measure process is even less studied than SLOs. At this juncture it is too early to know the extent to which either can serve as useful tools in teacher evaluation systems.

Conclusion

Developing effective ways to evaluate high school teachers holds the promise of improving the high school education experience for the millions of students who will be taught by these teachers. Based on what we know from research and experience, the following ideas should guide evaluation system development and implementation and the ongoing work in this area in the coming years.

The best evaluation systems will incorporate all available information from both practice-based information and student performance data into the ultimate evaluation of teachers

This means that both value-added measures and SLOs should be used when possible and that value-added measures should be used with as many high school teachers as possible. The tough decisions should be around how to weight the different components, including value-added, not whether or not to use them at all in teacher evaluation. Based on the research evidence, systems that do not use all of these measures are leaving information about the effectiveness of their teachers on the table.

At the same time, just because it may be hard to develop value-added measures for all high school teachers in a district, districts should not use this as an excuse to forego value-added evaluation for the subset of teachers where the data are available. Just as the standard pay scale across all subject areas that is common in education is shortsighted and reduces the supply of qualified math and science teachers who could be recruited into schools, refusing to use value-added information with some teachers because it may not be available for all is equally unwise.

A good evaluation system will not shortchange its practice-based evaluation component

There are very few systems that currently make the investments necessary to conduct high-quality practice-based evaluation. Well-designed systems use dedi-

cated and trained evaluators, master teachers who leave the classroom for two to three years and whose sole responsibility is evaluation of the district's teachers. Teachers are observed in the classroom multiple times during the year, the classroom observations are of sufficient length to gather meaningful data about that teaching episode, and the visits are mostly unannounced. Well-designed systems use research-based evaluation protocols and the training of evaluators on these protocols is taken seriously. While such systems can be costly, districts can offset some of the costs by only putting a portion of the teachers on the full, comprehensive evaluation cycle each year. It is arguably better practice to evaluate a portion of the teachers at a level of high quality each year than it is to evaluate all of the teachers with low-level evaluation every year. Furthermore, research suggests that high-quality evaluation can pay for itself by increasing teacher productivity, and that given this districts could further offset evaluation costs by redirecting some professional development dollars to evaluation.⁵⁷

It is likely that many districts that rely on principals for their teacher evaluation activities do so because they see the cost savings of this approach relative to having a cadre of master teachers who serve as evaluators. It is also likely that few of these districts take into account the opportunity costs of shouldering principals with this extra duty when considering the costs and benefits of principal-led evaluation relative to alternative approaches to evaluation. Given the rising importance of high-quality teacher evaluation going forward, districts should henceforth consider the full costs and benefits, not just the accounting costs, of using principals to carry the bulk of the evaluation load. When considering the full costs and benefits of having principals versus full-time, trained evaluators, or some combination of principals and evaluators, carry out evaluation, principal-led evaluations may not be as attractive as districts currently perceive them to be.

We need to continue the work that is being done to develop and test value-added models appropriate for use with high school teachers

An important test for resulting models will be the extent to which high school teachers' value-added scores are correlated with their classroom observation scores. This linkage has been established in some few but important instances at the elementary and middle school levels.⁵⁸ The takeaway from those studies is that low- and high-value-added teachers are doing something different in the classroom, and the evaluators observing these teachers are seeing that difference and scoring it, even though the evaluators themselves have no knowledge of where

a teacher might be in the value-added distribution. Finding this same kind of relationship between value-added and classroom observation scores at the high school level would provide an additional level of confidence in what are likely to be relatively complex high school value-added models.⁵⁹

More research focused on SLOs is needed

Given the increasing use of this evaluation method, we need more information on the extent to which SLO scores are related to a teacher’s ability to promote student learning gains. Absent this evidence we have to trust that SLOs are measuring effective teaching. We need research that can determine whether the SLO-setting process makes teachers better and raises the general level of teaching on a campus and in a district as some advocates claim. Given that SLOs may be used more at the high school level than at the elementary level, this research evidence is particularly important for high school evaluation.

More work needs to be done on the possibility of developing assessments that can gauge a teacher’s ability to teach his or her content

We first need to know more about the extent to which these content-specific pedagogical assessments are related to a teacher’s ability to promote student learning in the content. If we find there is promise on that critical dimension, then that would be cause for a push to develop more of these assessments across additional high school content areas.

Teacher evaluation should take the lead in finding a way to use technology for efficiency and productivity gains

Education has a dismal record in effectively using technology. If teacher evaluation is to have the impact envisioned by many, we may not be able to afford to follow this pattern. Especially in the realm of practice-based evaluation, technology holds the promise of allowing us to do more, do it better, and with less cost, particularly at the high school level. The promise of digital video technology is that all teachers in a high school could be “observed” several times during the year via video, these teaching episodes could be viewed by content specialists, and the total costs will

be less than the costs of training, providing administrative support, transportation, and salary for district-based, full-time evaluators. In addition, as the district builds a video collection of exemplary teaching sessions, it will develop an in-house bank of examples that novice teachers and others can view as professional development exercises.

The intense work being done across the nation is rooted in the belief that if we can do a better job at teacher evaluation, the ultimate result will be better outcomes for our kids and by extension our nation. A large part of answering that promise rests with how well new evaluation systems perform at the high school level. An impressive array of high-performing teachers and administrators, evaluation experts, economists and statisticians, technology leaders, and state and federal policymakers are currently working on teacher evaluation issues, with much of this work focusing on the high school question. The urgency with which this work is being done and the timing of much of the grant money tied to new evaluation initiatives suggests that in the next few years, new evaluation systems will be in place. It is fair to say that just as putting teacher evaluation front burner in a district can change the way teachers talk about their craft, the national effort currently in place is changing the way we talk about teaching and how we can and should evaluate teachers across the nation, including those in our high school classrooms.

About the author

John H. Tyler is associate professor of education, economics, and public policy at Brown University. He is also a research associate at the National Bureau of Economic Research. Tyler has conducted research on the linkages between school, skills, and labor market outcomes, particularly for low-educated, low-skilled individuals. His most recent work has examined questions pertaining to teacher evaluation, and in particular he has examined the relationship between teacher evaluation measures and teachers' measured ability to impact student achievement gains.

Acknowledgements

The author would like to thank the many people who took time to discuss or correspond on issues in this paper, especially Andy Baxter, Connie Casson, Charlotte Danielson, Brian Gill, Dan Goldhaber, Bridget Hamre, Julia Indelaccio, Rob Meyer, Christine Sargent, Maurine and Dave Sigler, Mary Ann Snider, Eric Taylor, and Tom Tomberlin.

The Center for American Progress thanks the Bill & Melinda Gates Foundation for generously providing support for this paper.

Endnotes

- 1 National Council on Teacher Quality, "State of the States" (2011).
- 2 For a summary of this research, see: E.A. Hanushek and S.G. Rivkin, "Generalizations about Using Value-Added Measures of Teacher Quality," *American Economic Review* 100 (2) (2010): 267–271.
- 3 Note that this is not a normative statement about the relative quality of high school versus elementary teachers in the nation but rather a statement about the relative performance of students across the two levels. It could be that the high school teacher workforce is on average equally or more effective than elementary teachers and we could still see worse student outcomes at the high school level as a result of factors outside of the control of teachers that are particular to older students. The best evidence on the disparities in outcomes between elementary and high school comes from the results on the National Assessment of Educational Progress, or NAEP, the so-called "nation's report card." In 2009 NAEP data, for example, 39 percent and 34 percent of fourth- and eighth-graders, respectively, were at or above the Proficient level in math while only 26 percent of the nation's 12th-graders scored in this range. See: National Center for Education Statistics, "The Nation's Report Card: Mathematics 2009" (2009).
- 4 For information on the dropout problem in general and data on the time trend in dropout rates, see: J.H. Tyler and M. Lofstrom, "Finishing High School: Alternative Pathways and Dropout Recovery," *America's High Schools* 19 (1) (2009): 77–103. For a discussion of the dropout problem and the role of student engagement in contributing to the dropout decision, see: J.M. Bridgeland, J.J. Dilulio, Jr., and K.B. Morrison, "The Silent Epidemic: Perspectives of High School Dropouts" (Washington: Civic Enterprises, 2006). For an example of research linking teacher behavior to student engagement, see: E.A. Skinner and M.J. Belmont, "Motivation in the Classroom: Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year," *Journal of Educational Psychology* 85 (4) (1993): 571–581.
- 5 Information on the prevalence of remediation in college is found in: National Center for Education Statistics, "Remedial Education at Degree-Granting Postsecondary Institutions in Fall 2000" (2003). For research indicating that remediation itself is effective in helping students succeed in college, see: E.P. Bettinger and B.T. Long, "Addressing the Needs of Underprepared Students in Higher Education: Does College Remediation Work?," *Journal of Human Resources* 44 (3) (2009): 736–771.
- 6 For a discussion of the skills required in the new economy, see: F. Levy and R.J. Murnane, *The New Division of Labor: How Computers are Creating the Next Job Market* (Princeton, NJ: Princeton University Press, 2004).
- 7 As evidence of union support at the national level, in a January 12, 2010, speech, Randi Weingarten, president of the American Federation of Teachers, or AFT, stated that the AFT wants "a fair, transparent, and expedient process to identify and deal with ineffective teachers." See: N. Anderson, "Union head to propose tying test scores, teacher evaluations," *The Washington Post*, January 12, 2010. Further evidence comes from the June 2011 Representative Assembly of the National Education Association, the nation's largest teachers' union, where delegates voted by a wide margin to adopt a policy statement on teacher evaluation and accountability, a move that put the NEA on record for the first time as calling for comprehensive overhaul of both teacher evaluation and accountability systems as vehicles for improving student learning.
- 8 Work the author has done in Cincinnati indicates that classroom observations make up half of a teacher's end-of-year summative score, while practices and artifacts relating to professionalism make up 25 percent and practices and artifacts relating to preparation and planning for teaching make up the remaining 25 percent. There is no research that provides this type of information on a national scale.
- 9 The three frameworks that have been correlated with student learning growth are: C. Danielson, *Enhancing Professional Practice: A Framework for Teaching* (Alexandria, VA: Association for Supervision and Curriculum Development, 2007); R. Pianta, K.M. LaParo, and B.K. Hamre, *Classroom Assessment Scoring System* (Charlottesville, VA: University of Virginia, 2004); "PLATO 3.0," available at <http://cset.stanford.edu/tools/plato/>.
- 10 Examples of exceptions where student test score gains are currently used to evaluate at least a portion of the teachers in the district include Washington, D.C.; Dallas; and Hillsborough County (Tampa, Florida).
- 11 Most districts that use trained evaluators instead of or in addition to building principals use the Peer Assistance and Review, or PAR, model where master teachers are removed from the classroom for one to three years to serve as teacher evaluators in the district. Only a small number of districts use PAR-based evaluation, however. According to information retrieved from the National Comprehensive Center for Teacher Quality, or NCCTQ, website, only about 70 districts nationwide use a PAR system. "Guide to Teacher Evaluation Products," available at <http://resource.tqsource.org/GEP/GEPTool.aspx?gid=23>. The NCCTQ website indicates that another 200 schools (not districts) use the TAP evaluation tool that also uses master and mentor teachers in addition to administrators to conduct classroom observations. Thus, while there is no study that provides information on the distribution of types of evaluation systems in place across the nation, what information is available makes it clear that in most of the 15,000 districts in the United States, teacher evaluation is being carried out by building administrators.
- 12 M. Kennedy, "Approaches to Annual Performance Assessment." In M. Kennedy, ed., *Teacher Assessment and the Quest for Teacher Quality* (San Francisco, CA: John Wiley & Son, Inc., 2010).
- 13 P.M. Tucker, "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 (2) (1997): 103–126.
- 14 The tendency of managers to overrate their employees when giving employee appraisal ratings is not unique to education. For a discussion of leniency bias in the private sector, see: C. Prendergast, "The Provision of Incentives in Firms," *American Economic Review* 37 (1) (1999): 7–63.
- 15 D. Weisberg and others, "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness," *Education Digest: Essential Readings Condensed for Quick Review* 75 (2) (2009): 31–35.
- 16 E.S. Taylor and J.H. Tyler, "The Effect of Evaluation on Teacher Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers." Working Paper w16877 (National Bureau of Economic Research, 2011).
- 17 As a part of a comprehensive evaluation in the Cincinnati system, teachers will be observed in the classroom at least four times during the year. All teachers who do not undergo comprehensive evaluation undergo an "annual" evaluation which constitutes one classroom visit by the principal. All new hires and all teachers coming up for tenure undergo comprehensive evaluation as do tenured teachers every five years.
- 18 It is noted here that Cincinnati was one of the 12 districts studied in the aforementioned "The Widget Effect," and like the other districts in that study, evaluation scores in Cincinnati were highly skewed to the positive end.

- 19 Taylor and Tyler, "The Effect of Evaluation on Teacher Performance."
- 20 This finding takes on added importance given the lack of rigorous evidence that professional development activities actually make teachers more effective. For a summary of the research on the impact of professional development for teachers, see: K. S. Yoon and others, "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement" (Washington: National Center for Education Evaluation and Regional Assistance, 2007), available at <http://www.eric.ed.gov/PDFS/ED498548.pdf>.
- 21 Numbers based on the author's examination of several websites of comprehensive high schools. Also, as is discussed later in the paper, the Hillsborough County School District has identified more than 500 courses that are taught in their school district.
- 22 Note that the Protocol for Language Arts Teaching Observation, or PLATO, is designed to evaluate language arts teachers in middle school and the ninth grade. The extent to which this content-specific tool can be used by evaluators who are not themselves content specialists, however, is unclear.
- 23 Information on the LMT project can be found at: "Learning Mathematics for Teaching (LMT) Project," available at <http://sitemaker.umich.edu/lmt/home>. Information on the MET Project's work on content knowledge assessments is at: MET Project, "Content Knowledge for Teaching and the MET Project," September 2010, available at http://www.metproject.org/downloads/Teacher_Knowledge_092110.pdf. Information on ATLAST is at: "Projects -- Horizon Research, Inc.," available at http://www.horizon-research.com/projects/current/current_project.php?project_id=3.
- 24 Information on this part of the MET project is at: MET Project, "Classroom Observations and the MET Project," September 2010, available at http://www.metproject.org/downloads/Classroom_Observation_092110.pdf. Information on Teachscape, the technology partner in the project, is at: "Teachscape - About - Business Partners," available at <http://www.teachscape.com/html/ts/nps/partners.html>.
- 25 These figures are from: I. Quillen, "Room with a View," *Education Week*, June 15, 2011.
- 26 Of course a district could have multiple evaluators visit the classroom for in-person classroom observations but this becomes more disruptive to the classroom environment and would increase costs.
- 27 For a discussion of some of the issues around videotaping teachers, see: M. Petrilli, "Lights, Camera, Action! Using video recordings to evaluate teachers," *Education Next* 11 (2) (2011): 85–87.
- 28 One way the federal Department of Education is attempting to address the test quality concern is via the Race to the Top Assessment competition where the two winning state consortia, representing 44 states and the District of Columbia, will share \$330 million to develop new student assessments in math and English language arts. The premise of the competition was to spur states to join together in creating assessments that were tied to the recently unveiled common core standards and could measure critical thinking skills and complex student learning.
- 29 Only a portion of the observed variation in value-added scores is due to true underlying differences in teacher effectiveness, with some of the observed variance being the result of random measurement error. Thus, care must be taken in interpreting the differentiation among teachers that comes from value-added models. Given that the measurement error can be estimated in most models, any given teacher's value-added estimate can be bracketed with standard confidence intervals.
- 30 For a discussion of cheating scandals across the years and in various districts, including Atlanta, see: C.A. Samuels, "Cheating Scandals Intensify Focus on Test Pressures," *Education Week*, August 4, 2011, available at http://www.edweek.org/ew/articles/2011/08/04/37cheating_ep.h30.html?r=468391506. For research into cheating in Chicago, see: B.A. Jacob and S.D. Levitt, "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118 (3) (2003): 843–877.
- 31 Evidence on the relationship between standardized test scores and labor market outcomes can be found in 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; E.A. Hanushek and others, "Education and Economic Growth: It's Not Just Going to School but Learning that Matters," *Education Next* 8 (2) (2008): 62–70.
- 32 The key assumption here is that a teacher's ability does not change radically from one year to the next. Given this assumption one would expect to see relatively stable value-added scores across the years even in the face of gradual improvement due to, say, experience.
- 33 D.F. McCaffrey and others, "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4 (4) (2009): 572–606.
- 34 D. Goldhaber, "When the Stakes Are High, Can We Rely on Value-Added?: Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions" (Washington: Center for American Progress, 2010).
- 35 D.A. Hoffman, R. Jacobs, and S.J. Gerras, "Mapping Individual Performance Over Time," *Journal of Applied Psychology* 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.
- 36 J. Rothstein, "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4 (4) (2009): 538–572.
- 37 C. Koedel and J.R. Betts, "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique," *Education Finance and Policy* 6 (1) (2011): 18–42.
- 38 J.G. Watson, S.B. Kraemer, and C.A. Thorn, *The Other 69 Percent* (Washington: Center for Education Compensation Reform, 2009).
- 39 No Child Left Behind required that, as of the 2005-06 school year, each state must measure every child's progress in reading and math in each of grades 3–8 and at least once during grades 10–12. By school year 2007-08, states must also have had in place science assessments to be administered at least once during grades 3–5, grades 6–9; and grades 10–12.
- 40 H. May and others, "Using State Tests in Education Experiments: A Discussion of the Issues" (Washington: National Center for Educational Evaluation and Regional Assistance, 2009), Table A.1.
- 41 S. Dietz and others, "State High School Tests: Exit Exams and Other Assessments" (Washington: Center on Education Policy, 2010).
- 42 Information on the tests developed by Hillsborough and on the Florida End of Course Exam Clearinghouse can be found at the Clearinghouse website <http://is4.sdhc.k12.fl.us/ece/doclist.html>.
- 43 I especially thank Rob Meyer at the Value-Added Research Center for conversations on this topic. Information on Meyer and the Value-Added Research Center can be found at: "VARC," available at <http://varc.wceruw.org/>.
- 44 In the fully specified model, these predictors will usually be included in addition to measures that capture student background characteristics along with class-level and school-level variables that may also affect the outcome.
- 45 Note that in the typical district administrative data, some students will be missing one or both of these tests for various reasons including absence on the test date(s), being a new student in the district, data entry errors, etc.
- 46 For an example of this approach, see: C.T. Clotfelter, H.F. Ladd, and J.L. Vigdor, "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects," *Journal of Human Resources* 45 (3) (2010): 655–681.

- 47 Note that in cases where all students in a given course have the same set of “other teachers,” as might be the case for example for students in an honors track, it is not possible to separate out the effects of other, contemporaneous teachers from the value-added estimate of the teacher in question.
- 48 Battelle for Kids, “The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness” (2010), available at http://www.battelleforkids.org/Services/Teaching_Effectiveness/Linking_Teachers_and_Students.html?sflang=en.
- 49 Information on the work these organizations are doing in the area of student-teacher data linkage can be found, respectively, at: “Battelle for Kids,” available at <http://www.battelleforkids.org>; “Data Quality Campaign,” available at <http://www.dataqualitycampaign.org>; “Teacher-Student Data Link Project,” available at <http://www.tsdl.org>.
- 50 S. Sawchuck, “Wanted: Ways to Measure Most Teachers,” *Education Week*, February 2, 2011.
- 51 “CTAC - Student Learning Objectives (SLOs),” available at <http://www.ctacusa.com/slos.html>.
- 52 Community Training and Assistance Center, *Catalyst for Change: Pay for Performance in Denver, Final Report* (2004).
- 53 Student achievement was measured in math and reading using both the Iowa Test of Basic Skills, or ITBS, and the Colorado Student Assessment Program, or CSAP, math and English tests.
- 54 Information in this section taken from: L. Schmitt and N. Ibanez, “AISD REACH Program Update: 2009-2010 Texas Assessment of Knowledge and Skills (TAKS) Results and Student Learning Objectives (SLOs)” (Austin, TX: Austin Independent School District Department of Program Evaluation, 2011).
- 55 Ibid.
- 56 Summary first year report from Delaware’s winning Race to the Top application found at <http://www.vision2015delaware.org/wp-content/uploads/2011/06/Progress-Report-2011.pdf>.
- 57 Taylor and Tyler, “The Effect of Evaluation on Teacher Performance.”
- 58 For this work as it relates to the Danielson Framework, see: T.J. Kane and others, “Identifying Effective Classroom Practice Using Student Achievement Data,” *Journal of Human Resources* 46 (3) (2011): 587–613; A. Milanowski, “The Relationship Between Student Performance Scores and Student Achievement: Evidence from Cincinnati,” *Peabody Journal of Education* 79 (4) (2004): 33–53; E. Holtzapple, “Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System,” *Journal of Personnel Evaluation in Education* 17 (3) (2003): 207–219. For this work as it relates to CLASS-S, the CLASS tool for use in high schools, see: J.P. Allen and others, “Observations of Effective Teaching in Secondary School Classrooms: Predicting Student Achievement with the CLASS-S,” University of Virginia, 2011. For this work as it relates to the PLATO tool, see: P. Grossman and others, “Measure for Measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers’ value-added scores.” Working Paper w16015 (National Bureau of Economic Research, 2010).
- 59 I note here that the aforementioned “validation” studies at the elementary level were generally seen as a test of whether practice-based evaluation was truly capturing teaching that was related to student test score growth. That is, the assumption that real effectiveness was captured by a teacher’s ability to promote student test score growth and practice-based scores were measured against this standard. What I am proposing here is that for the more complex and as yet still underdeveloped high school value-added models, their validation against teachers’ practice-based scores could help in both deciding on model specification and generating confidence and buy-in around the eventual models.

References

- Allen, J.P., and others. 2011. "Observations of Effective Teaching in Secondary School Classrooms: Predicting Student Achievement with the CLASS-S." Unpublished manuscript, University of Virginia.
- Anderson, N. 2010. "Union head to propose tying test scores, teacher evaluations." *The Washington Post*. January 12.
- Battelle for Kids. 2011. "The Importance of Accurately Linking Instruction to Students to Determine Teacher Effectiveness" (http://www.battelleforkids.org/Services/Teaching_Effectiveness/Linking_Teachers_and_Students.html?sflang=en [July 1, 2011]).
- Bettinger, E.P. and B.T. Long. 2009. "Addressing the Needs of Underprepared Students in Higher Education: Does College Remediation Work?" *Journal of Human Resources* 44 (3): 736–771.
- Bridgeland, J.M., J.J. Dilulio, Jr., and K.B. Morrison. 2006. "The Silent Epidemic: Perspectives of High School Dropouts." Washington: Civic Enterprises.
- Clotfelter, C.T., H.F. Ladd, and J.L. Vigdor. 2010. "Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects." *Journal of Human Resources* 45 (3): 655–681.
- Community Training and Assistance Center. 2004. "Catalyst for Change: Pay for Performance in Denver, Final Report." Boston.
- Danielson, C. 2007. *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Dietz, S., and others. 2010. "State High School Tests: Exit Exams and Other Assessments." Washington: Center on Education Policy.
- Goldhaber, D. 2010. "When the Stakes Are High, Can We Rely on Value-Added: Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions." Washington: Center for American Progress.
- Grossman, P., and others. 2010. "Measure for Measure: The relationship between measures of instructional practice in middle school English Language Arts and teachers' value-added scores." Working Paper w16015. National Bureau of Economic Research.
- Hanushek, E.A., and others. 2008. "Education and Economic Growth: It's Not Just Going to School but Learning that Matters." *Education Next* 8 (2): 62–70.
- Hanushek, E.A., and S. G. Rivkin. 2010. "Generalizations about Using Value-Added Measures of Teacher Quality." *American Economic Review* 100 (2): 267–271.
- Hoffman, D.A., R. Jacobs, and J. E. Baratta. 1993. "Dynamic Criteria and the Measurement of Change." *Journal of Applied Psychology* 78(2): 194–204.
- Hoffman, D.A., R. Jacobs, and S.J. Gerras. 1992. "Mapping Individual Performance Over Time." *Journal of Applied Psychology* 77 (2): 185–195.
- Holtzapple, E. 2003. "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System." *Journal of Personnel Evaluation in Education* 17 (3): 207–219.
- Jacob, B.A., and S.D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics* 118 (3): 843–877.
- Kane, T.J., and others. 2011. "Identifying Effective Classroom Practice Using Student Achievement Data." *Journal of Human Resources* 46 (3): 587–613.
- Kennedy, M. 2010. "Approaches to Annual Performance Assessment." In M. Kennedy, ed., *Teacher Assessment and the Quest for Teacher Quality*. San Francisco: John Wiley & Son, Inc.
- Koedel, C., and J.R. Betts. 2011. "Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique." *Education Finance and Policy* 6 (1): 18–42.
- Levy, F., and R.J. Murnane. 2004. *The New Division of Labor: How Computers are Creating the Next Job Market*. Princeton, NJ: Princeton University Press.
- May, H., and others. 2009. "Using State Tests in Education Experiments: A Discussion of the Issues." Washington: National Center for Educational Evaluation and Regional Assistance.
- McCaffrey, D.F., and others. 2009. "The Intertemporal Variability of Teacher Effect Estimates." *Education Finance and Policy* 4 (4): 572–606.
- Milanowski, A. 2004. "The Relationship Between Student Performance Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79 (4): 33–53.

- Murnane, R.J., J.B. Willett, and F. Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics* 77 (2): 251–266.
- National Center for Education Statistics. 2003. "Remedial Education at Degree-Granting Postsecondary Institutions in Fall 2000." Washington.
- National Center for Education Statistics. 2009. "The Nation's Report Card: Mathematics 2009." Washington.
- Petrilli, M. 2011. "Lights, Camera, Action! Using video recordings to evaluate teachers." *Education Next* 11 (2): 85–87.
- Pianta, R., K.M. LaParo, and B.K. Hamre. 2004. *Classroom Assessment Scoring System*. Charlottesville, VA: University of Virginia.
- Prendergast, C. 1999. "The Provision of Incentives in Firms." *American Economic Review* 37 (1): 7–63.
- Quillen, I. 2011. "Room with a View." *Education Week*. June 15.
- Rothstein, J. 2009. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy* 4 (4): 538–572.
- Samuels, C.A. 2011. "Cheating Scandals Intensify Focus on Test Pressures." *Education Week*. August 4 (http://www.edweek.org/ew/articles/2011/08/04/37cheating_ep.h30.html?r=468391506).
- Sawchuck, S. 2011. "Wanted: Ways to Measure Most Teachers." *Education Week*. February 2.
- Schmitt, L., and N. Ibanez. 2011. "AISD REACH Program Update: 2009-2010 Texas Assessment of Knowledge and Skills (TAKS) Results and Student Learning Objectives (SLOs)." Austin, TX: Austin Independent School District Department of Program Evaluation.
- Skinner, E.A., and M.J. Belmont. 1993. "Motivation in the Classroom: Reciprocal Effects of Teacher Behavior and Student Engagement Across the School Year." *Journal of Educational Psychology* 85 (4): 571–581.
- Taylor, E.S., and J.H. Tyler. 2011. "The Effect of Evaluation on Teacher Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers." Working Paper w16877. National Bureau of Economic Research.
- Tucker, P.M. 1997. "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 (2): 103–126.
- Tyler, J.H., and M. Lofstrom. 2009. "Finishing High School: Alternative Pathways and Dropout Recovery." *America's High Schools* 19 (1): 77–103.
- Watson, J.G., S.B. Kraemer, and C.A. Thorn. 2009. "The Other 69 Percent." Washington: Center for Education Compensation Reform.
- Weisberg, D., and others. 2009. "The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness." *Education Digest: Essential Readings Condensed for Quick Review* 75 (2): 31–35.
- Yoon, K.S., and others. 2007. "Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement." Washington: National Center for Education Evaluation and Regional Assistance.

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

