Policy Brief

OCTOBER 2010

Value-Added Measures of Education Performance:

Clearing Away the Smoke and Mirrors

Douglas N. Harris
University of Wisconsin at Madison

Policy Brief 10-4

Douglas N. Harris is an economist and Associate Professor of Educational Policy and Public Affairs at the University of Wisconsin at Madison and advisor to state and federal policymakers on teacher quality issues. He chaired the 2008 National Conferences on Value-Added and his research on teacher value-added has been published and cited in academic journals and in the *New York Times*, *Wall Street Journal*, and other media.

Measuring performance is challenging and holding people accountable for their performance is even more so. With almost two decades of the "new accountability" with high-stakes testing behind us, it is clear that measuring performance is an even greater challenge in education than in many other lines of business.

The federal *No Child Left Behind* (NCLB) is a strong example of well-meaning intentions gone awry. The NCLB designers understood that the country needed to make major improvements in its education system and they understood that measuring performance based on student outcomes could serve as a useful tool for school improvement. But they failed to adequately address some important complexities. For this reason, problems with NCLB are numerous: narrowing the curriculum, teaching to the test, focusing on "bubble kids" (those who fall right below the proficiency line) and so on.

Another critical problem is less widely recognized. Current federal policies do not account for the fact that student outcomes are produced by more than just schools. As a result, they fail to follow what I will call the "Cardinal Rule of Accountability": hold people accountable for what

Executive Summary

Current federal policies do not account for the fact that student outcomes are produced by more than just schools. As a result, they fail to follow what Douglas Harris calls the "Cardinal Rule of Accountability": hold people accountable for what they *can* control. California's policies are no better in this regard.

Indeed, measuring performance is challenging and holding people accountable for their performance is even more so. With almost two decades of the "new accountability" with high-stakes testing behind us, it is clear that measuring performance is an even greater challenge in education than in many other lines of business. We know that student outcomes are a product of more than just schools, yet current federal policies do not account for this fact. In this policy brief, Douglas Harris explores the problems with attainment measures when it comes to evaluating performance at the school level, and explores the best uses of value-added measures. These value-added measures, Harris writes, are useful for sorting out-of-school influences from school influences or from teacher performance, giving us overall better performance measures.

continued on page 2

Executive Summary continued

Value-added measures provide summative assessments of teacher performance. They indicate whether teachers are doing well or not, on one important measure of student performance. But value-added is often criticized for not providing information about *how to improve*. Value-added measures can be made more "formative" in this sense by providing multiple value-added measures for teachers on the subject or specific domains covered by tests.

No single measure can fulfill both the formative and summative functions very well, however. For this reason, any use of value-added, especially for individual teachers, should be coupled with observational information from school principals or peer assessors. These additional performance measures can provide more formative information to help teachers take concrete steps forward, and may also help reduce systematic and random errors by providing additional information for the overall performance assessment. Multiple measures provide more information, and more information is generally better. The issue is not whether to use value-added, but how to do so.

they can control. California's policies are no better in this regard.

Violating this principle of accountability, has important negative consequences, as I show in the next section. Fortunately, growth and "value-added" measures can help sort the out-of-school influences from school influences or teacher performance. After explaining the basics of value-added measures in the third section, I discuss the strengths and weaknesses of value-added measures, including evidence of how systematic and random errors can produce misleading results. I conclude by discussing the best uses of value-added measures for accountability based on existing evidence.

The Promises and Perils of Typical Test-Based Accountability

Historically, the focus of attention in the U.S. education system has been on resources and rule compliance rather than student test scores. Policymakers debated how much money to spend and how to distribute those funds. Districts debated teacher salaries, class size reduction, administration, textbooks, buses, and more. Together, state legislatures and school districts also created rules for how those resources were to be used by creating rules about, for example, financial accounting, the size and shape of new schools, student disciplinary practices, and the curriculum.

The resource and compliance focus also extended to teachers. States have long required teachers to be licensed to teach, and they have funded university-based programs to ensure a supply of teachers meeting certification requirements. Certified teachers are hired, and then work in the classroom for a 2-5 year probationary period. At that point, the vast majority of teachers are granted tenure, providing them considerable job security. Most teachers are paid based on a single salary schedule, based on their university degrees and years of experience. Many of these policies developed in the mid-1900s with the support of, and in negotiation with, local teacher unions.

In theory, the resource and compliance approach could be effective: 1) if the certification process ensures that teachers are well trained and if they put to good use the autonomy provided by tenure and salary schedules and; 2) if teacher evaluation systems eliminate the ineffective teachers who slip through the cracks of the certification system. However, none of these assumptions has proven to hold in practice. Certified teachers are, at best, slightly more effective than uncertified teachers.[1] One well respected scholar of teacher education criticizes the quality of the teacher evaluation system in which the vast majority of teachers are given the highest possible rating and where the "technical core" of teaching is largely ignored relative to general teacher activities.[2] Under these conditions, it is unsurprising that one recent survey found that "78% [of teachers] say their school has at least a few teachers who are simply going through the motions, and just 14% say it is easy to remove incompetent

teachers."[3] Indeed, it is important that teachers care about their students and have the appropriate formal training, as the vast majority do, but these factors alone are not enough to ensure success for students.

A central fact driving the current reform effort is the wide variation in individual teachers' effectiveness in producing academic growth.[4] Teachers and administrators have known this for a long time, as they see differences in performance first-hand. But the flaws in today's credentialing system are making it increasingly difficult for anyone, including educators themselves, to support it. As Randi Weingarten, President of the American Federation of Teachers, noted in a recent speech, "Our system of evaluating teachers has never been adequate. For too long and too often, teacher evaluation—in both design and implementation—has failed to achieve what must be our goal: continuously improving and informing teaching so as to better educate all students."[5]

The New Accountability and Unintended Consequences

The weaknesses of the resource and credentialing approaches helped create a broad consensus during the 1990s that the school system needed to focus more on outputs, specifically, on student academic outcomes. As part of this new accountability, states increased the frequency of student testing and required each school to have a report card summarizing, for example, the percentage of students who were proficient in reading and math, or on another measure of student attainment. But proficiency is just a snapshot of student outcomes at a single point in time that lacks a proper context.

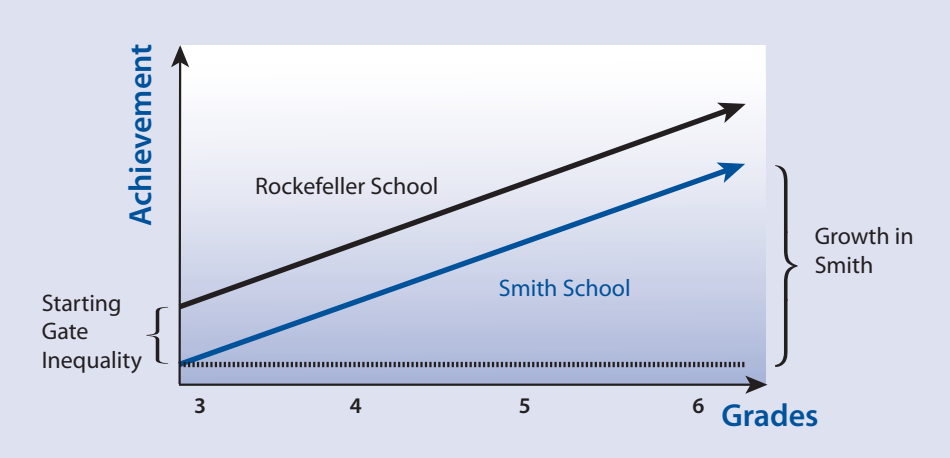
The problem with a snapshot approach has to do with a basic fact of human development: knowledge and skill accumulate over the full course of a person's lifetime. What adults know and can do depends on what they experienced as students, which in turn depends on their experiences prior to school, all the way back to their gestation and birth.[6] Kindergarteners arrive at the classroom door with vastly different early childhood experiences and levels of readiness for school. For example, at the very beginning of kindergarten, high-income children have average test scores that are 60 percent higher than low-income children.[7] Schools cannot have caused these "starting-gate inequalities," because most students haven't set foot in a classroom before kindergarten. Yet the inequalities are so large and persistent that even effective schools cannot completely overcome them. Non-school factors continue to influence children as they progress through school. These factors are outside the control of schools and failing to account for them, as attainment measures do, amounts to violating the Cardinal Rule of Accountability.

Does it really matter how we use student test scores to measure school performance? Unfortunately, yes. When school performance is measured poorly it creates a variety of perverse incentives, incentives to do things that are clearly inconsistent with a school's mission. Attainment measures are partly responsible, for example, for pressuring schools to exclude students from testing through "selective disciplining."[8] They can also help to push good teachers out of schools serving low-performing students, as these teachers become frustrated by a system that punishes them no matter how well they perform.[9]

Undue reliance on attainment measures can also lead low-attainment schools to spin in a never-ending cycle of reform, which may include frequent changes in curriculum, instruction, and leadership. A continual cycle of reform might make sense if those schools were in fact failing, but schools that are typically labelled "low-performing" are often not failing by any reasonable definition. As I will show below, many low-attainment schools are actually high-performing. The reverse is also true, though problems of poor performance are generally well hidden in high-attainment schools.

The new accountability was in some ways a step in the right direction, but the cascade of sanctions unleashed on schools by *No Child Left Behind* introduced more problems than it should have. Accountability should hold people and organizations accountable for what they can control, yet accountability polices have been violating this principle for decades. Fortunately, there is a better way.

FIGURE 1. School Performance Measurement under Attainment Models



Value-Added: The Basics

If the Cardinal Rule of Accountability is that people should be held accountable for what they can control, and if the goal of accountability policies is to improve student outcomes, then we need to find a measure of performance that captures what *each* school and *each* teacher contributes to student outcomes.[10] But how can we do that? As I show below, value-added measures offer two crucial advantages over attainment measures, by focusing on student growth and comparing similar schools.

The simplest value-added measure is growth. Subtract the initial level of student achievement from the end-of-year test score, and the answer is a measure of growth. In Figure 1, by way of illustration, each arrow represents the achievement growth of students over time in two hypothetical schools, Rockefeller and Smith. The difference in initial achievement levels (i.e., the starting-gate inequality) is indicated by the bracket in the lower-left-hand corner. The students at the Rockefeller School clearly start off in a stronger position than those in the Smith School. Because these differences are based on where students started when they walked in the door, however, they are outside the control of the schools. Using growth measures rather than attainment measures allows us to compare schools after having subtracted the starting gate inequality.

Notice that attainment is quite different for the two schools, but that growth is the same. This example highlights the fact that attainment and growth measures can yield different conclusions about performance. If we were to judge these schools based on student attainment, as we have done for decades, Rockefeller would be the clear winner. Based on a measure of value-added, however, the performance of the two schools is the same. Given that starting-gate inequalities are outside a school's control, growth provides a better measure of performance than attainment.

The situation described in Figure 1 is more than hypothetical. Studies show that there are many real schools that have low attainment but high growth.[11] In my own data analyses, for example, one school ranks at the 99th percentile on attainment, but at the 14th percentile in growth. This is an extreme case, but large differences between attainment and growth are not unusual.

Given how important the students themselves are to their own outcomes, and the fact that different schools have very different types of students, accounting for where students start is crucial. It is also consistent with the widely shared principle that schools should take students where they start and help them learn and grow as much as possible.

From Growth to Value-Added

While growth measures represent a substantial improvement over attainment, they still have weaknesses. First, schools differ in the resources they have to work with. The focus on resources in decades past was not completely unwarranted, because significant inequities in funding and other resources persist, largely outside the control of schools.

One possible way to account for these weaknesses is simply to compare student growth among similar schools. We could imagine one group of schools that has few disadvantaged students and low school resources, another group with half disadvantaged students and low school resources, and so on. This simple categorization would divide

schools into nine different "buckets" and we could make comparisons within each bucket. Unfortunately, if we want to make these comparisons in a more fine-grained way, we would need to compare thousands of schools. California is very large, but even California might not be large enough to make comparisons in every bucket. We therefore need a different way to approach the issue.

Another way to think about the usefulness of growth measures is that they allow us to make better predictions of future student achievement. What is the best way to predict how students will do next year? We can do this by looking at students' growth during the previous year.

Figure 2 illustrates this point for the Smith School (once again, shown by the solid line). Predicted achievement is shown by the dashed line. The difference between the actual and predicted scores shows the value-added of the school.

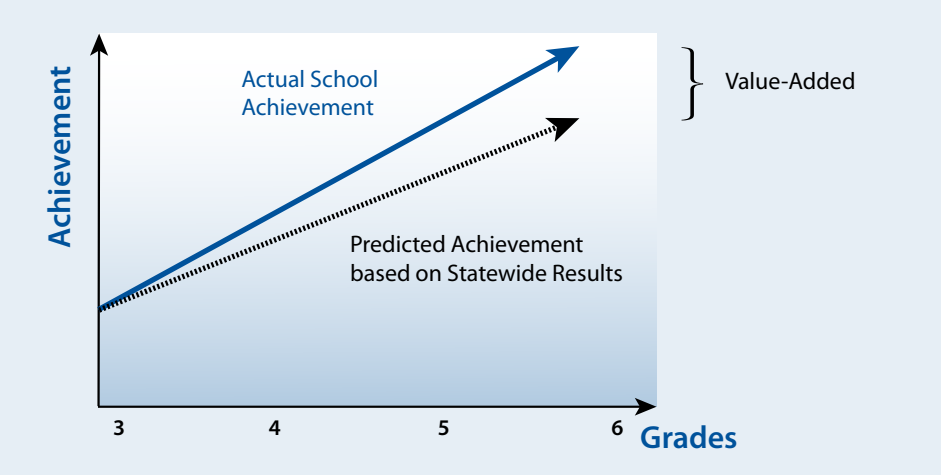
What factors should be taken into account in the prediction other than prior achievement? Class size, funding, and staff positions are some examples. If Smith Elementary has fewer of these important resources, then the prediction line in Figure 2 would be lowered and this would increase the value-added measure. Unfortunately, many critical school resources, like leadership, are difficult or impossible to measure, and if we cannot measure them, we cannot account for them in our predictions.

One of the most controversial issues in value-added is whether to account for student demographics. Strictly speaking, our definition of performance requires accounting for student background because growth alone does not account for all important differences among students. With limited engagement in learning activities during the summer, for example, some students experience a significant summer learning loss, forgetting what they learned during the school year. As with starting-gate inequalities, disadvantaged students experience much more summer-learning-loss than other students.[12] And the factors that generated these inequalities, such as family and community resources, may still influence students as they progress through school. This strengthens the case for taking student demographics into account in our predictions.

On the other hand, using student demographics as part of the prediction could be criticized because of the risk that it could involve setting lower attainment standards for disadvantaged groups. If the concern is that schools with more disadvantaged students will be able to reach the same performance rating with lower student growth, then the critics have a point. But if the concern is that schools may have an incentive to devote less *effort* to disadvantaged students, then the concern is unjustified. Value-added measures can be designed to give as much or as little weight to a given group as is deemed necessary.

The goal is to create a measure of performance that fits the Cardinal Rule of Accountability. Value-added does this in two ways: 1) taking into account where students start when they first walk into school and 2) comparing schools that are similar in terms of measureable school resources—or, more specifically, using a prediction approach that gives a reasonable head-start to schools that operate with fewer resources, making more reasonable comparisons possible.

FIGURE 2. Value-Added Measures of School Performance: Smith Elementary



Attainment, Value-Added, and Growth-to-Proficiency

Value-added measures yield very different performance results when compared with student attainment. Dr. Michael Weiss has estimated attainment and value-added measures and finds that the correlation between value-added and attainment (specifically, percent proficient) is only 0.47.[13]

Weiss also compares value-added and attainment measures to the "growth-to-proficiency" models adopted on a pilot basis by the U.S. Department of Education. In reality, growth-to-proficiency is essentially the same as proficiency itself (or any other version of an attainment model). Both models fail to take into account starting gate inequalities and therefore expect more value-added from some schools than from others. Weiss confirms this reasoning. He finds that there is a nearly perfect correlation between the likelihood that a school will be judged as failing under simple-proficiency models and under growth-to-proficiency models.

The Weiss result shows that the fact that federal accountability focuses on "adequate yearly *progress*" (AYP, italics added) does little to improve matters. One problem is that AYP pertains to progress made within each grade across years. This is quite different from tracking individual students over time. AYP is based on the idea that all students should obtain the same level of achievement, but the focus on proficiency requires schools with low-attainment students to generate more growth than schools with high-attainment students. This is the central problem with attainment measures.

The California Academic Performance Index (API) suffers from a similar problem. The API is a complex attainment measure, and schools are held accountable for growth in their API over time. This is quite different from value-added, however. Performance in value-added is based on how much *the same cohort* of students learns while going through school, compared with their predicted learning gains. In contrast, performance in API is based on how the performance of *different cohorts* of students at the same grade level changes from year to year. As was previously shown, there are many schools with high value-added and low attainment, including schools where API growth is low. The opposite is also true: schools where value-added is low may score high on attainment measures, including API.

We should hold people accountable for what they can control and value-added measures are better for this purpose than attainment. Moreover, there are a lot of imitators such as AYP, growth-to-proficiency, and API that look as if they measure growth but are really just more complicated attainment measures. We can do better.

Strengths and Weaknesses of Value-Added

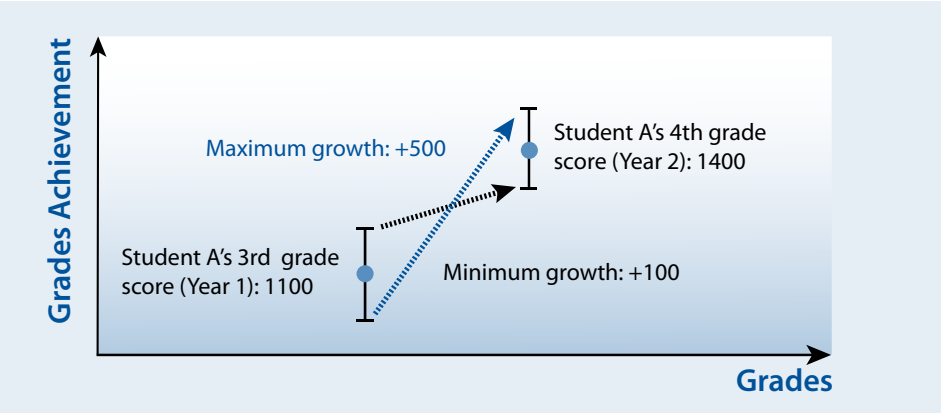
While value-added has important advantages over attainment and related measures, it is far from perfect. Below, I discuss two general types of errors we can make with performance measures and show how these come into play with value-added.

Random Error versus Systematic Errors

There are two general types of errors that are always present in varying degrees in all measures, not just value-added and not just performance measures. *Systematic error* is easy to explain because it requires only restating the problem with attainment measures. Attainment-based school performance measures like proficiency are *systematically* biased against schools serving low-attainment students. That is, by failing to account for factors affecting achievement that are outside the school's control, we systematically under-estimate the performance of low-attainment schools. *Random errors* in contrast are equally likely for any teacher or school. Some errors in value-added measures are also random.

One of the disadvantages of any type of growth measure is that random errors are larger than with attainment measures. To see why, consider the following example: suppose that a student scores 1100 on a test in the first year (an attainment measure). For various reasons, we are never 100% sure that this reported score accurately represents the student's true level of knowledge and skill. Therefore statisticians always encourage reporting "confidence intervals" which indicate a range of possible values. We can be reasonably certain that the true level of achievement for this student on the tested content is within the confidence interval of, say, plus or minus 100

FIGURE 3. How Random Error Increases with Growth and Value-Added Measures



points. That is, the confidence interval around 1100 is 1000-1200.

Now, suppose that the student scores 1400 on the test the following year, so that her observed growth is 300 points (from 1100 to 1400). This situation is illustrated in Figure 3. It would be tempting to think that the confidence interval for this growth measure is still plus or minus 100 points, but the range is actually much wider. The student might have scored as low as 1000 on the first test and as high as 1500 on the second test for a gain of 500 points. Or, the student might have scored as high as 1200 the first time and 1300 the second time, for growth of just 100 points. In short, random error gets much worse when we focus on growth. As I will discuss, since value-added is based on growth, random error is a major problem with value-added measures.

Random errors might seem innocuous because they are equally likely to arise with all teachers. But random errors are problematic because they call into question the conclusions we wish to draw from performance. Thus both systematic and random errors need to be taken into account when making decisions about performance measures.

Evidence on Systematic and Random Errors in Value-Added

The major strength of value-added measures, relative to attainment, is that they reduce systematic errors by taking into account factors that influence each student's achievement. But value-added measures are imperfect and still involve some systematic error. For example, summer learning loss is worse for disadvantaged students, creating systematic errors that favor schools with more advantaged students. Accounting for student demographics might help alleviate this problem, but would not solve it. Likewise, failure to account for school resources also introduces systematic error, in this case favoring schools with more resources.

Nevertheless, there are a few promising signs that value-added provides useful evidence about performance, at least for teachers. Several studies find a positive link between teacher value-added measures and principals' subjective assessments of teacher effectiveness, reporting correlations of 0.2-0.4.[14] These correlations might seem low, but note that the high end of the range is nearly as high as the correlation between value-added and attainment, two measures that are based on identical student achievement data. Moreover, it is clear that principals care about aspects of teacher performance that are unrelated to student achievement scores, so we would not expect the correlation between value-added and principal assessments to be perfect in any case.

Since what students bring to the classroom is important and varies widely, there is good reason to ask questions about how tracking students into different schools and classrooms might influence value-added measures.[15] While this issue briefly gained attention because of a study published by Jesse Rothstein, UC Berkeley, subsequent analysis by Julian Betts and Cory Koedel of UC-San Diego suggests that the problem may not be as great as it first appeared.[16] A third study by Tom Kane and Douglas Staiger addressed the tracking issue by identifying schools in Los Angeles that were willing to randomly assign students to teachers instead of tracking them.[17] They found that teacher performance in classrooms with randomly assigned students was a good predictor of teacher value-added, at least on average.

Since the reduction of systematic error is a strength of value-added, especially when compared with attainment, I return to the issue of random error. One consequence of large random errors is that it is hard to conclude that one teacher is truly better than another. That is, the confidence intervals around a teacher's true performance are wide. One teacher's value-added may appear larger than that of another, but we cannot be confident when implementing the usual statistical standards. Using data from North Carolina, Kane and Staiger conclude that about half the variation in grade-level achievement gains is due to random error rather than to real differences in performance.[18]

The problem is probably worse for individual teachers because each teacher has a small number of students and therefore has less test-score information.[19] Other researchers have shown that teacher value-added scores are imprecise enough that, by the usual standards of statistical significance, it is only possible to clearly distinguish very-low-value-added teachers from very-high-value-added teachers (Jacob & Lefgren, 2005). This is a problem for policies that intend to make nuanced distinctions among teachers for high-stakes decisions. It means, for example, that some average schools and teachers will be rewarded or punished unjustifiably.

Random error also means that value-added measures can be unstable over time even for individual teachers. Intuitively, we would expect that the actual effectiveness of each teacher should change little from year to year. Teachers might gradually improve over time, but it is unlikely that they would jump from the bottom to the top of the performance distribution. If we see value-added measures change significantly over short periods of time, random error is the most likely culprit.

Some of the earliest evidence on this topic, however, suggested that teacher value-added measures are much less stable than this intuition would suggest. Cory Koedel and Julian Betts found that only 50 percent of teachers who ranked in the top fifth of teachers on teacher value-added one year were still ranked in the top fifth in the subsequent year.[20] This suggests that half of high-performing teachers actually got worse relative to their peers over a short period of time, some dramatically worse. McCaffrey, Lockwood, Sass, and Mihaly (2009) show that stability increases by 40-60 percent when aggregating data across two years and by an additional 18-23 percent when adding a third year.[21] This is important, as it suggests that accountability policies calling for the use of teacher value-added measures should include requirements that many years of data be used for each teacher.

There are also significant and valid concerns about the achievement tests underlying value-added measures. In addition to the test content and scaling procedures, there are questions about administering tests mid-year (rather than at the end of the year) and about how to deal with students switching schools mid-year. These problems all contribute to the systematic and random errors mentioned above. Another limitation, that tests are not available in all subjects and grades, means that value-added can only be applied to a small percentage of teachers. This is not ideal, though teachers in different subjects and grades will always be evaluated differently, no matter what type of evaluation system is used.

Again, no performance measure is perfect and value-added is no exception. Random and systematic errors in these measures reduce their accuracy and make it less likely that the conclusions we draw are accurate.

Using Value-Added

A basic principle of measurement is that the validity of a measure depends on what conclusion one is trying to draw from it. Thus, value-added measures are good for some purposes and not for others.

School Versus Teacher Value-Added

Just as measuring school performance is less controversial and more widespread than measuring teacher performance, school value-added is less controversial than teacher value-added. Beyond this, the average school has far more students than the average teacher, which helps to reduce random error. With schools, therefore, the question is not whether but how to use student achievement scores. Value-added is preferable to attainment because this approach takes into account factors outside of school control.

The situation is more uncertain with teacher value-added. On the one hand, teacher value-added measures are unstable and probably do not fully account for student tracking. On the other hand, teachers are arguably the most important school resource, teacher performance appears to vary widely, and there is general agreement that current teacher evaluation systems do not work well. There is therefore a strong case for improving the evaluation of individual teachers, not just whole schools where low performance by some teachers can be hidden by the performance of their high-performing colleagues. The question is whether teacher value-added is better than the existing system of evaluation and other feasible policy alternatives such as improved principal and peer assessments.

A middle ground option is to evaluate teams of teachers by subject area (math or reading) or grade. This is an attractive option for many reasons. It may be seen as less threatening to individual teachers. It could also help to facilitate coordination among teachers (although teacher value-added measures, estimated correctly on a state-wide basis, should not create competition among teachers within schools). Team value-added, because it includes more students and avoids tracking concerns, would also reduce systematic and random errors.

Low-Stakes versus High-Stakes

Many discussions of value-added start with the assumption that the measures will be used for teacher merit pay, tenure, and other high-stakes decisions. These are possibilities, but they are not the only ones. The least controversial use of value-added is for program evaluation. Indeed, many of the studies cited above do not report value-added for any individual teacher or school. Instead, they look for patterns across large numbers of teachers and schools to learn about, for example, the effects of professional development programs or how well teacher certification distinguishes effective from ineffective teachers.

Other options include calculating value-added for individual teachers and providing this information privately to teachers and their principals, without attaching high-stakes, to inform professional development plans and provide teachers with a clearer sense of how well they are doing.

Formative versus Summative

Value-added measures provide summative assessments of teacher performance. They indicate whether teachers are doing well or not, on one important measure of student performance. But value-added is often criticized for not providing information about *how to improve*. Value-added measures can be made more "formative" in this sense by providing multiple value-added measures for teachers on the subject or specific domains covered by tests (though such measures have greater random error).

No single measure can fulfil both the formative and summative functions very well, however. For this reason, any use of value-added, especially for individual teachers, should be coupled with observational information from school principals or peer assessors. These additional performance measures can provide more formative information to help teachers take concrete steps forward, and may also help reduce systematic and random errors by providing additional information for the overall performance assessment. Multiple measures provide more information, and more information is generally better. Again, the issue is not whether to use value-added, but how to do so.

Moving Forward: Recommendations

Two things are clear from this discussion. First, value-added measures are preferable to attainment measures (including the API and current federal measures) when it comes to evaluating performance at the school level. Second, the current system of teacher evaluation and accountability is sufficiently discredited that it is at least worth experimenting with some uses of teacher value-added. This should be done as part of a comprehensive system that includes other measures.

The analysis and evidence also point toward several recommendations as to how policymakers should move forward:

- ***Build data systems that track individual students over time and link teachers to students.*** Regardless of whether and how value-added

measures are used, school systems should be able to track the trajectory of every student, even students who switch schools and districts. Linking students to teachers remains controversial, but doing so is prerequisite to a more informative use of student achievement scores.

- ***Use value-added techniques to evaluate school programs.*** We need to hold schools and teachers accountable for their performance, but we also want to know whether the programs that we implement in schools are working. Value-added can be a powerful tool for program evaluation, helping us to learn whether new professional development programs or alternative compensation policies have positive effects on teacher performance or student learning. For example, many studies focused on teacher credentials have applied value-added logic to determine whether additional pre-service training leads to improved student outcomes.

- ***Experiment with, and evaluate, different uses of teacher and team-based value-added measures.*** Using value-added for program evaluation has obvious potential advantages, but there is very little evidence on how value-added measures can best be used for purposes of accountability and diagnosis. The state government and school districts should carry out pilot programs and carefully evaluate them to learn which approaches are most effective.

- ***Combine value-added with other measures.*** All forms of evaluation contain systematic and random errors. Value-added is no different in this respect, though the errors in value-added measures are more widely reported. Value-added measures are also summative. Combining value-added with measures of the quality of teacher practice can help address all of these limitations by providing more formative feedback, reducing systematic error, and reducing random error.

- ***Tread carefully.*** As policymakers move forward toward productive experimentation with value-added, they should avoid becoming overconfident in the ability of these measures to accurately distinguish performance with any degree of nuance. Value-added measures have potential, but we cannot lose sight of their limitations or of their larger purpose: measuring performance in a way that drives genuine improvement in teaching and learning.

Endnotes

[1] See the following for evidence that traditional certification does not distinguish highly effective teachers: Dan D. Goldhaber and Dominic J. Brewer, "Does Teacher Certification Matter? High School Teacher Certification Status and Student Achievement," *Education Evaluation and Policy Analysis.* Vol. 22, No. 2 (2000) 129-145. For evidence of a similar result with regard to National Board Certification, see: Dan Goldhaber and Emily Anthony, "Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching," *Review of Economics and Statistics.* Vol. 89, No. 1 (2007) 134-150 and Douglas N. Harris and Tim R. Sass, "The effects of NBPTS-certified teachers on student achievement," *Journal of Policy Analysis and Management.* Vol. 28, No. 1 (2009) 55–80. For exceptions to the rule regarding state certification, see the following studies finding a positive association between certification and performance in high school: Charles T. Clotfelter, Helen F. Ladd, and Jacob L. Vigdor. *Teacher Credentials and Student Achievement in High School: A Cross-Subject Analysis with Student Fixed Effects.* Center for the Analysis of Longitudinal Data for Education Research (CALDER). (Washington, DC. Urban Institute: 2007).

[2] Mary M. Kennedy, "Recognizing Good Teaching When We See It," in *Handbook of Teacher Assessment and Teacher Quality,* Ed. Mary Kennedy (San Francisco: Jossey Bass, 2010).

[3] Farkas, S., Johnson, J., & Duffett, A. (2003). *Stand by Me: What Teachers Really Think About Unions, Merit Pay, and Other Professional Matters.* New York: Public Agenda.

[4] Sanders, William L. and Sandra P. Horn. "Research Findings from the Tennessee Value-Added Assessment System (TVASS) Database: Implications for Educational Evaluation and Research." *Journal of Personnel Evaluation in Education.* Vol. 12, No. 3 (1998): 247–256. Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools and Academic Achievement." *Econometrica.* 73 (2005): 417–458.

[5] Randi Weingarten (2010) *A New Path Forward: Four Approaches to Quality Teaching and Better Schools.* January 12, 2010. Accessed September 10, 2010 from: http://www.aft.org/pdfs/press/sp_weingarten011210.pdf

[6] Heckman, James J. 2006. "Skill Formation and the Economics of Investing in Disadvantaged Children." *Science.* Vol. 312, No. 5782 (June 30): 1900-1902.

[7] Valerie Lee and David Burkam. *Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School.* 2002. Washington, DC: Economic Policy Institute.

[8] David N. Figlio, "Testing, Crime and Punishment," *Journal of Public Economics.* Vol. 90, No. 4-5 (2006), 837-851.

[9] Charles T. Clotfelter, Helen F. Ladd, Jacob L. Vigdor, and Roger Aliaga Diaz, "Do School Accountability Systems Make It More Difficult for Low-Performing Schools to Attract and Retain High-Quality Teachers?" *Journal of Policy Analysis and Management.* Vol. 23, No. 2 (2004) 251–271.

[10] McCaffrey et al. (2003, p.10-13) define value-added as follows: "We suggest defining a teacher's effect as the average causal effect on student achievement across all students of interest."

[11] Michael Weiss "Can We Project Future Proficiency? Examining the Measures Used in the Federal Growth Model Pilot Program." Paper presented at the 2008 Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC, March 3, 2008.

[12] Harris Cooper, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. "Summer Learning Loss: The Problem and Some Solutions," *Review of Educational Research* Vol. 66, No. 3 (1996) 227-268. Karl L. Alexander, Doris R. Entwisle, and Linda S. Olson. *Schools, Achievement, and Inequality: A Seasonal Perspective Educational Evaluation and Policy Analysis.* Vol. 23, No. 2 (Summer, 2001), 171-191. David T. Burkam, Douglas D. Ready, Valerie E. Lee, and Laura F. LoGerfo, "Social-Class Differences in Summer Learning Between Kindergarten and First Grade: Model Specification and Estimation," *Sociology of Education.* Vol. 77, No. 1 (Jan., 2004) 1-31.

[13] Weiss, Michael J. "Examining the Measures Used in the Federal Growth Model Pilot Program." Paper Presented at the Annual Meeting of the Society for Research on Educational Effectiveness, Washington, DC, March 3, 2008.

[14] Douglas N. Harris and Tim R. Sass, "What Makes for a Good Teacher and Who Can Tell?" National Center for the Analysis of Longitudinal Data in Education Research (CALDER). Working Paper #30. (Washington, DC: Urban Institute, 2009). Brian Jacob and Lars Lefgren "Principals as Agents: Subjective Performance Assessment in Education," *Journal of Labor Economics.* Vol. 26, No. 1 (2007) 101-136.

[15] Rothstein's (2009) study found that he could "predict" student achievement gains in the past by knowing which teachers students had in the future. Since teachers cannot change the past, this suggests that tracking is occurring. Koedel and Betts (2007) found that when using three years of data for teachers, in contrast to one year as in Rothstein's study, there was no longer evidence of tracking. Jesse Rothstein (2009) "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy.* Vol. 4, No. 4 537-571.

[16] Koedel, Cory, and Betts. "Re-Examining the Role of Teacher Quality in the Educational Production Function." Working Paper #2007-03. Nashville, Tenn.: National Center on Performance Initiatives, 2007.

[17] Thomas Kane and Douglas Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," National Bureau of Economic Research. Working Paper #14607 (2008).

[18] Thomas J. Kane and Douglas O. Staiger "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives.* Vol. 16, No. 4, (Autumn, 2002) 91-114.

[19] It is not entirely certain that random error is a larger problem for teachers than for schools. The reason is this: if teacher performance varies widely, then the differences between teachers are large and random error is less problematic. Conversely, if much of the variation in teacher performance is within schools, then the differences in school performance are probably small, making even a small amount of random error problematic.

[20] Koedel, Cory, and Betts. ibid.

[21] Daniel F. McCaffrey, Tim R. Sass, J. R. Lockwood, and Kata Mihaly, "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy.* Vol. 4, No. 4 (2009) 572–606.

PACE

Policy Brief

School of Education, Stanford University
520 Galvez Mall
CERAS Rm. 401
Stanford, CA 94305-3084
(650) 724-2832
http://www.edpolicyinca.org

We would like to thank The James Irvine Foundation, The William and Flora Hewlett Foundation, and the Carnegie Corporation of New York for financial support for the publication of this policy brief. The views expressed are those of the authors, and do not necessarily reflect the views of PACE or its funders.

Recent PACE Publications

- Jennifer L. Steele, Richard J. Murnane and John B. Willett. *Do Financial Incentives Draw Promising Teachers to Low-Performing Schools?* Policy Brief 10-3, May 2010.
- Martin Carnoy. *California's Impending College Graduate Crisis and What Needs To Be Done About It.* Policy Brief 10-2, April 2010.
- Charles Taylor Kerchner. *There's Lots to Learn from L.A.: Policy Levers for Institutional Change.* Policy Brief 10-1, January 2010.
- Susanna Loeb and Jon Valant. *Leaders for California's Schools.* Policy Brief 09-4, September 2009.
- Sean F. Reardon and Michal Kurlaender. *Effects of the California High School Exit Exam on Student Persistence, Achievement, and Graduation.* Policy Brief 09-3, September 2009.
- Heather J. Hough. *The Quality Teacher and Education Act in San Francisco: Lessons Learned.* Policy Brief 09-2, May 2009.
- Heather J. Hough and Susanna Loeb. *The Development of a Teacher Salary Parcel Tax: The Quality Teacher and Education Act in San Francisco.* May 2009.
- Julia E. Koppich and Jessica Rigby. *Alternative Teacher Compensation: A Primer.* March 2009.
- Katharine Strunk. *Collective Bargaining Agreements in California School Districts: Moving Beyond the Stereotype.* Policy Brief 09-1, January 2009.
- *Conditions of Education in California,* October 2008.
- Susanna Loeb and David Plank. *Learning What Works: Continuous Improvement in California's Education System.* Policy Brief 08-4, August 2008.
- Julia Koppich. *Reshaping Teacher Policies to Improve Student Achievement.* Policy Brief 08-3, March 2008.
- Susanna Loeb, Tara Beteille, and Maria Perez. *Building an Information System to Support Continuous Improvement in California Public Schools.* Policy Brief 08-2, February 2008.
- Jennifer Imazeki. *Meeting the Challenge: Performance Trends in California Schools.* Policy Brief 08-1, February 2008.
- Anne K. Driscoll. *Beyond Access: How the First Semester Matters for Community College Students' Aspirations and Persistence.* Policy Brief 07-2, August 2007.
- W. Norton Grubb and David Stern. *Making the Most of Career-Technical Education: Options for California.* Policy Brief 07-1, April, 2007.