Great teachers matter enormously to the success of their students. While this seems like common sense, it has also been borne out by decades of empirical research: Several studies show that to which teacher a student is assigned makes a huge difference in determining how much that child will learn in a given school year. And many of the benefits of good teachers are more long-lasting. A new study by researchers at Harvard and Columbia, for instance, shows that a student assigned to a great teacher is less likely to have an early pregnancy, is more likely to attend college, and will earn a higher salary as an adult.

In response to such findings, policymakers across the nation are aggressively pursuing policies to address teacher quality. At the federal level, teacher-quality reforms played a central role in President Obama’s Race to the Top program. And several states—including Colorado, Florida, Indiana, Connecticut, New Jersey, and New York—have recently adopted or seriously pursued reforms that would connect crucial employment decisions, such as those regarding salary and tenure, to teachers’ performance (rather than to seniority).

But of all the current reform proposals related to teacher quality, the most fundamental are those that would change the way public-school teachers are evaluated. The reason is straightforward: Any reform policy linked to teacher quality requires first being able to accurately assess teacher performance. It would be impossible to compensate teachers
based in part on their performance, or to improve the system’s ability to assign tenure, without an evaluation tool capable of distinguishing between the system’s best and worst teachers.

Unfortunately, the current system for evaluating public-school teachers makes no meaningful attempt to assess the influence that a teacher has on his students’ outcomes. Consequently, just about all of the nation’s public-school teachers are deemed “satisfactory” on their official evaluations.

In order to change this state of affairs, many policymakers and researchers have called for school systems to employ a statistical technique known as “value added” analysis, which uses standardized-test scores to assess a teacher’s performance in the classroom. Careful analysis of student test scores can provide valuable information about an individual teacher’s competence that is not accounted for under the current system, and research shows that, while not perfect, value-added analysis is a very promising improvement on today’s rubber-stamp approach to teacher evaluations.

Nevertheless, despite the serious flaws in the current system and the opportunities value-added analysis creates, this approach to teacher evaluation has turned out to be one of the most controversial questions facing reformers. Wary of being judged by their students’ scores on standardized tests, teachers have fought the use of value-added assessments in their performance evaluations through their unions. In this sense, the debate over using test scores to inform teacher evaluations is also a clear example of the fundamental problem facing education reformers: The interests of teachers do not always align with those of their students, and when those interests collide, the current system works to the teachers’ benefit.

The lack of a rigorous teacher-evaluation method is hardly a new problem in American education: It long pre-dates collective bargaining and powerful teachers’ unions. Consider that, as recently as 2006, only 0.89% of New York City’s teachers received “unsatisfactory” ratings on their official evaluations. This is roughly in keeping with the results from nearly a century ago: A 1914 study of New York City’s teacher-evaluation system found that only 0.5% of elementary-school teachers evaluated in the city’s schools were considered to be deficient in instruction, and only 0.9% were considered non-meritorious in discipline.

The author of that long-ago study argued that New York’s evaluation system would surely have to change in order to more accurately assess
teachers’ effectiveness. But it wasn’t until earlier this year that the city finally began taking some incremental steps toward improvement—when Mayor Michael Bloomberg proposed a new evaluation tool that would incorporate teachers’ observed performance in the classroom as well as their contributions to students’ standardized-test scores.

Meanwhile, most of the country is still languishing under the old rubber-stamp evaluation system that does nothing to meaningfully distinguish between excellent teachers and terrible ones. According to most evaluation systems in use today, there are in fact hardly any low-performing public-school teachers in the United States; in any given school system, nearly all of the teachers are rated as effective. Indeed, a 2009 report by the New Teacher Project looked at outcomes from teacher evaluations in 12 school districts across the United States and found that these systems rated fewer than 1% of their teachers as “unsatisfactory.”

Even obviously struggling urban public-school systems rate the vast majority of their teachers as performing well in the classroom. According to the National Assessment of Educational Progress—a highly respected standardized test administered by the U.S. Department of Education to representative groups of students—more than half of fourth graders in Chicago read below the “basic” level. Nevertheless, it is common for more than 60% of the city’s teachers to be rated as “superior,” and fewer than 1% are typically identified as “unsatisfactory.” In the Houston Independent School District, 43% of fourth graders read below the “basic” level and more than a third of eighth graders do. Yet, from 2005-06 to 2008-09, only 3.4% of public-school teachers were labeled “below expectations” or “unsatisfactory” on their performance evaluations. Even more discouraging, only two of the 661 teachers working in schools that Houston had formally recognized as being “academically unacceptable” were rated as performing unsatisfactorily on any single component of their official evaluations.

Such results are inconsistent with both empirical research and simple common sense. No one believes that 98% or more of today’s public-school teachers are effective. Indeed, both principals and teachers know that official evaluations are inflated. A 2009 survey of teachers in four large public-school districts found that 43% of teachers believed that there was at least one tenured teacher in their schools who should be removed from the classroom. In another survey, conducted by the New Teacher Project in 2007, more than half of veteran principals in Chicago
reported that they assigned undeservedly high evaluations to their teachers. In most public schools, which teachers are effective and which are not is an open secret.

It is so obvious that many undeserving teachers receive passing marks, in fact, that the issue has finally attracted the attention of Democratic policymakers at the national level. Shortly after he was sworn in, President Obama held a town-hall meeting at the White House, during which he had the following exchange with a Philadelphia teacher:

Obama: How long have you been teaching?
Teacher: Fifteen years.
Obama: Fifteen years. Okay, so you’ve been teaching for 15 years. I’ll bet you’ll admit that during those 15 years there have been a couple of teachers that you’ve met—you don’t have to say their names—(laughter)—who you would not put your child in their classroom. (Laughter.) See? Right? You’re not saying anything. (Laughter.) You’re taking the Fifth. (Laughter.) My point is that if we’ve done everything we can to improve teacher pay and teacher performance and training and development, some people just aren’t meant to be teachers, just like some people aren’t meant to be carpenters, some people aren’t meant to be nurses. At some point they’ve got to find a new career.

President Obama, the audience, and this teacher all know that there are bad teachers in America’s schools today. The uncomfortable laughter in the exchange occurred only because the president pointed to a fact that most people politely ignore. But as Obama’s remarks illustrate, more and more policymakers are willing to have an open and frank discussion about the controversial subject of teacher evaluation.

**WHY THE CURRENT SYSTEM FAILS**

That discussion must begin with *why* it is that, even though everyone knows some teachers don’t belong in the classroom, the current system continues to rubber-stamp them as qualified. After all, in any profession, trade, or job, there will be some duds and some superstars. That such distinctions should exist among the nation’s 3.3 million public-school
teachers seems self-evident. And yet the current system nevertheless officially rates all teachers as being equally effective.

One significant part of the problem is that, precisely because poor teacher evaluations are given so rarely, the designation implies that the recipient is not merely “unsatisfactory” but egregiously incompetent. This is often a much stronger message than a principal intends to send, and he is then left with no way to distinguish on an official evaluation between an ineffective teacher who could use some remediation and a hapless one who should find a new career. The default is just to list the teacher’s performance as “satisfactory” or higher. This practice hurts not only schools and students, but also struggling teachers in need of constructive feedback.

Another major problem is the basis for teacher evaluations: The current system relies entirely on limited observations of a teacher’s performance in the classroom. Typically, a principal sits in on a teacher’s classroom session at a time both have agreed on and assesses the teacher’s performance according to a protocol, evaluating the teacher’s classroom-management techniques, the set-up of the room, and the quality of his lesson plan.

Direct observation of a teacher’s performance in the classroom is a potentially valuable tool that should be used as part of a teacher’s evaluation. But observations alone are inadequate for assessing a teacher’s overall performance. In fact, the manner in which observations are typically conducted under the current system makes them laughably insufficient. To begin, current classroom observations are far too infrequent to be informative; indeed, teachers are hardly evaluated at all. Historically, novice teachers have been observed only once or twice during the school year, and in some school districts today, tenured teachers are observed only once every few years. For instance, in Los Angeles, tenured teachers who have been in the district for at least ten years and have been deemed “highly qualified” are observed once every five years.

When observations do take place, they are often brief: The official observation in the Miami-Dade school system need not last longer than 20 minutes. In its study of teacher evaluations in Arkansas, Colorado, Illinois, and Ohio, the New Teacher Project found that tenured teachers were observed for an average of only 75 minutes during the school year. This is not because principals are devoting most of their attention to novice teachers: The survey found that probationary teachers were
evaluated for an average of 81 minutes in a school year. Clearly, a teacher’s job is far too complex for his performance to be adequately assessed by an hour or so of observation once every year—or every few years.

Exacerbating the problem is the fact that the current system discourages principals from accurately rating even their worst teachers on official evaluations. Research shows that principals can in fact tell which teachers are the best and worst: A recent study found that principals who were asked to rate their teachers on a scale from 1 (inadequate) to 10 (exceptional) were very good at identifying the teachers marked by empirical measures as the best- and worst-performing. So if principals can distinguish between good and bad teachers, why do they routinely give them all high ratings?

One reason is that flunking a teacher on his evaluation produces major headaches without any tangible rewards. In many school systems, a teacher who is identified as exhibiting below-standard performance may file a grievance according to the rules of the union contract. Many school systems have explicit policies requiring principals to fill out extensive paperwork and allowing teachers several appeals in cases of poor ratings. Avoiding paperwork is not a good excuse for an administrator to keep a bad teacher in the classroom. Then again, why should a principal go through the hassle that results from giving a teacher a poor rating if today’s evaluations have no meaningful consequences?

And when it comes to those teachers who do have tenure—virtually every teacher with more than three to five years of service in the school system—the process is even more futile. Technically, tenure entitles a teacher only to due process before being fired, which may sound harmless. The problem, however, is that the due process required to fire a tenured teacher is so burdensome and so unlikely to succeed that few principals attempt to remove even their worst teachers. In New York City, for example, only 45 tenured teachers were fired for any reason during the 2008-09 and 2009-10 school years. On average, only about two tenured teachers in the entire state of Illinois are fired for poor classroom performance each year. During the 18-year period between 1987 and 2005, 93% of districts in Illinois did not attempt to fire a single tenured teacher. Between 1995 and 2005, the struggling Chicago public-school system formally remediated a total of only 231 teachers—a tiny fraction of the city’s 27,039 public-school teachers. In California, the situation is much the same: Los Angeles fires, on average, roughly one teacher out of every
1,000 per year; Long Beach fires about six teachers out of every 1,000; and San Diego terminates some two teachers out of every 1,000.

Firing a tenured teacher is so difficult, in fact, that the process is often reserved only for those teachers who pose a physical threat to students. Rarely is poor classroom performance cited as a reason for dismissal; for instance, in Los Angeles between 1994 and 2009, 80% of the dismissals upheld in the district did not list poor teacher performance as a factor. Only eight out of the 45 terminations in New York City in 2008-09 and 2009-10 were related to teacher effectiveness, and six of those cases included other charges such as insubordination or misconduct.

Given that removing an ineffective teacher is almost impossible, many principals decide that they might as well avoid the significant red tape it involves. A survey of principals in Chicago by the New Teacher Project found that 30% of principals who admitted to inflating their evaluations did so in part because giving a low rating wasn’t worth the cost: The bad teacher wouldn’t be dismissed anyway. Indeed, teacher-performance ratings are so meaningless that they aren’t even used when financial constraints force school districts to let some teachers go. When budget cuts require districts to lay teachers off, most states require that the terminations be decided explicitly according to seniority (with some exceptions for grade-level and subject-matter needs). Even in these cases in which some teachers must be fired, school districts are not able to remove the 2% of teachers whose performance is so poor as to receive a bad rating under the current rubber-stamp system.

In short, the “unsatisfactory” rating is a distinction embarrassing enough for teachers to fight with powerful weapons, but far too meaningless for administrators to defend. Given the many disincentives, it’s a wonder that any teachers are rated “unsatisfactory” at all.

THE QUANTITATIVE APPROACH

Recognizing the ludicrous results of the current evaluation system, many researchers and policymakers have called for using a more data-driven approach to assessing the performance of individual teachers. Though imperfect, these quantitative measures of teacher quality can dramatically improve today’s rubber-stamp evaluation system.

This turn to quantitative assessments is part of a broader shift that, over the past two decades, has changed how researchers and policymakers think about public schools. In the past, education research was dominated by
those who believed that schools were too complex and heterogeneous for empirical evaluations of entire systems to hold meaning. Rather than follow the scientific revolutions taking place within other disciplines, researchers in education followed what sociologist Thomas Cook has described as an R&D model based on various forms of management consulting. This “sciencephobia” (Cook’s phrase) in education research during the 1980s and ’90s left a major void in our understanding of the effectiveness of policies operating within public schools and created a culture in which teachers and school systems were suspicious of quantitative measurement.

Today, however, quantitative researchers—particularly economists—are at the cutting edge of research in education policy. They now hold significant positions within the U.S. Department of Education and are frequently hired to faculty positions at prestigious education colleges. Such quantitative researchers were previously uninterested in education largely because there were no meaningful data to consider. But they were drawn into the education discussion when the boom in standardized testing produced extensive data on student academic achievement. The economists in particular treated education as a production process: Organizations mix inputs (such as curriculum, class size, and teacher quality) in order to produce an output (student proficiency). This worldview values scientific procedures and quantitative measurement.

As the most important school-based input during the learning process, teachers have received considerable attention from this new crop of researchers. If the goal is to improve student learning, it is only logical that differences in teacher quality should be identified and addressed. And if quantitative measures exist to identify those differences, so much the better.

Economists and statisticians have developed just such a quantitative measure of teacher quality: “value added” assessment. This approach uses a statistical model to estimate the teacher’s independent contribution to student learning, as measured by standardized-test scores. Value-added measurement (VAM) generally relies on a common statistical technique known as multiple regression. In this case, the regression analysis estimates how differences in observed characteristics about a student, his school, and his teacher are related to changes in his math or reading test scores in a particular year.

Value-added analysis predicts how well a student should perform in a given year based on a series of observable factors that are related to
his academic achievement, but are beyond a teacher’s control—factors such as race, gender, and family income. The analysis then compares for each teacher the estimate of how well his students were expected to perform at the end of the school year given the characteristics they brought into the classroom with their actual test scores in the spring. The teacher’s VAM score represents his performance in standard-deviation units relative to the average teacher (the mean VAM score) in the school system; the mean score is set at zero. If a teacher’s students tend to outperform expectations on average, then the teacher’s VAM score will come back as positive; if students perform worse than expected given their characteristics, the teacher will receive a negative VAM score.

The value-added method requires access to data that follow the test scores of individual students over time and match the students to their teachers. A decade ago, such data simply did not exist. But thanks to the ubiquity of standardized testing imposed by the No Child Left Behind Act, this information should be available in all states and school districts.

The analysis itself is carried out by a data center within the governing department of education office. In this way, the use of VAM represents an important consolidation of the administrative structure: Under the current system, the on-site principal has nearly complete control over his teachers’ evaluations, but a VAM-based system takes a good deal of that discretionary power out of the principal’s hands.

The adoption of value-added measurement is still in the very early stages, though the Washington, D.C., school system is already using value-added measurement to evaluate teachers and make employment decisions. The district’s IMPACT evaluation tool, adopted under former D.C. schools chancellor and aggressive reformer Michelle Rhee, uses both frequent classroom observations and value-added analysis to identify ineffective teachers. Washington’s approach is already rooting out underperformers: The district dismissed 98 teachers this summer based on poor evaluation results, bringing the total number dismissed for poor performance to nearly 400 since 2009.

Washington is the school system furthest along in the process of embracing value-added measurement, but other reform-minded school systems across the nation have recently begun their own experiments. For instance, teachers in Colorado, Nevada, and Tennessee now revert back to probationary (i.e., non-tenured) status if, for two years in a row, they
receive poor performance ratings based in part on VAM scores. New York, New Jersey, and Connecticut have also recently passed legislation to use value-added measurements as an important part of teacher evaluations.

Since the use of VAM is so recent, there are not yet enough data to measure the effects of policies that base employment decisions on value-added analysis. By looking back in time, however, we can consider the likely results of VAM-based policies. For instance, in a recent report for the Manhattan Institute, I examined the measured effectiveness of Florida teachers in 2009. This research shows that, had a VAM-based tenure system been in place in 2007, the teachers who would have been removed were far less effective in 2009 than were teachers who would have been retained. This evidence suggests that more widespread adoption of value-added policies has significant potential to accurately identify and remove underperforming policies from the nation’s schools.

The objectivity inherent in quantitative analysis is perhaps one of the most attractive attributes of value-added assessment. Unlike principals who can be swayed by the thought of having to deal with a disgruntled teacher when considering whether to issue a poor evaluation, the value-added analysis provides its measure of the teacher’s performance completely irrespective of the possible consequences. The procedure used to evaluate teachers is determined well before the school year begins; the computer used to run the model doesn’t care that the teacher is friends with the principal or is well liked by the students. Nor does the analysis consider that the teacher is a whistleblower within the school. All that matters is whether the students in the teacher’s classroom appear to be making academic improvement.

Further, by considering the entirety of the student’s measured academic growth during the school year, value-added analysis provides a much broader view of the teacher’s effectiveness than do brief, infrequent observations. A teacher surely has good and bad days, but value-added assessment evaluates the end product of a long and challenging school year. As the primary concern in judging teachers is determining whether, at the end of the year, students have made academic progress, value-added analysis helps us to focus on what matters most.

AN IMPERFECT TOOL

To be sure, value-added analysis is not a complete or perfect measure of teacher performance. It should not—and, in many cases, cannot—be
used in isolation to evaluate a teacher’s performance when making employment decisions.

The most important problem with using value-added analysis is that there are many teachers whose performance simply cannot be assessed by standardized tests. Because the No Child Left Behind law required states to test students in grades three through eight, many states test only in these grades—and value-added scores can be calculated only when students are tested. Even in cases in which value-added measures can be calculated, these analyses tell us only part of what we want to know about a teacher’s performance. Value-added analysis can tell us in broad terms whether a teacher is performing well in the classroom, but is far too blunt a tool to provide us with useful information about what exactly the teacher is doing right or wrong.

Moreover, because the standardized tests mandated by NCLB examine only reading and math skills, value-added assessments tell us how well a teacher’s students are performing only in the basic subjects of reading and math. Emphasis on these subjects is warranted because they are foundational: If a fourth-grade student is not improving as a reader during the course of the school year, he probably isn’t learning much of anything else, either. And far too many students lack even basic proficiency in these bedrock subjects. Nevertheless, we do ask more from our teachers than that they produce math and reading gains, and many teachers—particularly in the middle- and high-school grades—do not teach reading or math at all.

Perhaps the most important common criticism of value-added assessment is that, as a statistical calculation, it is necessarily influenced by random error. Some teachers are worried that their value-added scores will not accurately measure their true contributions to their students’ learning. The concern is justified: No statistical model can account for every factor that produces a student’s test score at the end of the year.

For instance, one common complaint about the value-added method is that students walk into a teacher’s classroom with varying levels of skill. Some worry that those teachers who are assigned the best students will receive higher value-added scores by default—and non-random classroom assignments are indeed an issue that researchers continue to grapple with. In this case, however, the concern can be greatly alleviated: Value-added models control for the ability students bring into the classroom by accounting for their test scores at the end of the previous
school year. This procedure is common in education-policy research, and ensures that the value-added model considers factors most related to a child’s academic growth — the best indicator of a particular teacher’s input — rather than his overall academic achievement.

More worrisome are the random factors contributing to test scores for which value-added analysis cannot control. Perhaps a paragraph on the reading exam happened to address a topic that the student knew a lot about; maybe there was a dog barking outside during the test; perhaps a few kids in the teacher’s class happened to guess a group of questions right or wrong. The unambiguous consequence of random error — what researchers refer to as “noise” — is that it will always be the case that some average teachers will appear to be ineffective when they should not, and that some bad teachers will score better than they should.

We can limit the influence of such noise in teacher evaluations by incorporating multiple years into the value-added calculation. Furthermore, we can statistically calculate how much error appears to be included in the analysis and take the precision of the estimates into account.

Still, the influence of random error cannot be entirely eliminated. And it is this influence that leads teachers to argue so strongly against using standardized-test scores to evaluate their performance. Teachers are concerned that they will be among the unlucky whose value-added scores dramatically underestimate their effectiveness. But they are not the only ones who will be harmed by mistaken evaluations: Students, too, will suffer if good teachers are misclassified and are incorrectly forced to leave the classroom.

Clearly, value-added analysis is not perfect; no evaluation system is. When considering whether to use value-added measures of teacher quality, however, we must think about whether the error they introduce is outweighed by the accurate information they provide. And the debate over how to make that judgment highlights one of the most serious problems with today’s public-school system: When the interests of students and their teachers collide, the current system favors the teachers.

Weighing Interests

Everyone wants a fair evaluation system. The real question is, “Fair for whom?” The decision whether to use value-added assessments to evaluate teachers requires us to weigh the interests of teachers against those of their students.
The entities charged with protecting teacher interests—their unions—commonly claim that no such division of interests exists. According to American Federation of Teachers president Randi Weingarten, “teachers and kids are totally and completely interconnected. Teachers advocate for things that they need that are in the interest of kids and vice versa. Trying to divide teachers from kids is only a way of hurting what parents and students need to create opportunity in this country.”

It is a mistake to accept at face value such claims that the interests of teachers and students always align. The division of interests is a natural one that occurs in just about any employment relationship. Though we prefer to think of teachers as special, they are, after all, employees working within a large bureaucratic system that ought to value student achievement over everything else.

As the employer, the public-school system is required to determine its overall objective. The unions want the system to prioritize the needs of teachers: to protect jobs and provide teachers with substantial leeway to practice their craft as they see fit. Parents and students, on the other hand, should prefer a system that employs the highest quality public-school teachers and produces the least amount of variation in performance. The debate over using value-added analysis to evaluate teachers illustrates how these interests can collide.

Individual teachers primarily want their own evaluation scores to be “fair” — which for teachers means not receiving bad scores if they are actually performing well in the classroom. And because value-added measures would surely lead to some average teachers receiving poor evaluations, teachers would rather not see them used.

The problem is that the current system is even more error-prone than value-added assessment is — though the error is of a different kind, and one that works to teachers’ advantage. By uniformly classifying just about all teachers as effective, the current system produces many “false positives”: bad teachers who are incorrectly labeled effective. Because it is very unlikely to underestimate their performance, teachers prefer the current evaluation system — even if it keeps some bad teachers in the classroom.

For students, however, this wide variation in teacher quality means that they suffer when bad teachers are not identified and removed. And if we take the position that student achievement — rather than teacher job protection — is the primary purpose of the public-school system,
then the question to answer is whether the error introduced by value-added analysis is so large that its results provide no useful information for differentiating between the system’s best and worst teachers. In other words, it is crucial to determine whether, on balance, the positive effect of accurately identifying and removing bad teachers outweighs the negative effect of misclassifying some average and good teachers.

This is a very hard and important test, and yet it is one that value-added analysis passes. Though researchers are still working on improvements to the method, even simple value-added assessments of teacher quality have been demonstrated to provide useful information about a teacher’s performance. For instance, research by economists Dan Goldhaber and Michael Hansen found that value-added measures of a teacher’s performance in his first three years on the job—usually the years before tenure—are far more accurate predictors of how well that teacher’s students will perform in later years than are conventional measures of teacher quality (such as years of experience and the obtainment of advanced degrees). Goldhaber and Hansen further showed that a policy of removing low-performing teachers based on these value-added scores from the first three years in the classroom would substantially improve teacher quality by increasing the number of bad teachers who exit the system early in their careers.

If the school system’s primary purpose is to serve the interests of teachers—ensuring that no average or good teachers are mistakenly marked as ineffective and removed from the classroom—it should not use value-added measures. If, however, the public-school system is designed to serve students, using value-added analysis to improve teacher quality is clearly the wise course of action.

**Minimizing Error**

Deciding to use value-added assessments as a part of teacher evaluations does not mean that they should be used blindly. As discussed above, value-added analysis does have some very real limitations—and policymakers should use commonsense strategies in order to minimize them.

Because of these limitations, no serious researcher or policymaker proposes that value-added analysis be used in isolation to evaluate teachers. Rather, the hope is to use test-score analysis in combination with assessments from rigorous classroom observations to get a more complete picture of a teacher’s performance than is provided by the current
evaluation system. After all, Goldhaber and Hansen showed that a policy of removing ineffective teachers based on their value-added scores alone would substantially improve teacher quality in public schools. An evaluation system that used value-added as one of many tools to evaluate teachers would do an even better job of identifying bad teachers while also protecting good teachers from the error inherent in value-added analysis.

Value-added measures can be further improved upon by increasing the number of students and years of student test scores included in the analysis in order to arrive at greater precision. When value-added models are used to help inform employment decisions—setting a teacher’s pay, or determining whether to remove him from the classroom or grant him tenure—the teacher’s rating should be based on at least three years of classroom data. This measurement period allows for enough data to minimize error without imposing an undue burden on the school system. And since most school systems decide whether to offer a teacher tenure after the third year, this restriction should be relatively easy for school systems to adopt.

School systems should also be wary of policies that use value-added measures to make fine distinctions between the effectiveness of individual teachers. Value-added scores are useful for differentiating between very good and very bad teachers, but the error involved in their calculation means that value-added measures are not particularly good at rating teachers in the middle of the pack. This would be analogous to weighing 1,000 people and ranking them in order from heaviest to lightest: Since body weights vary slightly day by day, it would be impossible to determine whether the 503rd heaviest person was in fact heavier than the 504th. It would be easy, however, to distinguish between the thin and the obese. This is the sort of broad classification for which value-added analysis should be used.

AN IMPORTANT MEASUREMENT

Policymakers on both sides of the aisle have either recently realized the truth or have finally become emboldened to state the obvious: There are some bad teachers in our schools. This conclusion has been supported by a robust body of research, which consistently finds that the differences among bad, average, and great teachers are substantial. For example, a study by Stanford University economist Eric Hanushek using data from Texas public schools found that, on average, students assigned
to teachers whose quality is at the 25th percentile tend to make about half of a grade level’s worth of reading achievement gains during the course of a school year. In contrast, if those same students are assigned to a teacher at the 75th percentile, they tend to gain about one and a half grade levels in reading during the year. The difference between a child’s being assigned to one teacher or another can thus reasonably be as much as a grade level’s worth of learning during the school year. Studies have found similar results using data from school systems in several other states.

And the consequences of teacher assignment can last far longer than just one school year. A student lucky enough to be assigned to great teachers several years in a row will have an enormous advantage over other children, while unlucky students who are assigned to bad teachers a few years in a row are likely to fall behind their peers and stand little chance of catching up. Recent research shows that the effects of teacher assignments during the course of a student’s career are sustained throughout the child’s life.

Given what is at stake, it seems obvious that the nation’s public-school system should want to do everything in its power to make sure that children are instructed by the best teachers possible — using pay, tenure, and other incentives to reward quality, not simply longevity. But such reforms are not possible without a reliable, empirical measure of teacher quality — one rigorous and objective enough to withstand the opposition of teachers concerned primarily with their own comfort and job security.

Fortunately, value-added analysis offers just such a measure. While not perfect, the value-added approach does provide important information that is missed by the current system and can be used to identify our best- and worst-performing public-school teachers. And reforming that system — which is now so stacked in the teachers’ favor as to be completely ineffective — is the first step toward the ultimate aim of ensuring that all public-school students receive a quality education.