Value-added measures of teachers: Research and policy

Value-added models in education are used to attempt to measure the contributions to student achievement of individual teachers. Test scores for a particular teacher's students are compared to those of the same students in the previous year, as well as to those of students with other teachers in the same grade, in an effort to isolate the contribution of the given teacher. Advocates of these methods argue that these measures provide objective information that can be used to improve instruction, while critics counter that their validity as an indicator of teacher quality is still in question. School districts from Washington D.C. to Los Angeles have started to use value-added measures, and some teachers' ratings have been made publicly available, including recently in the Los Angeles Times *and the* New York Times.

On June 27, 2012, as part of IRP's annual Summer Research Workshop, three researchers participated in a roundtable discussion of teacher value-added measures. Raj Chetty presented outcomes from a long-term study of the effects of teachers on students from elementary school through early adulthood. Jesse Rothstein explored the potential dangers of using value-added measures to make teacher personnel decisions. Finally, Eric Hanushek looked at policy implications from a different perspective, exploring why and how value-added measures can be used most effectively. This set of articles summarizes the three presentations.

The long-term effects of teachers

There is considerable debate about the best way to measure and improve teacher quality. One method is to rate teachers based on their students' test score gains, known as the "value-added" approach.[1] School districts have begun to use these measures to make personnel decisions about teachers. For example, District of Columbia Public Schools lay teachers off or offer them bonuses using a teacher-performance assessment system that puts 50 percent weight on value-added scores. This article describes an extensive study designed to estimate the effects of teachers on student outcomes through early adulthood.[2]

Potential issues with value-added models

The debate about using teacher value-added models stems primarily from three issues. First is concern about the potential for bias in value-added estimates; do differences in test-score gains across teachers capture causal effects, or are they instead driven by student sorting? Second is lack of evidence on the long-term effects of teachers; do teachers who raise test scores also improve students' long-term outcomes, or are they simply better at teaching to the test? The third issue is measurement error; are estimates based on only a few years of data accurate enough to be used for policy decisions?

Assessing teachers over the long-term

The study done by Raj Chetty, John Friedman, and Jonah Rockoff addresses all of the above issues by using data on a million children, from childhood through early adulthood.[3] The researchers developed new quasi-experimental tests to assess bias in value-added estimates. They look at whether those who had high value-added teachers as children have better outcomes in adulthood. Finally, they assess the monetary gains to be made by selecting teachers with higher estimated value-added scores, given observed measurement error.

The estimated teacher effects include both direct and indirect outcomes. For example, having a good teacher in the fourth grade can improve a student's labor market outcomes in adulthood directly, but it can also have indirect effects, if students with good teachers in the fourth grade receive better teachers in subsequent grades.[4]

Are value-added estimates unbiased?

Since students are not assigned to teachers randomly, it is necessary to adjust for the composition of students assigned to a classroom; the standard approach is to control for prior year variables. Recent studies have reached conflicting conclusions about whether this approach is sufficient for obtaining consistent estimates of teacher effects.[5]

In this study, the researchers evaluate whether or not value-added estimates are biased by (1) testing for selection on observable characteristics, and (2) using quasi-experimental methods that make use of natural teacher turnover. In order to test for selection on observable characteristics, they look at whether parent characteristics are correlated with teacher value-added scores, and find no relationship. For example, the children of wealthier parents are no more likely to get higher value-added teachers. In order to test for selection on unobservable characteristics, they looked at changes in students' scores in the year before and after a switch in teachers due to teacher turnover. Again, they find no evidence of selection. Therefore, they conclude that their value-added measures provide unbiased estimates of teachers' causal effects on student test scores.

This article summarizes the presentation given by Raj Chetty.

Raj Chetty is Professor of Economics at Harvard University.

Effects on outcomes in adulthood

Next, the researchers assessed whether teachers who raise test scores also improve their students' outcomes in adulthood. They analyze the effects of teachers on three sets of outcomes; college attendance, earnings, and other indicators such as teenage birth rates.

Being assigned to a higher value-added teacher in a single grade significantly raises a student's likelihood of attending college. A one standard deviation increase in the value added of a teacher appears to increase the probability of that student attending college by age 20 by 1.25 percent. Students with higher value-added teachers are also more likely to attend a better college, as measured by projected average earnings at age 30.

Having a higher value-added teacher has a clear statistically significant effect on earnings. An increase in teacher value added of one standard deviation increases annual earnings at age 28 by $182. The lifetime financial value of having a teacher one standard deviation higher is approximately $4,600 per grade.[6]

Having a teacher one standard deviation higher in value added in a single year from grades 4 through 8 reduces the probability of a teen birth for female students by 1.25 percent.[7] Students with higher value-added teachers are also more likely to live in higher socioeconomic status neighborhoods as adults.

Measurement error and policy relevance

Any evaluation of teachers based on value-added measures must rely on only a few years of classroom data. This limited amount of data adds uncertainty to value-added estimates, thus potentially reducing their utility for performance evaluation. In order to evaluate how much the utility is reduced, it is necessary to look at a policy example. Thus, the researchers analyze the effects of retaining or firing teachers on the basis of their value-added scores.

On average, replacing a teacher in the bottom 5 percent with an average teacher for one year raises a child's cumulative lifetime income by $50,000. For a class of average size (28 students), the cumulative lifetime income gains from a high value-added teacher exceed $1.4 million. This is equivalent to $267,000 in present value at age 12, discounting future earnings gains at a 5 percent interest rate. Of course, data limitations do not allow certainty about which teachers are in the bottom 5 percent. In estimating the gains of deselecting teachers based on their *estimated* value added, there is still a substantial potential lifetime earnings gain. The present value of earnings gain from deselecting teachers below the fifth percentile increases with the number of classes observed per teacher. While the gain with even ten observed classes is still below the $267,000 value achievable with perfect knowledge of teacher rank, with even three or four observed classes, the lifetime gain is still around $200,000.

Policy implications

While the Chetty and colleagues study supports the idea that existing value-added measures are useful in identifying long-term effects of teachers, this conclusion alone is not sufficient to assess value added as a policy tool, for at least two reasons. First, it is necessary to weigh any potential gains against the cost of firing teachers. The researchers' calculations suggest that the financial benefits of such a policy far outweigh the costs. A second and more serious concern not addressed in this study is potential negative behavioral responses to testing when the stakes are so high, such as teaching to the test or even cheating.[8] It is possible that such responses, if sufficiently large, could completely counter any policy gains.

Parents should be interested in knowing the value added of their child's teacher, whether or not that information is useful as a policy tool. This analysis shows that high value-added teachers improve students' achievement and long-term outcomes. The most important lesson of this study is that finding policies to raise the quality of teaching—whether through the use of value-added measures, or through other tools such as salary structure changes or teacher training—is likely to have substantial economic and social benefits in the long run.■

---

[1]See, for example, E. A. Hanushek, "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," *American Economic Review Papers and Proceedings* 61, No. 2 (1971): 280–88; and R. J. Murnane, *The Impact of School Resources on the Learning of Inner City Children* (Cambridge, MA: Ballinger, 1975).

[2]The study summarized here is described in detail in R. Chetty, J. N. Friedman, and J. E. Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," NBER Working Paper No. 17699, National Bureau of Economic Research, 2011. http://www.nber.org/papers/w17699.

[3]The data link two large databases: student records from a large school district, including teachers, class assignment, and test scores from 1991–2009 for students in grades 3–8; and U.S. income tax records including both student outcomes (such as earnings, college, and teenage birth) and parent characteristics (such as income, savings, home ownership, mother's age at childbirth, and marital status).

[4]See, for example, T. J. Kane and D. O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607, National Bureau of Economic Research, 2008.

[5]Kane and Staiger, "Estimating Teacher Impacts on Student Achievement;" and J. Rothstein, "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125, No. 1 (2010): 175–214.

[6]A one standard deviation increase in teacher value added in a single grade results in increased earnings at age 28 of $182, which is 0.9 percent of mean earnings in the regression sample. The researchers assume that the percentage gain in earnings remains constant at 0.9 percent over the lifecycle, and that earnings are discounted at a 3 percent real rate (that is, a 5 percent discount rate with 2 percent wage growth) back to age 12, the mean age in the sample. Under these assumptions, the mean present value of lifetime earnings at age 12 in the U.S. population is approximately $522,000. Thus, the financial value of having a one standard deviation higher value added teacher is 0.9 percent of $522,000, or approximately $4,600 per grade.

[7]The "teenage birth" measure indicates whether a tax return was filed that included a dependent born while the mother was a teenager.

[8]G. Barlevy and D. Neal, "Pay for Percentile," *American Economic Review* 102, No. 5 (2012): 1805–1831.

Effects of value-added policies

Jesse Rothstein

Jesse Rothstein is Associate Professor of Public Policy and Economics at the University of California, Berkeley, and Research Associate at the National Bureau of Economic Research.

It is important to distinguish between two topics that have often been mixed together: (1) the properties of value-added models and (2) the effects of value-added-based policies. Most research to date has focused on the first, nearly always in low-stakes settings, and many researchers and others have drawn strong policy conclusions from that research. But at this point we know very little about the effects of policies that would use value-added scores to make decisions about teachers. That should be the focus going forward. What really matters is not the effect of individual teachers, which is what most research estimates, but the effect of a policy.[1]

What do we know about the properties of value-added models?

A considerable amount of research has been devoted to developing models to estimate the contributions of individual teachers to student achievement. It is important to note that the things we have learned about the properties of value-added models nearly always come from low-stakes settings; that is, the value-added calculations for individual teachers have not generally been used to make decisions about teacher retention or bonuses. While much has been learned, there are still many unanswered questions. I'll review here what I see as a few of the most important outstanding issues.

Value-added measures have been shown to have substantial measurement error, although averaging a few years of data does help. The measures are also sensitive to student assignments. We know that assignment of students to teachers is not random, but it remains an open question whether assignment practices introduce large biases in individual teachers' evaluations. In a paper a few years ago, I showed that the available data were consistent with substantial biases or with essentially no bias.[2] Important papers by Kane and Staiger and Chetty, Friedman, and Rockoff have narrowed the plausible range somewhat.[3] However, both the Kane-Staiger and the Chetty and colleagues estimates have had very wide confidence intervals, so we still do not know the importance of biases due to student assignments.

The Chetty and colleagues study revealed an important fact that has not been incorporated into most thinking about value-added models to date. Specifically, they found that teacher effectiveness changes over time: Some teachers are ineffective at first but improve as they age, while others start better and then burn out. Under a policy that uses value-added measures to fire poor teachers and reward good ones, some teachers fired early for poor student achievement would have improved over time, while some teachers who receive early raises will continue to receive them even if the quality of their teaching declines. Both modeling and policy calculations will need to change to accommodate this fact, which could have important implications for the kinds of cost-benefit analyses that have been done to date (including in the Chetty and colleagues study).

Another unresolved issue is the choice of value-added specifications. Each author tends to focus on his or her preferred value-added model, and it isn't clear how much it matters. An important aspect of this issue is the distinction between within- and between-school comparisons. Researchers typically focus on within-school comparisons, including fixed effects to absorb any between-school differences. There is good reason for this, as while it is barely possible that students are randomly assigned to teachers within schools, it is clearly not the case that students or teachers are randomly assigned to schools. Proposed policy applications of value added, however, will need to make both within- and between-school comparisons. We do not have a consensus about how to do this, nor much evidence about how much it matters.

Finally, Chetty and colleagues show that teacher value added is predictive of students' future wages. However, the strength of this correlation is unknown. If we could measure teachers' impacts on student wages, would we find that their test score impacts (as measured by value added) were good proxies for them? We don't know. We also know very little about the interactions across grades; if a student has a high value-added teacher two years in a row, how should the values be combined to calculate the joint effect? Researchers typically treat the effects as additive, but there's no evidence for this and a good deal of reason to think it is incorrect.

What do we know about the effects of value-added-based policies?

Much less is known about the effects of value-added-based policies than about how to measure the contributions of teachers to student achievement. It is difficult to find studies that show that offering significant bonuses to high value-added teachers in the United States produces significant effects, and some of the highest-quality studies of the issue find no evidence of such effects.[4] And there is essentially no evidence on the effects of policies that use value added for deselection (i.e., firing) or for professional feedback.

A study by Carrell and West provides a cautionary tale: adjunct Air Force academy professors, whose continued employment depends on their measured teaching performance,

outscored their regular faculty peers on value-added-type measures based on end-of-year tests, but their students performed poorly in follow-on classes.[5] These results suggest the potential for teacher responses that improve the teacher value-added measure without improving future student outcomes.

What would we expect to happen if teacher policy is based on value added?

In the absence of extensive evidence on the effects of value-added policies, we can still make an educated guess using a long-standing principle in the education field known as Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[6] Campbell also states that "achievement tests may well be valuable indicators of general school achievement under conditions of normal teaching aimed at general competence. But when test scores become the goal of the teaching process, they both lose their value as indicators of educational status and distort the educational process in undesirable ways."

Thus, if teachers are told that their jobs depend on having a high value added, we should expect that value added will be high, but also worry that that might come at the cost of teachers not doing things that we would really like them to do, but that are not directly related to value-added scores. For example, since teachers are evaluated based on math and reading scores, they might spend less time teaching subjects that are not covered in achievement tests, such as history. Even within a tested subject, teachers might spend more time on topics that are covered on the test such as analogies, and less on topics that are not such as composition. There is anecdotal evidence that some teachers are unwilling to teach students whom they believe will not improve their value-added score. Teachers might also focus more on short-term learning (such as drills on multiple-choice questions) that is likely to be reflected in test scores, rather than on long-term learning that will serve students better after the tests are done. The Air Force Academy results mentioned above appear to indicate that these kinds of responses can be important.

David Figlio has done a lot of work looking at the unintended effects of school accountability, ranging from suspension of students who are expected to do poorly, to changing the food offered in the cafeteria on test day.[7] There are a great deal of factors that may affect test scores without affecting learning, and this may not be how we want our school resources to be used. We do not currently have a sense of how large these distortions would be, and thus how much they would undermine a policy that was based on value-added measures, but it does appear possible that they could completely negate the effects of a teacher policy based on value added.

Personnel economists have spent years studying incentive compensation, and there are lessons from that field that clearly apply to education. When a task is multidimensional, as teaching certainly is, and when a performance measure is subject to influence, as I believe value added is, it is important to ensure that the stakes are low for a particular measure; that multiple measures be used; that human discretion be part of the process; and finally that the process for helping employees improve be separate from the process through which personnel decisions are made. I believe that describes a viable teacher personnel policy, albeit one that looks quite different from what many districts are implementing. What would it take to implement this kind of policy? First there must be lots of administrators, all highly trained and carefully selected. It seems unreasonable for a single principal to be solely responsible for 40 teachers, accompanying staff, and all other aspects of a given school. While the consulting-world standard of one manager for every five workers is not likely to occur in the world of education, perhaps one administrator for every ten teachers is achievable? It is important that the administrator be capable—there is no reason to think that principal quality is any less important than is teacher quality. We should also be thinking at least as much about the best ways to develop and improve staff, rather than firing them. Finally, there should be an incentive pay component, but stakes need to be relatively low so as not to cause too much distortion of outcomes.■

---

[1]This point was made by D. B. Rubin, E. A. Stuart, and E. L. Zanutto, "A Potential Outcomes View of Value-Added Assessment in Education," *Journal of Educational and Behavioral Statistics*, 29, No. 1 (Spring 2004): 103–116.

[2]See J. Rothstein, "Student Sorting and Bias in Value Added Estimation: Selection on Observables and Unobservables." *Education Finance and Policy*, 4, No. 4 (Fall 2009): 537–571.

[3]See: T. J. Kane and D. O. Staiger, "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation." Working Paper No. 14607, National Bureau of Economic Research, 2008; and R. Chetty, J. N. Friedman, and J. E. Rockoff, "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood," Working Paper No. 17699, National Bureau of Economic Research, 2011. In addition, the Gates Foundation's Measures of Effective Teaching (MET) Project recently released results of a large-scale experiment along the lines of that carried out earlier, on a smaller scale, by Kane and Staiger. Unfortunately, the experiment was plagued by high rates of noncompliance, which limited its ability to answer the question at hand. See: T. J. Kane, D. F. McCaffrey, T. Miller, and D. O. Staiger, *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment*, MET Project Research Paper, Bill and Melinda Gates Foundation, Seattle, WA, January 2013; and J. Rothstein and W. J. Mathis, Review of *Have We Identified Effective Teachers?* and *A Composite Estimator of Effective Teaching*: "Culminating Findings from the Measures of Effective Teaching Project," National Education Policy Center, Boulder, CO, January 31, 2013.

[4]See M. G. Springer, D. Ballou, L. Hamilton, V. Le, J. R. Lockwood, D. F. McCaffrey, M. Pepper, and B. M. Stecher, *Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching*, National Center on Performance Incentives at Vanderbilt University, Nashville, TN, 2010.

[5]S. E. Carrell and J. E. West, "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy*, 118, No. 3 (2010): 409–432.

[6]D. T. Campbell, "Assessing the Impact of Planned Social Change," *Evaluation and Program Planning*, 2, No. 1 (1979): 67–90.

[7]See, for example, D. N. Figlio and S. Loeb, "School Accountability," in *Handbook of Economics of Education*, Volume 3, eds. E. Hanushek, S. Machin, and L. Woessmann (The Netherlands: North-Holland, 2001).

Use of value added in teacher policy measures

Eric A. Hanushek

Eric A. Hanushek is Paul and Jean Hanna Senior Fellow at the Hoover Institution of Stanford University, and an IRP affiliate.

I would like to offer a different take on the policy issues related to value-added estimates than that provided in Jesse Rothstein's article. I believe that the primary value of these estimates is in illustrating how much difference there is between teachers. When the estimates are made in low-stakes situations where there is little incentive to teach to the test, estimates of the variance in teacher quality are very precise. In this article, I discuss the implications of the results of these types of studies, and then explore the implications for teacher policy. I believe that where Rothstein's argument falters is that there are not currently any school systems that make teacher personnel decisions solely on the basis of value-added estimates, nor am I aware of any current proposals for such a system. For example, in regard to the District of Columbia policy described by Raj Chetty, only 18 percent of teachers in the system have value-added scores available, so this information is clearly only a relatively small part of what goes into making firing and bonus decisions.

Magnitudes of effects

Estimates of the average standard deviation in gains in student achievement over one year attributable to higher value-added teachers within a given school range from 0.13 to 0.17. Any between-school differences in teacher effectiveness would need to be added on top of this. Although Chetty has already discussed some of the implications of these differences, I will very briefly offer my own calculations.[1] Estimates of the effect of test scores on earnings indicate that a standard-deviation increase in scores translates into a 13 to 20 percent annual increase in earnings. Figure 1 illustrates the effect on student lifetime income by class size and teacher effectiveness, allowing for some depreciation in scores over time. This figure shows the estimated marginal effect, compared to an average teacher, of having a teacher in various percentiles. Calculations for individual students are multiplied by class size. So, for example, the present value at the beginning of high school for a 75th percentile teacher with a class of 30 students is $430,000, while that for a 25th percentile teacher with the same class size is $425,000. These numbers appear large enough to suggest that, although there may be some error in particular teacher personnel policies, having no personnel policy at all cannot be the correct answer.

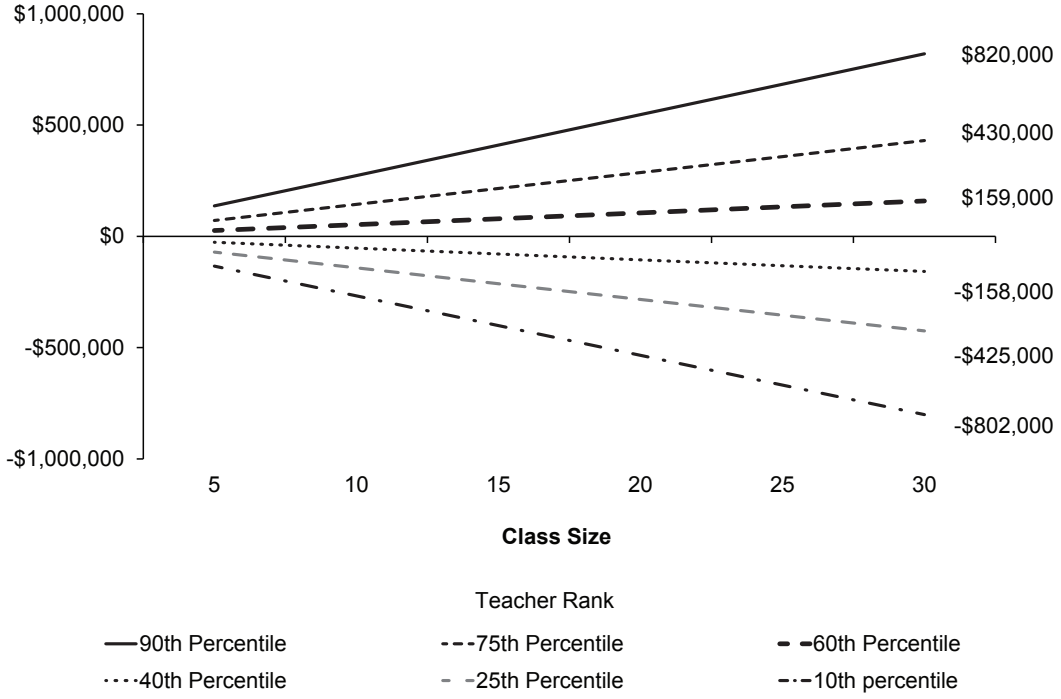

Figure 1. Effect on student lifetime incomes by class size and teacher effectiveness (compared to average teacher).

Source: Calculations by author relying on estimates of teacher quality using 0.2 standard deviations, and reflecting between-school calculations.

School districts have needed to lay off teachers in substantial numbers only quite recently, as a result of the recent recession. The standard policy for determining layoffs is to use teacher seniority. A recent simulation comparing this policy to one that used a measure of effectiveness found some differences between the two approaches.[2] Since seniority-based layoffs generally mean that those with lower salaries are more likely to lose their jobs, more layoffs are required to achieve a given budget reduction. In this simulation, a system based on value-added results in about 25 percent fewer layoffs than one based on seniority. In addition, the typical teacher laid off using a value-added system is less effective than the typical seniority-based layoff, by 26 percent of a standard deviation.

Another mental exercise is to imagine ranking all teachers in the United States based on effectiveness, and look at the performance gains that would result from deselecting some percentage of the lowest-ranked teachers, and replacing them with an average teacher. In this case, unlike the one-year effects that Rothstein estimated, I am looking at lifetime effects. I find that, depending on whether a high or low estimate of teacher effectiveness is used, a deselection rate of between 5 and 8 percent would result in achievement levels similar to that of Canada, a country that currently ranks 0.42 standard deviations above the United States. According to calculations I have made along with Ludger Woessman, such an increase in achievement is worth $72 trillion in GDP.[3] Larger estimates of the variation in teacher effectiveness result in even higher estimates. Although the precise value can certainly be argued, it is clear to me that the value of having policies based on teacher effectiveness is enormously higher than having no policy at all, and that policies based on teacher effectiveness in fact represent the future of the U.S. economy.

Use of value-added measures in teacher personnel policy

There has been a great deal of discussion about errors in estimating value added, and whether it is acceptable to, for example, have a 5 percent error rate in determining which teachers contribute the most to student achievement. I believe that the current state of having no policy translates to a 100 percent error rate, and that we should be striving not for perfection, but for a policy that improves teacher effectiveness overall.

Rothstein discussed some of the implications of making teacher-retention decisions based on imperfect value-added scores. If the rate of dismissal and replacement is somewhere between 5 and 8 percent, that translates to 2 to 3 individuals in a school of 30 teachers. I have found in all of my dealings with teachers, administrators, parents, and staff in numerous schools, that there is very little uncertainty about who the 2 to 3 least-effective teachers in any given school are. I believe that an evaluation process that allowed decisions based on this type of common knowledge would not necessarily need to depend on value-added data that might not be available in a timely manner, and that the evidence suggests that such a policy would likely result in substantial gains in student achievement.

As has been mentioned, both the *Los Angeles Times* and the *New York Times* have recently published teacher value-added scores for their respective school districts. This was extremely controversial, and the aftershocks are still being felt. I was one of the few researchers to support the idea of publishing value-added scores, not because I think that personnel policy should be done through newspapers, but because within a week of these publications, unions and school officials were meeting to discuss teacher-evaluation policy. This is an issue that had been on the agenda forever with no progress. It seems that providing a strict value-added ranking as one (extreme) option prompts people to develop better personnel systems that incorporate other teacher-evaluation tools, and this is exactly what is needed.

Issues and areas for further study

One could ask whether the currently available achievement tests are really up to the task of providing reliable value-added scores. I would say certainly not, and that value-added measures should never be the sole basis for personnel decisions. Rothstein also raised the possibility that value-added measures can become less reliable when used for consequential purposes. While this and the accompanying loss in reliability and validity is certainly possible, I believe such problems can be dealt with in feasible ways.

On the question of whether value-added measure can be used to rate principals, I agree with Rothstein that a parallel system is required. There are some indications that reliable value-added measures can be constructed. Preliminary estimates from work that I have been involved in suggest that principal quality is extremely important and that a one standard deviation increase in principal quality results in an increase of approximately 0.05 standard deviations in average student growth.[4] While this effect is much smaller than that seen for teachers within a given school, principals affect all students in a school, so an increase in principal quality will have effects much greater than a similar increase in the quality of a single teacher.■

---

[1]E. A. Hanushek, "The Economic Value of Higher Teacher Quality," *Economics of Education Review*, 30, No. 3 (June 2011): 466-479.

[2]D. Boyd, H. Lankford, S. Loeb, and J. Wyckoff, "Teacher Layoffs: An Empirical Illustration of Seniority versus Measures of Effectiveness," *Education Finance and Policy*, 6, No. 3 (Summer 2011): 439–454. The simulation was conducted using fourth- and fifth-grade math and language arts achievement scores for students in New York City public schools.

[3]E. A. Hanushek and L. Woessmann, *The high cost of low educational performance: The long-run economic impact of improving PISA outcomes* (Paris: Organisation for Economic Cooperation and Development, 2010).

[4]G. F. Branch, E. A. Hanushek, and S. G. Rivkin, "Estimating the Effect of Leaders on Public Sector Productivity: The Case of School Principals," NBER Working Paper No. 17803, National Bureau of Economic Research, 2012.