

The Promise of Value-Added Testing

by Jonathan Crane

Standardized testing, a key component of public education, is about to become even more important. The Elementary and Secondary Education Act (ESEA) reauthorization of 2002, better known as the No Child Left Behind Act, mandates that children in grades 3-8 be tested every year in both reading and math. It requires schools to ensure that all students meet minimum proficiency standards on these tests and to reduce the achievement gap between disadvantaged and minority children and other students. Schools that fail will lose federal money, and their students will have the option of going to other public schools. In short, schools, school districts, states, and indeed, the nation as a whole will now find themselves in the position, known so well to every student, of nervously awaiting their final exam grades each year.

This emphasis on accountability begs the question of whether the testing regimes now in place are up to the task at hand. Many people question the validity of any kind of standardized test. Clearly, no tests are perfect. Nevertheless, they are necessary. Without some sort of systematic evaluation process, there is simply no way to determine which approaches are working and which ones are not. In that environment, ineffective schools can persist without change *ad infinitum*, which is essentially the situation we've had in public education for many years. The ESEA seeks to change this. For it to work, we have to ensure that the tests we use are accurate, that they actually measure how much and what children are learning in school.

The ESEA allows schools to choose their own testing regimes. Every regime has its own strengths and weaknesses. However, virtually all of them suffer from one major flaw. They focus only on achievement *levels*, and ignore *changes*.

The testing regimes, and probably most people, assume that good schools are those with students who have high test scores and bad schools are those with students who have low test scores. But that isn't entirely true. Not all students enter a school at the same level. A school that admits students with low scores and raises them up to average is better than one that admits high-scoring students and merely keeps them at the same level. Ultimately, the best way to measure school quality, or teacher quality for that matter, is to determine how much they change their students' test scores each year. This is nothing more than the simplest and most direct way to measure how much students are learning. This approach is called "value-added," because it focuses on how much value a school or a teacher is adding to what students bring with them.

This does not imply that achievement levels are irrelevant. They are useful in determining what material students are ready for at any given point in time, and where they stand in relation to proficiency standards. But they do not tell the whole story.

Using a value-added testing regime in addition to the standard one could yield at least three important benefits:

- ▶ It would make ESEA more effective by providing a more accurate picture of which schools, school districts, and states are and are not making progress.
- ▶ It would generate objective measures of teacher quality that could be used to improve teaching.
- ▶ It would lend itself more readily to evaluating school reform programs.

Value-added testing is not perfect. It has a harder time distinguishing among students, teachers, and schools in the middle than among those at the extremes. It cannot necessarily distinguish between true learning and “teaching to the test.” And there is some disagreement among proponents of value-added testing about its power as a teaching tool. But standard testing regimes suffer from these same problems.

This paper will examine both the promise and the problems of value-added testing. It will show that there are still a number of issues that need to be addressed, so we should be cautious in its application. We should use it in combination with other approaches. Nevertheless, it is clear that the potential benefits of value-added testing are significant.

The Basics of the Value-Added Approach

The value-added approach to testing is very simple. It focuses on changes in test scores over time, rather than on a single test score at a given moment. The whole point of school is for children to learn and progress over time. Yet, most current testing regimes do not measure progress. Virtually every city newspaper publishes the average test scores of local schools at some point during the year, often accompanied by an article that highlights the best schools (those with the highest average scores), and the worst schools (those with the lowest). But are the children in the schools with the highest scores actually learning more than those in the schools with the lowest? Not if the high-scoring schools are just taking in children with the highest IQs from the most enriched environments and simply spitting them out at the same advanced levels years down the road.

A value-added system can use the exact same tests as a standard system. The only difference is that the value-added system keeps track of every student’s previous scores (whether or not they were at the same school), so that the change in their scores from year to year can be calculated. In this kind of system,

good schools are not necessarily the ones with the highest average test scores. They are the ones that show the *greatest average increases* in their students’ scores. They are the ones that are “adding the most value.”

Under their current systems, many states estimate and publish changes in schools’ average test scores at particular grade levels. But that is not the same as value-added analysis. They are making comparisons across cohorts. For example, they might show that this year’s seventh grade had higher reading scores than last year’s seventh grade at a particular school. That might mean the school is improving. But it might also mean that this year’s seventh grade was simply ahead of last year’s at the start. The essential element of the value-added approach is that it measures the progress of each and every child as he or she moves through school.

Because of the need to keep track of students across time and place, even if they change schools, value-added regimes require a more sophisticated data collection system than most current regimes. But this problem is by no means insurmountable. Tennessee has had a value-added system since 1991 and has been testing every student in grades 2-8 each year in math, reading, language arts, science, and social studies. (Testing high school students began in 1995.) Every year, each student’s results are calculated and compared to his or her scores from the previous year. Since 1993, school and district-wide averages have been reported to the public. The average score increases of all the students taught by each teacher are reported to administrators for use in evaluations and personnel decisions.

The Benefits of a Value-Added Testing System

Making ESEA More Effective

There are at least three major benefits of a value-added system. First, it would increase the effectiveness of ESEA, which is intended to encourage schools to improve by sanction-

ing those that are doing poorly. The hope is that these sanctions will force bad schools to abandon approaches that are not working and adopt approaches used by good schools. Over time, successful strategies will spread and unsuccessful ones will fall by the wayside. Obviously, the legislation will only work if the methods of identifying good and bad schools are accurate.

Under most current testing regimes, some schools that are actually doing a reasonably good job teaching large numbers of severely disadvantaged students may be sanctioned. Some poor schools that are merely shepherding top students may be imitated. And, perhaps worst of all, some excellent schools that are using effective strategies to help high-risk students progress may not be imitated, because their test score levels remain below average.

Value-added testing regimes would help us accurately identify those schools where students are learning the most and those where they are learning the least. Thus, we would be able to sanction those schools that are truly failing. And struggling schools would have an easier time identifying successful schools to model themselves after. This would be particularly important to schools with high proportions of disadvantaged students. It is unlikely that the techniques used in tony suburbs are directly applicable to impoverished neighborhoods of inner cities or struggling rural schools. These schools need to find schools that are dealing successfully with similar problems. They shouldn't be looking to imitate schools with the highest average test scores, but instead try to emulate schools in similar situations that are adding the most value.

This principle also holds true at the district and state levels. There is a good deal of debate about which school districts and states are really doing well and which ones are failing. A value-added testing regime would enable us to determine which districts and states are doing the best job of teaching their students. This, in turn, would help other districts and states identify those policies and practices from around the country that would do the most to help their schools meet ESEA mandates.

Raising Teacher Quality

The second benefit of value-added testing is that it can help improve teaching by providing an objective measure of teacher quality. In a value-added system, the average annual gain in test scores can be calculated for all the students of every teacher. Thus we can determine which teachers are adding the most value.

A fair amount of research has been done in this area already, and it turns out that this information is potentially very useful. There is a great deal of variation in the amount of value that individual teachers add for students. Simply put, good teachers help their students learn more than bad teachers.

In a study of second to fifth grade students in Tennessee between 1991 and 1995, William L. Sanders found, with all other things being equal, the top one-fifth of teachers raised their students' achievement test scores 39 percentile points more than teachers of the bottom one-fifth.¹ A percentile score indicates where a student stands relative to others on a particular test. So, for example, a test score in the 40th percentile means that a student scored higher than 40 percent and lower than 60 percent of the other test takers. A 39-percentile difference is enormous. It implies that a child who winds up in the bottom one-third of the distribution (say in the 30th percentile) after a year with a poor teacher, might have wound up in the top one-third (the 69th percentile) if they had been lucky enough to get an excellent teacher for that year. What's more, these large effects were found in all kinds of classrooms with all kinds of students. Good teachers raised test scores whether they taught in heterogeneous or homogeneous classrooms.² They helped struggling students as well as those who were excelling.

The effects of teacher quality also tended to persist and accumulate. In another study, Sanders and June C. Rivers found that good teachers raised their students' math test scores at least two or three years into the future. For example, suppose that two children from the same school start with equal scores

at the beginning of second grade. They are assigned to different teachers in the second grade, but they then have the same teachers in third, fourth, and fifth grades. All other things being equal, the child with the better second grade teacher will still have higher test scores at the end of fifth grade. That good second grade teacher will still have an effect three years later. Also, children who were lucky enough to have a succession of excellent teachers had a huge advantage over those who had a succession of poor ones. The difference between having three consecutive teachers in the top one-fifth of the quality distribution and three in the bottom one-fifth was *53 percentiles*.³ Effects of this magnitude are rarely seen in the social policy world. Education researchers Rivkin, Hanushek, and Kain found smaller effects, but limitations of their data forced them to make highly conservative “lower-bound” estimates (while noting that the true effects are probably much greater).⁴ Even so, the effects of teacher quality dwarfed those of other factors, such as class size.

In short, this growing body of research, which has been driven by the availability of value-added data, suggests that teacher quality may be the single most important in-school factor determining how much students learn. If true, the implications are extremely important. It suggests that the best way to raise students’ test scores is to improve teacher quality. This could be done in two ways. We could use value-added data on teacher quality to help make personnel decisions, and/or we could apply the lessons learned from value-added analysis to teacher education and professional development.

Focusing on personnel decisions would bring the quickest results, but this approach is fraught with thorny problems. The mechanics of applying value-added criteria to personnel decisions are straightforward. States would report the average test score gains of students in every teacher’s classroom to school administrators. The administrators would use these data in their decisions on hiring, firing, promotion, and salary. Teachers who raised their students’ test scores the least would be offered

help. But ultimately, if they did not improve, they would be replaced. Teachers who increased scores the most could be promoted to instructional leadership positions. They could receive raises and be given additional responsibilities to help peers who were struggling. Faced with pressures to meet more rigorous performance requirements, schools would have incentives to hire teachers who add the most value, thus creating a more competitive climate for good teachers. Teachers would also face strong incentives to improve under this approach. And, while many would succeed, some of the lowest performing teachers would leave the profession, which would also improve average teacher quality.

Quantitative value-added data would not be the sole criterion used to make personnel decisions. Nevertheless, it is possible that any use of value-added criteria in hiring, firing, and promotion would raise objections among teachers and teachers’ unions. Ironically, the notable effects of teacher quality found using value-added methodology demonstrate the importance of teachers and could be used to make the case more persuasively for higher salaries and other incentives to attract and retain good teachers.

A less contentious approach would be to apply the lessons learned from value-added analysis to teacher education and professional development. That way, we could raise the quality of the entire distribution of teachers. One problem, however, is that we do not know much about how to teach teachers to add value. Teachers with master’s degrees do not raise their students’ test scores any more than teachers without them.⁵ This suggests that the things that are being taught to teachers in school, at least at the graduate level, do not add value. The evidence is preliminary and mixed with regard to the effects of other advanced certifications like the National Board for Professional Teaching Standards. The Department of Education must support research to determine what kinds of curricula and pedagogical techniques are most effective in teaching teachers how to add value. Value-added testing regimes would provide the data needed to fuel such research.

Finding School Reform Models that Really Work

The third benefit of a value-added testing regime is that it could help us find school reform models that really work. The best way to improve schools is to rigorously test a wide variety of different strategies and models and then systematically winnow out the failures and build on the successes. Policymakers and school administrators are beginning to take such an evidence-based approach to education seriously. The Obey-Porter Comprehensive School Reform Demonstration Program, which provides resources to implement and evaluate school reform models, was an important first step in this direction. The Department of Education is undertaking a major initiative to promote the use of evidence in educational policymaking at the federal, state, district, and school levels.

To be able to test the efficacy of school reforms, evaluators need to be able to compare changes in the test scores of students in schools carrying out reforms to changes in the scores of students in control or comparison groups. Under a value-added testing regime, data on test score changes is already collected for every student. This makes it much easier to carry out large-scale evaluations. By increasing the amount of data available to researchers, a value-added system can increase the reliability and lower the costs of evaluations.

For example, in 1995, the city of Memphis began restructuring its elementary schools, using several different comprehensive school reform models. Restructured schools were compared to matched control schools, in terms of their performance on the Tennessee Value-Added Assessment System over the subsequent five years. Because it routinely collected value-added data for the entire district, the city was able to test a number of different models at once, comparing them to each other and to traditional schools.

Issues Surrounding Value-Added Testing

Value-added testing regimes have much to offer, but they are not cure-alls. We need to understand the limits and problems of this ap-

proach. First of all, value-added analysis is statistical. By definition, all statistical approaches are imprecise to one degree or another; there is always a margin of error. Thus, they are able to make distinctions at the extremes more accurately than they can in the middle. In other words, value-added analysis can reliably identify schools that are well above or below average, but not necessarily those that are slightly above or below average. It is particularly important to keep this caveat in mind if value-added analysis is used to evaluate teachers. While it could be of great benefit in this regard, it should never be used as the sole criterion in making personnel decisions, because there is always the possibility of error in any individual case, especially for teachers near the average. However, when the analysis identifies teachers who are either outstanding or very poor at adding value, the results should carry a lot of weight.

Second, value-added analysis will not solve one problem that all testing regimes face—the difficulty of distinguishing between true learning and teaching to the test. As schools and teachers are being held increasingly accountable for test scores, the temptation to devote class time to drilling in answers to questions expected to be on the test will grow. This strategy can raise test scores and thus give the impression that schools and teachers are adding value. But the students are not really learning, they are just memorizing answers. The best way to limit this practice is to make it harder to teach to the tests by changing them frequently. Large national testing companies tend to do a better job of this than smaller, local ones, simply because they have greater resources.

Third, there is still a fair amount of disagreement over the explanatory power of value-added analysis in relation to other factors. As noted above, Sanders and his colleagues have used value-added analysis to demonstrate that teacher quality has huge effects on student test scores. Sanders argues that teacher quality and other characteristics of the school are so powerful that socioeconomic characteristics of students are, for all practical purposes, irrelevant to determine test scores in a value-added

framework. That is why Tennessee's value-added system, which was heavily influenced by Sanders' work, makes no attempt to adjust for such factors. However, Hanushek and his colleagues contend that this view overstates the case. They agree that teacher quality may very well be the single most important factor determining annual changes in test scores. But they believe that socioeconomic factors still play a role. In line with this thinking, the city of Dallas uses a value-added system that adjusts for socioeconomic factors.

At this point, it would be prudent for states and districts wishing to implement a value-added system to use a less extreme version, similar to the Dallas model. Sanders has not released all of his data, so some of his findings cannot be replicated. At least until this is done, we should not assume that teacher and school quality are the only factors determining how much students learn each year. The more states and districts to develop value-added systems that use a variety of factors, the more we will learn about the relative effects of each factor.

Implementing a Value-Added System

Since ESEA mandates that all schools must test their students annually, the additional costs and requirements of implementing a value-added system would actually be relatively limited. And while the framers of ESEA did not have value-added systems in mind when they drafted the legislation, there is certainly nothing in the bill that precludes its implementation. A value-added system uses the same raw data that is collected in standard testing regimes. Only two other things are needed. First, schools need database software to link students' scores in one year both to previous scores and to their teachers' identities. Second, a statistical model is needed to generate estimates of the effects of schools and teachers on the annual changes in scores.

Another tricky thing involved in implementing a value-added system is the need to keep track of students who change schools and districts (and, in an ideal world, even states). Since the system focuses on the change in test

scores from year to year, it needs to know how each student tested the previous year regardless of where they went to school. Movement across schools within the same district and across districts is fairly common, so it is important that these movements are tracked in the database. This is not difficult to do, but it does necessitate some involvement at the state level to ensure that there is information sharing across schools and districts. Ideally, different states would share information with each other as well. But since relocation across states is less common, it is not an absolute necessity. Data required to make value-added estimates would be missing for a student the first year he or she entered a state, but some data are always missing in any testing regime. Estimates are still valid, as long as the amount of missing data is reasonably small.

Value-added estimates would be a little more accurate in states that use statewide tests, rather than those that let each district choose their own. Different tests are not always directly comparable, so value-added estimates would be somewhat less reliable for students who switched between schools using different tests. However, norm-referenced measures and statistical adjustments can minimize these problems, so a value-added regime would still be tremendously beneficial even in states that do not use statewide tests.

Where Do We Go From Here?

While valued-added testing has its problems and some issues remain unresolved, the potential benefits of implementing such a system are clear. As discussed above, there are at least three major benefits.

Value-added systems would provide a more accurate picture of which schools, school districts, and states are succeeding and which ones are failing. They would generate objective measures of teacher performance that could be used to raise the quality of teaching in our schools. And they would provide data that could be used in evaluating school reform programs.

To implement a value-added system, a state or district needs only to connect students' test

scores within a relational database software program to link their scores in previous years and to the identities of their teachers. Then they need to develop a statistical model for analyzing the data. Since ESEA already mandates that test scores be collected annually, the additional cost would be small for urban school districts and trivial for any state. Problems associated with a value-added system could be minimized by implementing it statewide and keeping careful track of students as they move across schools and districts. A value-added system will not solve every problem or answer all questions, and thus states should continue collecting and reporting test score levels. They

must also consider other factors, such as socioeconomic characteristics, in their analysis.

We have entered a new era in American education. The emphasis on accountability has the potential to improve public education dramatically. But it will only work if the system we use to hold schools accountable is accountable itself. Standard testing regimes are, quite simply, inadequate to the task at hand. Adopting value-added testing regimes nationwide would go a long way toward ensuring that the results we get from the army of tests that we are currently mustering will provide reliable information that we can actually use to make schools better.

Jonathan Crane is a senior fellow with the Progressive Policy Institute and the editor of Social Programs that Work.

Endnotes

¹In comparison, estimates from the Lasting Benefits Study of Tennessee's Project STAR experiment suggest that a one-third reduction in class size over three years would, at most, increase an average child's test scores by 10 percentiles. Hanushek, Eric A., "The Evidence on Class Size," in Susan E. Mayer and Paul E. Peterson (Eds.), *Earning and Learning: How Schools Matter*, The Brookings Institution Press, 1999, p. 131-168.

²Wright, S. Paul, Sandra P. Horn, and William L. Sanders, "Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation," *Journal of Personnel Evaluation in Education*, Vol. 11, 1997, p.57-67.

³Sanders, William L. and June C. Rivers, "Cumulative and Residual Effects of Teachers on Future Student Academic Achievement," *Research Progress Report*, University of Tennessee Value-Added Research and Assessment Center, 1996.

⁴Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. "Teachers, Schools, and Academic Achievement," National Bureau of Economic Research Working Paper #6691, 1998.

⁵*Ibid.*