



From the Snapshot to the Full Picture: Measuring School Performance with Value-Added

By Douglas N. Harris

The development of the horse and buggy was a necessary first step toward the development of the automobile; in fact, the first cars were built by putting engines on buggies. So it is with school accountability. The failure of No Child Left Behind (NCLB) to measure school performance is well known among researchers and, to some degree, among policymakers and their staffs. There is a potential solution: “value-added” measures. Shifting from “adequate yearly progress” (AYP) to a value-added approach called “school performance tables” (SPTs) is possible due to the provisions of NCLB. The law’s expansion of standardized testing, the focus on accountability for results, and the goal of 100 percent proficiency were all important first steps. Now is the time to bring our horse and buggy accountability into the twenty-first century and upgrade to value-added measures.

Almost every education group in the country has weighed in about how to improve NCLB and provide better ways to turn around failing schools, but most of these ideas amount to putting antilock brakes on a horse and buggy. There is nothing wrong with antilock brakes, but first things first.

The first thing—the foundation of NCLB or any other accountability system—is performance measurement. This is not a new issue in education. Back in the 1840s, Boston schools administered a standardized test and ranked schools according to their performance with much public fanfare. In rural schools, spelling bees and other public exhibitions were common. Our approach to measuring school performance is not much different today. Under NCLB, we still take the basic spelling bee approach and judge schools based on their scores at a single point in time.

The problem is that federal AYP does not measure school contributions to student learning any more than spelling bees measured how much

Douglas N. Harris (dnharris@wisc.edu) is a professor of education policy at the University of Wisconsin.

teachers helped students improve their spelling. To see why, let us start at the very beginning of students’ school careers. Research is clear that students start kindergarten at vastly different levels of academic skill. These starting-gate inequalities are obviously not the fault of the schools—many students have not even set foot in a classroom before kindergarten.¹

Key points in this Outlook:

- The foundation of No Child Left Behind is performance measurement.
- Students start school with different achievement levels, and we must account for these.
- Value-added approaches, using school performance tables, address this issue.
- Combining value-added and proficiency components will make government interventions more successful.

Because knowledge and skill accumulate over students' entire lives, this problem does not go away in later grades. On top of starting-gate inequalities, there is a natural progression from elementary to middle school and then to high school. So, high school achievement depends on what happened in middle school. High mobility in some schools also means that some students switch between different elementary schools, often right in the middle of the school year. For these and other reasons, each school's students start off at different levels of achievement, and if we fail to account for that, then the student's current school will be punished (or rewarded) for achievement differences that are outside their control.

We are already having trouble attracting teachers to low-proficiency schools, which is partly why these schools do poorly to begin with. The focus on proficiency in NCLB accountability makes that problem even worse.

This is a longstanding problem created by state accountability systems that, while less aggressive than NCLB, still looked only at the end-of-year test scores and failed to account for where students started. Given the rapid expansion of testing, one might think the problem has been solved, but it has not. The federal law requires collecting a lot of new information, but then essentially throws the information away by continuing to evaluate schools based on snapshots of student performance, sometimes called "status models." This creates a wide range of perverse incentives that are bad not only for the schools, but also for the students—especially for the disadvantaged students the law is intended to help.

Yet, the debate on NCLB is focused on questions like: How do we turn around failing schools? Should we change the policy that uses federal funds for private after-school tutoring? Should school turnarounds be left to teams of state government experts? Should extra compensation be provided for teachers in failing schools? These are all interesting questions, but good answers are impossible without first improving the school performance measures. We are not only judging school performance unnecessarily crudely, but are also failing to target the intensity and type of intervention to the specific performance level. Forget about putting the

antilock brakes on NCLB. First, we need something that has an engine.

Most educators have heard of value-added approaches, and many school districts and states are already using them in one form or another. But federal policymakers have been slow to move. I propose replacing AYP with SPTs.

Principles for Measuring School Performance

My critique of NCLB and proposal for SPTs are based on five main principles:

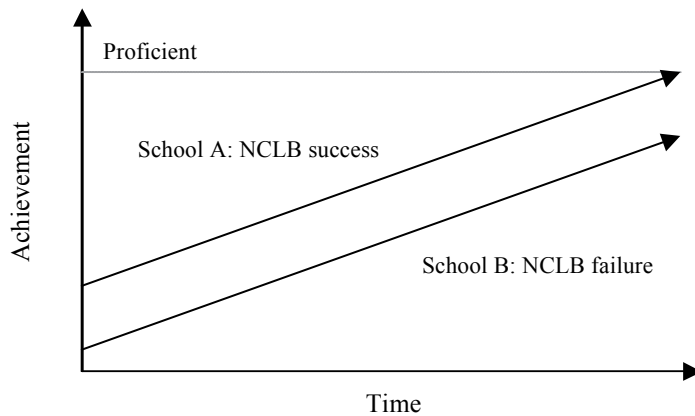
Principle One: Schools should be judged based on what they contribute to student learning. While this may sound obvious, NCLB rewards schools for who they teach, not how well they teach. A good school under current law is one in which students had high achievement before they entered the classroom—or, in practical terms, a school in which parents in BMWs drop their kids off. Instead, schools should be judged by what they contribute to student improvement.

We are already having trouble attracting teachers to low-proficiency schools, which is partly why these schools do poorly to begin with. The focus on proficiency in NCLB accountability makes that problem even worse. Teaching low-income students is challenging enough by itself; why should teachers put up with the added and misguided punishments that current law imposes?

Another perverse incentive is that the law encourages schools to focus on the "bubble kids" who are near proficiency level and need less attention to pass the test and provides little incentive to help students who are far above or below proficiency. Schools under the current system also have incentives to abandon programs whose effectiveness is not properly reflected in the inaccurate federal school evaluations.

Figure 1 illustrates this problem. Students in school A start off at a higher level of achievement compared with school B but generate the same amount of learning. Yet, school B is considered failing for a reason that is outside the control of the school—because its students started school with lower achievement. School B generates just as much learning but, under current law, gets no credit for it. There are some exemptions in the law that make the NCLB approach more complicated than figure 1 suggests, but these do not eliminate the problem illustrated here. Moreover, the array of

FIGURE 1
SCHOOL PERFORMANCE MEASUREMENT UNDER NCLB



SOURCE: Author's calculations.

exceptions makes even less sense as a way of measuring school performance. The bottom line is that it is difficult to justify treating these two schools as differently as we do given that they are generating the same amount of learning.²

Principle Two: The goal of 100 percent proficiency should be maintained as a broad national goal. The first principle shows that proficiency is a poor way to measure school performance and that doing so creates many counterproductive incentives. But proficiency is a useful way of focusing our attention on a broad national problem—the large number of students who consistently fail to reach even the most basic standards of academic preparation. For that reason, the national goal of 100 percent proficiency is perfectly reasonable. But the law's use of proficiency confuses the means (measuring school performance for accountability) with the ends (getting all students to proficiency). It is as if we decided to set a national goal of zero automobile traffic injuries and tried to achieve it by banning all cars in which injuries ever occur. It is a nice goal, but I can only imagine the horrible cars we would end up with.

Principle Three: Schools doing very poorly in raising achievement should be treated differently than those doing very or moderately well. There is no question that there is a continuum of school performance—some do very well, others do very poorly, and others are somewhere in the middle. Common sense dictates

that school performance should be met with a proportional response (sanctions, interventions, or rewards). This is not just out of a sense of fairness, but because more targeted responses will make genuine school improvement more likely.

Principle Four: Standardized tests should remain the primary measure of student learning and progress, but other measures should be added to the accountability system. Businesses, when they judge the performance of their managers and workers, consider multiple factors, including the bottom line of profit and specific activities or intermediate outcomes that contribute to long-term business success. For example, it is difficult to determine how much the accounting department of a company contributes to overall profit, but the performance of accounting personnel can be judged based on factors such as the percentage of bills they pay on time and how often they have to restate earnings reports.

In education, school climate and student and teacher absences are important intermediate outcomes in schools, and these provide better indications about how well the school is functioning and how much achievement growth the school will generate in the future. The bottom line in schools is more complex than it is in business, in which profit is the main motive. For this reason, it is important to include graduation rates and other measures that are predictors of long-term life outcomes of students.

Principle Five: School performance measures should be relatively simple and easy to explain to teachers, parents, and other key stakeholders. Supporters of the current model argue that, even though it is probably an inaccurate measure of school performance, the system is easier to explain than the alternatives. However, the current model has so many flaws that the U.S. Department of Education has had to put in place a wide variety of band-aids—rules and regulations that make little sense and are anything but simple. While there are some technical issues with measuring achievement growth, these can be addressed in ways that make this approach clearly superior to the current AYP model, and the methods used to make these calculations can be explained in simple and intuitive language for teachers, parents, and others.

In short, while NCLB made several critical improvements in federal education policy, it is inconsistent with the five principles above. It does not evaluate schools based on what they contribute to student learning, it confuses means with ends, it does not facilitate proportional responses, it uses insufficient measures of student outcomes, and it is incredibly complex. We can do better.

The School Performance Table

The SPT is built on these five principles. It borrows some of the useful elements from the current NCLB model, while avoiding its weaknesses. The focus of the SPT is on the value-added index. For all the reasons listed above, this is the key measure for accountability purposes. In addition, there is a proficiency index, which is used to determine how close schools are to the 100 percent proficiency goal and perhaps to identify schools eligible for special programs that might help failing students catch up. The value-added and proficiency indices would also be calculated for specific racial and program subgroups, such as special education.

The value-added index can be calculated in a variety of ways, ranging from simple changes in student test scores over time (“growth”) to versions making sophisticated statistical adjustments to improve validity and reliability. Perhaps the most important issue is that tests in math and reading are only administered in grades three through eight, plus one high school grade. Without adjustments, the simple value-added measures could create incentives for schools to pay less attention to students in kindergarten through grade three, precisely the years when schools seem to have the most influence on students. To address this issue, the value-added indices would be adjusted to include significant focus on third-grade proficiency. Because growth in elementary schools can be calculated only for three of the seven grades in K–6 schools, roughly half the weight would go to third-grade proficiency and the other half to growth in fourth, fifth, and sixth grades. While not ideal, these adjustments are necessary given the limits in achievement information that is currently available.

Indices for four hypothetical elementary schools are provided in table 1. The lowest scoring school in a state on “value-added” would have a score of zero, and the highest would have a score of one hundred. Likewise, the lowest possible “percent proficient” is zero and the maximum is one hundred.

Walker Elementary and Hoover Elementary both have low proficiency, but Hoover has much higher value-added. Both schools are serving disadvantaged students, but only one—Hoover—is serving those children very well. Likewise, Roosevelt Elementary and Wilson Elementary are serving the more advantaged children, but only Roosevelt is doing well in contributing to learning.

Under current law, Walker and Hoover would surely be classified “in need of improvement” because they both have low levels of proficiency. In Walker’s case,

TABLE 1
SUMMARY OF SCHOOL PERFORMANCE INDICES
FOR FOUR EXAMPLE SCHOOLS

Measure	Walker	Hoover	Roosevelt	Wilson
Value-added				
Reading	4.0	26.0	30.4	12.8
Math	2.0	23.6	32.0	16.0
Subgroups and other	4.2	12.2	11.8	4.0
Index	10.2	61.8	74.2	32.8
Percent proficient				
Reading	5.5	7.0	16.3	16.0
Math	7.5	8.8	14.8	19.0
Science	2.7	3.6	11.9	11.0
Social studies	1.8	3.9	10.8	10.7
Subgroups and other	4.2	12.2	11.8	4.0
Index	21.7	35.5	65.6	60.7

SOURCE: Author’s calculations.

that makes sense—the school has low value-added. But Hoover is well above average (fifty) on value-added, so it makes little sense to treat these schools the same way. When it comes to government intervention, Walker and Hoover should be treated differently, as should Roosevelt and Wilson.

Interventions, Rewards, and Resources

The value-added and proficiency components of the SPT provide different types of information about schools, all of which are important in deciding how the government should respond. How would the federal government judge and respond to the two indices for Walker Elementary? Table 2 shows the performance of schools on the two indices above—value-added and

proficiency. Ideally, all schools would be in the bottom right hand corner in which proficiency and value-added are both in the highest categories.

TABLE 2
SCHOOL PERFORMANCE TABLE

	1	2	3	4	5
Proficiency categories					
1 (0–19)					
2 (20–39)	Walker			Hoover	
3 (40–59)					
4 (60–79)		Wilson		Roosevelt	
5 (80+)					
	Intervention			Rewards	

SOURCE: Author’s calculations.

One of the advantages of having separate value-added and proficiency indices is that they can be used to target responses appropriately. The schools in columns one and two have low value-added. In such cases, it is reasonable for the government to intervene to make sure they improve. There is considerable debate about the best ways to intervene and what role should be played by the federal government versus lower levels of government. The point here is not to resolve those debates but only to emphasize that interventions and rewards intended to improve school performance should be based on school contributions to achievement—value-added.

Columns four and five in table 2 indicate high value-added. Current law includes little if anything that could be considered a reward for high performance. A good case can be made for doing so, both to make sure there are good incentives for improvement in all schools and to show appreciation for high performance. Rows one and two in table 2 include schools with low proficiency, though not necessarily low value-added. Since one of the goals is 100 percent proficiency, the government might also consider putting in place programs targeted to raising proficiency.

These three types of responses—interventions, rewards, and targeted programs—yield various combinations of responses. In the case of Walker, at which achievement levels and value-added are low, a combination of targeted

programs and interventions would be in order. While the targeted programs would involve additional resources and might be seen as a reward, the fact that it is coupled with strong government interventions means that schools will have a strong incentive to avoid being in a low-proficiency category, while still having the wherewithal to get more students above the proficiency bar.

Schools like Wilson in the low value-added category would receive interventions but no targeted programs because proficiency is relatively high already. Roosevelt, because it is doing well in both dimensions, would receive only rewards, but Hoover would see a combination of targeted programs and rewards. Most important, all schools would have incentives to improve. Even schools in the highest value-added category would have an incentive to maintain their annual rewards.

Notice also that the shading is darker in the far right and far left columns. This reflects principle three: “Schools doing very poorly in raising achievement should be treated differently than those doing very or moderately well.” This means the interventions would be more intensive for the schools in category one versus category two and the rewards larger for schools in cate-

Interventions and rewards intended
to improve school performance should be
based on school contributions to
achievement—value-added.

gory five versus category four. A good accountability system is based on proportional responses to performance.

Because the interventions and rewards are determined by the school’s value-added index, any student outcome, or other leading and lagging indicator that policymakers might consider important, could and should be included in the value-added index. It is tempting to ask how much weight or focus is given to proficiency versus value-added in the SPT. In some ways, this is the wrong question. Again, the value-added and proficiency indices measure different things and should be used for different purposes. Value-added provides a better indication of each school’s contributions to student learning, which should be the primary basis for interventions and rewards. Proficiency, in contrast, provides information about a combination

of factors—school readiness of students when they entered kindergarten, community characteristics, and the performance of previous schools students attended. Low proficiency suggests a different problem than low value-added and therefore requires a different response—specifically, programs targeted to the needs of low-achieving students.

Low proficiency suggests a different problem than low value-added and therefore requires a different response—specifically, programs targeted to the needs of low-achieving students.

While the value-added measures are clearly the most useful for determining government interventions, are they also the most useful for parents? Yes. As a parent, I ask, “Given where my child is starting off, how much will the school help my child learn?” The answer is determined by measuring a school’s value-added. So, there should be little conflict between the way we should measure performance for the purposes of government and the way we should measure it for the purposes of parents.³

Good Policy and Good Politics

School value-added measures like the SPT are gathering support among researchers and policymakers. In a recent survey of education finance experts and economists (the term value-added and much of the research actually come from economics), 90 percent said that value-added measures are among the best ways to gauge school performance, compared with 9 percent who favored the NCLB model of “test score levels.”⁴ Lest you think this group is of one political disposition or another, the same group was split in their support for school vouchers.⁵ Value-added is anything but controversial among the politically moderate people who study schools.

Value-added is also good politics. Sandy Kress, former president George W. Bush’s education adviser and one of the primary architects of NCLB, had been a school board member in the Dallas Public Schools, which was one of the first in the nation to use value-added to measure the performance of all its schools. He said

that he had wanted to use value-added in NCLB, but he decided it was not technically feasible at the time because most states did not have the necessary testing procedures in place, and educators and policymakers were less familiar with the value-added idea.⁶ So, advocates of the current law can rest assured that this will improve rather than undermine the law’s goals—even the law’s main architect thinks so.

The main potential concern is that value-added will eliminate the focus on the lowest scoring and most disadvantaged students, but there is no need for this to happen. The SPT rewards schools twice when they raise achievement for the lowest-scoring students. It shows up in the overall value-added index and in the subgroup value-added index. And the government will still report proficiency rates by school and provide targeted programs to schools in which proficiency is low. The SPT therefore maintains the focus on the most disadvantaged students but does so in a way that provides smarter incentives for schools. It also takes away the excuse that the accountability system is unfair.

Some might say we have already fixed this problem. The Bush administration approved “growth-to-proficiency” pilot models in fifteen states to address the above problems with proficiency. But growth-to-proficiency is actually very similar to the current AYP and very different from value-added. To see why, look back at figure 1. Under growth-to-proficiency, school A students are growing fast enough to reach proficiency, but school B students are not—the exact same conclusion we reached with AYP, before growth-to-proficiency. The reason the problem remains is that both approaches take proficiency as the ultimate arbiter of school performance. Research shows that schools are judged almost exactly the same way under AYP and growth-to-proficiency and very differently from value-added.⁷ The problem is anything but solved.

Beyond the Horse and Buggy

How we measure school performance matters—a lot. This short discussion is not enough to do justice to all the issues. How can we deal with the limitations of state standardized tests? Could value-added be calculated without annual assessments (which is common in high school)? How much room is there for local flexibility? There is not enough room here to go into the answers, but none of the issues raised in these questions should stop us from using the SPT.

The developers of the current model were well intentioned, but it is hard to dispute that the current model is a muddled mix of compromises that creates needless perverse incentives that are good for no one—students or teachers. Worst of all, it makes it difficult for those educators who want to support test-based accountability to stand up for it. For those who oppose accountability, it provides an easy excuse to ignore it. It is time to take away that excuse by addressing the law's real flaws and the legitimate concerns they raise, time to move beyond the horse and buggy approach and use more credible performance measures. We criticize schools often enough for being relics of the past. There is no need to make school accountability policies yet another bad example.

For their useful comments, the author wishes to recognize Kevin Carey, Linda Darling-Hammond, Adam Gamoran, Sara Goldrick-Rab, Cathy Loeb, Robert Manwaring, Robert Meyer, Howard Nelson, Andrew Rotherham, Thomas Toch, and Michael Weiss.

Notes

1. Academic skill in very young children is often called “readiness.” See Valerie E. Lee and David T. Burkham, *Inequality at the Starting Gate* (Washington, DC: Economic Policy Institute, 2002).

2. Federal rules require schools “in need of improvement” to select from a menu of intervention options, so there is some local discretion; however, as we get closer to the 2014 deadline, by current estimates, a very large percentage of schools will be

“in need of improvement” and subject to a single set of federal responses.

3. This statement assumes, of course, that the government and parents agree on the goals of education. Even then, an evaluation of school personnel by the government would differ slightly from an evaluation by parents because students' outcomes are also influenced somewhat by the other students in their classrooms. Who attends the school is largely outside the control of school personnel and, in principle, should be excluded from the performance calculation. Research suggests that it is somewhat beneficial for students to attend schools with higher-achieving peers, though the effects are complicated and the benefits probably relatively small. For a review of evidence on this topic, see Douglas N. Harris, “How Do School Peers Influence Student Educational Outcomes? Theory and Evidence from Economics and Other Social Sciences,” *Teachers College Record* (forthcoming).

4. Anne K. Rotenberg, Amy E. Schwartz, and Leanna Stiefel, “The Views of AEFA Members on Issues in Education Finance and Policy” (paper, annual meeting of the American Education Finance Association, Nashville, TN, March 19–21, 2009).

5. Forty-eight percent were in favor of vouchers for students in low-performing schools, and 34 percent were against such measures (18 percent were neutral).

6. Thomas Toch, “Measure for Measure,” *Washington Monthly* (October/November 2005), available at www.washingtonmonthly.com/features/2005/0510.toch.html (accessed July 14, 2009).

7. Michael J. Weiss, “The ‘Growth Model’ Pilot Isn’t What You Think It Is,” *Education Week* 27, no. 42 (2008): 28–29.