In one suburban school district, teachers across the system were ranked and evaluated according to the contribution they had made to student learning—based on a value-added analysis of state test results. When they were ranked again the next year, the results were very similar except for one teacher who moved from a top rank to a very low rank. When the school superintendent looked at the results, he immediately identified the teacher—her husband had died during the second year of rankings.

What does this anecdote illustrate? For starters, that value-added assessments can misidentify good and bad teachers—and that the discrepancies can be cleared up by local administrators. More important, however, was the fact that the findings were robust; teachers’ rankings were similar across three years of analysis. Moreover, these rankings were used to grant a standard reward to a top group of teachers, not to make fine distinctions in the amount of compensation that teachers received. In other words, the results of value-added analysis, when examined over a period of two or three years, were stable and were used in a way that respected the margin of error involved in using these statistical techniques.

Critics of value-added assessment tend to embrace the concept but don’t want the results gleaned from such analysis to be used for accountability purposes—and especially don’t want to use the results to reward or sanction teachers. But teachers are the dominant school input, in terms of both spending and impact on student learning. Excluding them essentially leaves the education system without accountability.

The main concern with value-added assessment is that the technique exacerbates the amount of random error involved in measuring student performance. The risk is that teachers and schools may be
wrongfully rewarded or punished because value-added techniques either over- or underestimated their students’ learning gains. However, no intelligent users of standardized testing would make policy choices based on a single year’s result or small differences among schools and teachers. Texas, for example, rewards schools based on value-added achievement gains calculated using two-stage regression analysis. The state does not hand out awards based on decimal-point differences among schools; officials reward a previously set percentage of top-ranking schools. The point here is that most of the statistical objections to value-added measurements assume a misuse of the analysis. The statistical “noise” involved in measuring value-added should preclude decisions that are based on small, unreplicated analysis; it should not preclude decisions that are based on gross findings.

Confidence in gross findings can be developed by replication, by averaging results over several time periods, and by using several measures of the development of human capital—not tests alone, but also attendance rates, dropout rates, and promotion rates (a very high-quality assessment will track indicators of human capital such as post-secondary school earnings and higher-education outcomes as well). The richer the measures used, the less weight there is on the psychometric concerns involving test scores. The alternative is to rest teacher compensation on factors that have little to do with student learning. It is now well established, for example, that the number of degrees teachers possess and the number of hours teachers spend in education courses are unrelated to student learning. Put another way, an important criterion in the determination of a teacher’s salary does not have any bearing on the ability of the teacher to develop human capital. We know that because it has been replicated in many studies in many school districts, even though salary schedules have yet to reflect this information.

Critics often cite the difficulty of comparing the results of large and small schools and comparing one subject with another. It is essential in the debate over the usefulness of the value-added assessment approach that the unit of observation and the comparison group be specified. If, as has been the case in a number of places, the comparison group is all the teachers in a given grade in the school district—with, say, the top 15 percent of the 4th grade teachers receiving an award—what is the significance of a big or small school? The class is the unit, and class size tends to be uniform within a school district. At the high-school level, the comparison group is likely to be, for example, all history teachers or all science teachers across the district. The award-receiving group is a percentage of that comparison group, and it is not affected by the test scores in another subject.

If multiple factors are used for assessment, efforts to check on robustness are made, only extreme performances are rewarded or sanctioned, and comparison groups are selected carefully, the technical psychometric points raised by critics should not swamp the incentive and information benefits of performance-based compensation plans for teachers. If we are ready to determine reading programs, language labs, class sizes, and the use of computerized learning on the basis of value-added assessments, we should be ready to reward teachers using the same techniques.

There is already considerable evidence from several places—such as Tennessee and Florida, where value-added analysis has been used for accountability purposes—that low-achieving students are the main beneficiaries of the changes that occur when these techniques are implemented. When low-achieving students are taught the same body of knowledge over and over again, and when they are taught how to work under a time constraint, they benefit. Value-
Critics of value-added assessment say that it is just too complicated for teachers and the public to understand the results these systems will generate. In other words, the results will not be transparent. But if transparency is the criterion that trumps all other criteria, we would be compelled to use an assessment method that we know to be wrong. For example, there is no satisfactory way to make judgments about which method of teaching reading is superior—whole language or phonics—without factoring in the socioeconomic, school, and teacher characteristics of each of the groups of students in the experiment. Statistical controls must be used if the assessments of teachers, schools, or programs are to be accurate, even though very few educators understand the statistical principles and methods involved. I do not require a transparent understanding of the efficacy of the flu shot I take, nor do I require a transparent understanding of the operating characteristics of my car; I trust the experts on the techniques. So must it be in educational evaluation.

The problems with the use of value-added assessments, even for teachers, are greatly exaggerated, and the alternatives are simply untenable.

One proposed alternative is the use of one or more methods of subjective evaluation. Other teachers, students, and/or parents can be surveyed to make the judgments. Most of those surveys focus on whether respondents like the teacher, are happy in the classroom, or equally soft attributes. They do not really determine what kind of learning gains a certain teacher is eliciting relative to other teachers. Most people look back at their primary and secondary schooling and identify an extremely demanding teacher (whom they disliked at the time) as the one who made the biggest contribution to their educational development.

A second possibility is to set up a standard—some threshold—of student achievement as an absolute hurdle to define adequate performance. The problems with this are legion. How should the threshold be determined? Those who want to be rewarded set lower thresholds than those who watch budgets. The threshold becomes a major bargaining tool, quite divorced from the informed decisionmaking objective. What incentives for improvement are there for those who crossed the threshold?

Basicly, subjective evaluation allows the information essential to the rational allocation of educational resources to be derived politically rather than scientifically.

The state of American public education has been deplored by critics for many years. The notion that student learning has remained somewhat stagnant over the past century, though the country has spent steadily more real resources on education, is a matter of profound concern. This surely indicates that we do not know a lot about what works and what doesn’t.

We need to know if we are to change the pattern. We must have calibrated results; we must use sophisticated statistical methods to interpret the data; we need to use multiple measures of performance; and we need to implement the analysis in ways that are appropriate to the quality of the information. And teachers, the most important school-controlled input into the educational process, cannot be exempt from this information-gathering activity. Our current unhappy results are consistent with the subjective and/or unsophisticated tools we use for assessing effectiveness and creating incentives.

—Anita A. Summers is a professor emeritus of public policy and management at the W.arton School at the University of Pennsylvania.