

Sizing Up Test Scores

by DALE BALLOU

Illustrations by John Berry.

One of the basic critiques of using test scores for accountability purposes has always been that simple averages, except in rare circumstances, don't tell us much about the quality of a given school or teacher. The high scores of students in a wealthy suburban New Jersey school will reflect the contributions of well-educated parents, a communal emphasis on academic achievement, a stable learning environment at home, and enriching extracurricular opportunities. Likewise, the low scores of students in an inner-city Newark school will reflect the disadvantages of growing up poor. The urban school might have stronger leadership and a more dedicated teaching staff, yet still score substantially lower than the suburban school. As a result, in the past decade researchers have grown interested in ways of measuring and comparing the *gains* in academic achievement that a school or teacher elicits—in other words, a school or teacher's "value added." Say, for instance, that a school lifts its students from the 35th percentile on national tests to the 50th percentile. An accountability system that uses value-added assessment might judge this school more effective than a school whose students consistently score at the 60th percentile. A value-added system might also identify a school's best and worst teachers by tracking their students' gains in the course of a year. The prospect of measuring the contribution made by schools and teachers to their students' progress is winning a growing number of converts to value-added assessment. However, some practical complications stand in the way.

It is important, first, to distinguish between assessment for diagnostic purposes and assessment as a mechanism of accountability. Value-added assessment has demonstrated its value in the former capacity. Pioneering work in Dallas and in Tennessee has shown that value-added assessment provides information that can be useful when viewed in context by educators who understand local circumstances.

The more serious difficulties arise when value-added assessments are used to hold schools and teachers accountable, with high-stakes personnel decisions to follow. The danger is that such assessments will be used to supplant local decisionmaking, rather than to inform it. Unfortunately, our instruments of assessment are not precise or dependable enough for this purpose. I will discuss three problems: 1) current methods of testing don't measure gains very accurately; 2) some of the gains may be attributable to factors other than the quality of a given school or teacher; and 3) we lack a firm basis for comparing gains of students of different levels of ability.

- *Measured gains are noisy and unstable.*

Tests are not perfect measures of student ability or achievement. A student's performance on any given test will be due partly to true ability and partly to random influences (distractions during the test, the student's emotional state that day, a fortuitous selection of test items, and so on). This test "error" causes problems enough when we attempt to assess a student's *level* of achievement. The problems are significantly compounded when we take it a step further, to measuring achievement *gains*. A gain score is the difference between two test scores, each of which is subject to measurement error. The measurement errors on the two tests, taken months apart from each other, are unlikely to be related (after all, these are random influences). When we subtract one score from another, the measurement errors do not cancel out. However, a student's true ability does not change that much from one test occasion to another. When we subtract one score from another, a good deal of the portion of the scores that represents true ability *will* cancel out. The result: the proportion of a gain score that represents measurement error is magnified vis-à-vis the initial scores. In statistical parlance, gain scores are much noisier than level scores.

Statisticians facing this problem have adopted procedures that adjust raw measures of gain to minimize the contribution of statistical noise. The noisier the data, the less weight is placed on measured gains for any one school (or teacher). In extreme cases, the school or teacher in question is simply assigned the average

level of effectiveness. Of course, the amount of noise in the data is itself something that must be estimated. As a result, these statistical methods are quite sophisticated. Virtually no one who is evaluated by these methods—teachers or administrators—will understand them. Thus, value-added systems that adjust for the unreliability of raw test scores will fail one of the criteria that educators have deemed important for accountability: that they be *transparent*. Measured performance (as determined by the statistical models) will not accord with the raw data. It will be impossible to explain to the satisfaction of educators why two schools (or teachers) with similar achievement gains nonetheless received different ratings of their effectiveness.

Moreover, inequities will arise simply because measured gains are more dependable for schools and teachers for whom there are more data. There will not be enough information about teachers who are new to a school system to obtain reliable estimates of their effectiveness based on past performance—they will simply be deemed “average.” Likewise, it will be considerably more difficult for a small school to rank high: it will have to outperform larger schools in order to appear equally effective. (In the same way, a small school’s inferior performance may go undetected.) Discrepancies will also arise across subjects. For reasons probably due to the home environment, more of the variation in student reading performance is independent of school quality than is the case in math performance. As a result, it is harder to detect particularly strong (or weak) performance by reading instructors than by math teachers.

In the end, using sophisticated methods of value-added assessment may not be worth the trouble if the object is to identify and reward high performance. William Sanders, formerly of the University of Tennessee and now at the SAS Institute, has done pioneering work to develop a system of value-added assessment, using the results of annual tests administered to all elementary and middle-school students in Tennessee. The great majority of teachers assessed by this system do not differ from the average at conventional levels of statistical significance. A recent investigation of achievement in one large Tennessee school district (in which I am collaborating with Sanders and Paul Wright of the SAS Institute) has found that 20 percent of math teachers are recognizably better or worse than average by a conventional statistical criterion. By the same criterion, the percentage falls to 10 percent in language arts instruction and to about 5 percent among reading teachers. Those who want to reward teachers on the basis of measured performance should consider whether it is worth the trouble and expense to implement value-added assessment if the only outcome is to reward small numbers of teachers. Of course, it is possible to disregard statistical criteria and reward the top 10 percent of teachers in all subjects willy-nilly. But then many rewards will be made on the basis of random fluctuations in the data.

- *Gain scores may be influenced by factors other than school quality.*

Value-added assessment has one signal merit: it is based on student progress, not on the level of achievement. Schools and teachers are accountable for how much students gain in achievement. They are not given credit for students entering at a high level or penalized when their students start far behind. In effect, value-added assessment “controls for” the influence of family income, ethnicity, and other circumstances on students’ initial level of achievement.

However, this may not be enough. The same factors may influence not just the starting level, but also the rate of progress. Thus, even in value-added assessment, it may be necessary to control explicitly for these factors (or demonstrate that they do not matter).

The socioeconomic and demographic factors that might influence student progress make a long list. In practice it is unlikely that an assessment system will have access to data on student backgrounds beyond what is routinely collected by school systems: the percentage of students with limited English proficiency, the percentage eligible for free and reduced-price lunch, and the ethnic and racial composition of the student population. Clearly other factors also matter. Critics of high-stakes assessments will object that without an exhaustive set of controls, the assessment system will end up penalizing some teachers and schools for circumstances beyond their control. Unless it can be shown that value-added assessment need

not account for these other influences, schools that receive low marks will have an obvious excuse: the assessment did not recognize that “our students are harder to educate.”

Some progress has been made in this task. Sanders, Wright, and I have found that introducing controls for students’ race, eligibility for free and reduced-price lunch, and gender (together with the percentage of a teacher’s students eligible for free and reduced-price lunch) usually has only a minor impact on teachers’ measured effectiveness. However, this study was limited to one school district and one series of achievement tests. Whether the results will generalize remains to be seen.

Moreover, even small differences in measured effectiveness can have practical consequences for schools and teachers, depending on how these assessments are used. For example, suppose it is school policy to reward teachers who score in the top 10 percent. Whether a specific teacher falls into this category can be rather sensitive to the inclusion (or omission) of controls for student background. Even relatively modest changes in measured effectiveness, such as our research has found, can have a decisive influence on whether a teacher falls above or below a cut-off point defined in this manner. A teacher who would rate in the top 10 percent on one measure has only to fall slightly below the cut-off on the other measure to drop out of the category of teachers who are recognized for their excellence. Our findings suggest that this will happen with some frequency: more than one-third of the teachers who ranked in the top 10 percent when our assessments included socioeconomic and demographic controls no longer belonged to that category when these controls were omitted from the analysis.

- *Similar gain scores are not necessarily comparable.*

To practice value-added assessment, we must be able to compare the achievement gains of different students in a meaningful way. We need to be assured that the scale on which we measure achievement is one of equal units: one student’s five-point increase on an achievement test, from 15 to 20, must represent the same gain as another student’s five-point increase from 25 to 30 (see [Figure 1](#)). If it does not, we will end up drawing false conclusions about the relative effectiveness of these students’ teachers and schools.

Mathematicians who specialize in measurement in the social sciences, together with experts in the construction and interpretation of tests—psychometricians—have devoted considerable attention to this matter. Their findings are highly unfavorable to value-added assessment. First, it is clear that a simple tally of how many questions a student answered correctly will not have the desired property. Test questions are generally not of equal difficulty. Raising one’s score from 15 to 20 might well represent a different achievement gain than an increase from 25 to 30, depending simply on the difficulty of the additional questions that have been answered. This objection also applies to several popular methods of standardizing raw test scores that fail to account sufficiently for differences in test items—methods like recentering and rescaling to convert scores to a bell-shaped curve, or converting to grade-level equivalents by comparing outcomes with the scores of same-grade students in a nationally representative sample.

In the 1950s, psychometricians began to deal with this issue in a systematic way. The result has been the development of “item response theory,” which is used to score the best-known and most widely administered achievement tests today, such as the CBT/McGraw-Hill Terra Nova series and the National Assessment of Educational Progress. In item-response theory, the probability that a student will answer a given item correctly is assumed to depend on the student’s ability and on the difficulty of the item, as expressed in a mathematical formula. Neither a student’s ability nor the difficulty of the item can be directly observed, but both can be inferred from the pattern of answers given by a particular student as well as by other students taking the same test. For example, as one would expect, the more students who answer a given item correctly, the “easier” the item is judged to be. The estimate of a student’s ability (known as the *scaled score*) is expressed on the same scale as item difficulty.

The critical question, given that neither ability nor item difficulty can be measured directly, is whether the procedures of inference are powerful enough to put the resulting ratings of ability and difficulty on equal-unit scales. Has student A, whose scaled score rose from 500 to 600 (using item-response theory methods), truly learned less than student B, whose scaled score rose from 300 to 450? Does 100 points at one range of the scale really represent less learning than 150 points at another point on the scale? The fact that we express both scores numerically predisposes us to answer affirmatively. The dubiousness of such a response can be appreciated by approaching the question from another angle, taking advantage of the fact that ability is measured on the same scale as item difficulty. Suppose a student has answered test item A correctly, which has a difficulty rating of 500. Item B is harder, with a difficulty of 600. Clearly the student needs to know more to answer question B than question A. But is the extra knowledge required to answer question B truly less than the extra knowledge required to answer item C, with measured difficulty of 450, compared to item D, with a difficulty of 300? Does a numerical difference of 100 between items A and B really mean that the latter item “contains” 100 more units of something called difficulty—and that these are the same units in which the difference between items C and D is judged to be 150? Do we possess such a scale for difficulty, or are we merely able to determine the order of difficulty, assigning higher numbers to items judged to be harder?

Generally speaking, the latter account of the matter is the correct one. And because ability is measured on the same scale as difficulty, the same holds true of it. In practice, psychometricians usually act as if ability scores are on an equal-unit scale (or, in technical terms, an “interval” scale). But this is merely an assumption of convenience. As prominent psychometricians have pointed out, many of the usual procedures for comparing achievement gains yield meaningless results if the ability scales lack this property.

In the previous example, the size of the intervals, 100 and 150, depended on the choice of the mathematical function expressing the probability of a correct response. Different choices for that function will produce different scales. And choice is just what takes place. There is no “correct” choice—or, more precisely, we possess no criteria for determining whether one choice is more correct than another. Statistical properties, such as the “fit” of the data to the model, are of no help here. As noted in a 1986 article in the *Journal of Educational Measurement* by Wendy Yen, formerly the chief research psychologist with CBT/McGraw-Hill and now with the Educational Testing Service (ETS), there are infinitely many nonlinear transformations of the ability scale that will fit the data equally well, yielding the same probabilities. Yet these transformations will shrink the scale over some ranges and expand it over others, so that student A appears to make more progress than student B using one scale, but less using another.

These conclusions cut the ground out from under value-added assessment. Our efforts to determine which students gain more than others—and thus which teachers and schools are more effective—turn out to depend on conventions (arbitrary choices) that make some educators look better than others. This does not mean that testing is of no value. It is still possible to rank students’ performance and to ask how many students exceed a specified benchmark. But the finer kinds of measurement required to compare the progress of students at different levels of initial ability exceed the capacities of our instruments. As Henry Braun of ETS wrote in a 1988 article for the *Journal of Educational Measurement*: “We have to be very careful about the questions we ask and very sensitive to the possibilities for obtaining misleading answers to those questions. . . . In particular, we should probably give up trying to compare gains at different places on the scale for a given population.”

Conclusion

There is a simple idea behind value-added assessment: schools and teachers should be evaluated based on student progress. As the foregoing discussion shows, however, successful implementation of this concept is far from simple. It is much harder to measure achievement gains than is commonly supposed.

Notwithstanding these problems, policymakers, educators, and the public will continue to look at indicators of student progress to see how schools are doing. For some purposes, this seems entirely reasonable.

Sanders's assessment system has been a beneficial diagnostic tool in Tennessee. But those who look to value-added assessment as the solution to the problem of educational accountability are likely to be disappointed. There are too many uncertainties and inequities to rely on such measures for high-stakes personnel decisions.

–Dale Ballou is an associate professor of economics at the University of Massachusetts at Amherst.
Published by the [Hoover Institution](#) © 2002 by the Board of Trustees of [Leland Stanford Junior University](#)

[home](#) | [past issues](#) | [search](#) | [about](#) | [subscriptions](#)