ILLUSTRATION / MICHAEL WARAKSA

Choosing the RIGHT Growth Measure

Methods should compare similar schools and teachers

State education agencies and school districts are increasingly using measures based on student test-score growth in their systems for evaluating school and teacher performance. In many cases, these systems inform high-stakes decisions such as which schools to close and which teachers to retain. Performance metrics tied directly to student test-score growth are appealing because although schools and teachers differ dramatically in their effects on student achievement, researchers have had great difficulty linking these performance differences to characteristics that are easily observed and measured.

The question of how best to measure student test-score growth for the purpose of school and teacher evaluation has fueled lively debates nationwide. This study examines three competing approaches to measuring growth in student achievement. The first approach, which is typical of systems using the popular student growth percentile (SGP) framework, eschews all controls for differences in student backgrounds and schooling environments. The second approach, typically associated with value-added models (VAM), controls for student background characteristics and under some conditions can be used to identify the causal effects of schools and teachers on student achievement. The third approach is also VAM-based, but fully levels the playing field between schools and teachers by eliminating any association between school- and teacher-level measures of test-score growth and student characteristics.

by MARK EHLERT, CORY KOEDEL, ERIC PARSONS, and MICHAEL PODGURSKY

We examine the appeal of these three approaches in the context of a system for evaluating schools, although the substance of our findings also applies to evaluations of teachers and districts. We conclude that the third approach is preferable in the context of educational evaluations for several reasons: it encourages educators in all schools to work hard; it provides performance data useful for improving instruction system-wide; and it avoids exacerbating labor-market inequities between schools serving advantaged and disadvantaged students. The key distinguishing feature of our preferred approach, and the reason we advocate for its use in evaluation systems, is that it ensures that the comparisons used to measure performance are between schools and teachers that are in similar circumstances. Similarly circumstanced comparisons are well suited to address the policy goals listed above, and in an evaluation context this is a more important consideration than perfectly capturing the school's or teacher's true causal effect on student achievement. Simply put, comparisons among similarly circumstanced schools send more useful performance signals to educators and local decisionmakers than the alternatives.

Student Growth Measures

The three approaches we examine in this article represent the range of options that are available to policymakers. The first approach, based on aggregated student growth percentiles, has been adopted for use in evaluation systems in several states. SGPs calculate how a student's performance on a standardized test compares to the performance of all students who received the same score in the previous year (or who have the same score *history* in cases with multiple years of data). For example, an SGP of 67 for a 4th-grade student would indicate that the student performed better than two-thirds of students with the same 3rd-grade score. An SGP of 25 would indicate that the student performed better than only one-quarter of students with the same 3rd-grade score.

To produce a growth measure for a district, school, or teacher, the SGPs for individual students are combined, usually by calculating the median SGP for all students in the relevant unit. The number of years of student-level data used to calculate median SGPs can vary. In our analysis, we use the median SGP of students enrolled in a given school over five years.

A key feature of the SGP approach is that it does not take into account student characteristics, such as race and poverty status, or schooling environments. Advocates of SGPs, and of "sparse" growth models more generally, view this as an advantage; they worry that methods that do take into account student or school-level demographic characteristics effectively set lower expectations for disadvantaged students. Critics of SGP-type metrics counter that not taking these differences into account may in fact penalize schools that serve disadvantaged students, which tend to have lower rates of test-score growth for reasons that may be at least partly out of their control.

Our preferred approach ensures that the comparisons used to measure performance are between schools and teachers that are in similar circumstances.

A second approach, by far the most common among researchers studying school and teacher effects, is a one-step value-added model. Many versions of the value-added approach exist. The version we use takes into account student background characteristics and schooling environment factors, including students' socioeconomic status (SES), while simultaneously calculating school-average student test-score growth. Specifically, we calculate growth for schools based on math scores while taking into account students' prior performance in both math and communication arts; characteristics that include race, gender, free or reduced-price lunch eligibility (FRL), English-language-learner status, special education status, mobility status, and grade level; and school-wide averages of these student characteristics.

Researchers have gravitated toward the value-added approach because, under some assumptions, it provides accurate information on the causal effects of individual schools or individual teachers on student performance. But interpreting growth measures based on the one-step value-added approach in this way requires assuming that the available measures of student and school SES, and the specific methods used to adjust for differences in SES, are both adequate. If the measures are insufficient and the academic growth of disadvantaged students is lower than that of more advantaged students in ways not captured by the model, the one-step value-added approach will be biased in favor of high-SES schools at the expense of low-SES schools.

The third approach we consider is also based on value-added but is carried out in two steps instead of one in order to force comparisons between schools and teachers serving students with similar characteristics. In the first step, we measure the relationship between student achievement and student and school characteristics. In the second step, we calculate a growth measure for each school using test-score data that have been adjusted for student and school characteristics in the first step.

By design, this third approach fully adjusts student test scores for differences in student and school characteristics. In fact, it may overadjust for the role of such differences. For example, suppose that students eligible for free or reduced-price lunch attend schools that are truly inferior in quality, on average, to the schools attended by ineligible students. The average gap in school quality between these groups would be eliminated in the first step of the two-step value-added procedure, and thus would not carry over to the estimated growth measures. Consequently, it is important to interpret the results using this approach accurately, as they do not necessarily reflect differences in the causal effects of schools and teachers on student performance. We argue below, however, that this approach is still the best choice for use in an evaluation system aimed at increasing student achievement.

Comparing Results from the Three Approaches

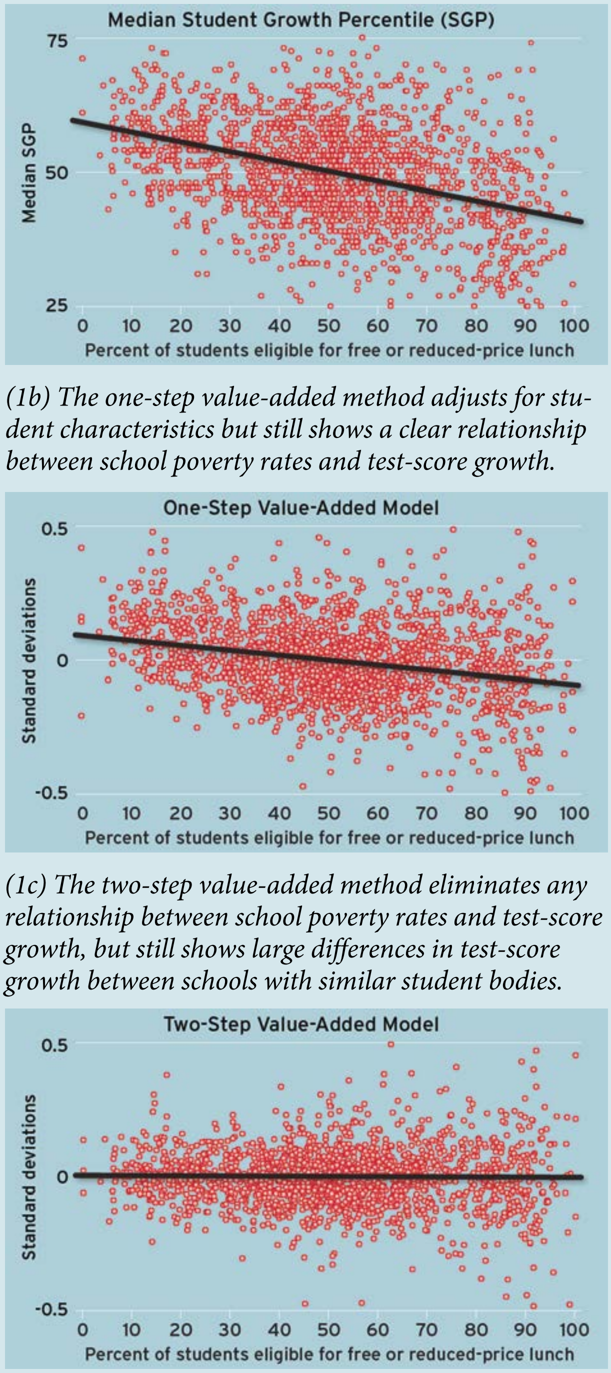
We calculate growth measures in mathematics for 1,846 Missouri schools serving grades 4 to 8 using each of the three approaches. The data available for our study are from the Missouri Assessment Program (MAP) test results, linked over time for individual students. They include nearly 1.6 million test-score growth records for students (where a growth record consists of a linked current and prior score) covering the five-year time span from 2007 to 2011 (2006 scores are used as prior-year scores for the 2007 cohort).

For both the SGP and one-step value-added approaches, we find a clear relationship between the school growth measures and the socioeconomic status of the student body, as measured by the percentage of FRL students. Figures 1a and 1b show that schools with more FRL students tend to have lower growth measures. In the case of the SGP approach, this reflects the fact that low-SES students make less progress, on average, than high-SES students, even after conditioning on prior test performance. The one-step value-added approach corrects for SES effects to some degree but a relationship still remains.

Figure 1c shows the same data for the two-step approach. Because of how it is constructed, this approach ensures that there is essentially no relationship between the growth measures and aggregate measures of student poverty. As a

Three Methods, Three Patterns (Figure 1)

(1a) Comparing schools using median student growth percentiles (SGPs) confirms that students in schools with many poor students make less progress on standardized tests.



(1b) The one-step value-added method adjusts for student characteristics but still shows a clear relationship between school poverty rates and test-score growth.

(1c) The two-step value-added method eliminates any relationship between school poverty rates and test-score growth, but still shows large differences in test-score growth between schools with similar student bodies.

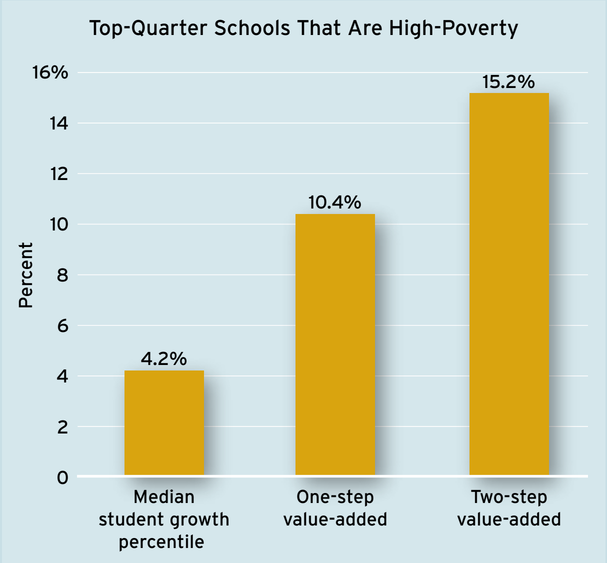
Two-Step Value-Added Model
Standard deviations
0.5
0
-0.5
0 10 20 30 40 50 60 70 80 90 100
Percent of students eligible for free or reduced-price lunch

NOTE: Each chart is based on mathematics test scores from the Missouri Assessment Program covering the five years from 2007 to 2011. Student growth percentiles and value-added figures cannot be directly compared because SGPs are calculated in percentiles and value-added measures are calculated in standard deviation units.

SOURCE: Authors' calculations

A Fair Comparison (Figure 2)

More high-poverty schools appear in the top quarter of growth estimates in the two-step approach than in either of the other two approaches.



NOTE: We define high-poverty schools as those in which at least 80 percent of students are eligible for free or reduced-price lunch. Using this definition, the share of high-poverty schools in Missouri is 13.3 percent.

SOURCE: Authors' calculations

result, high- and low-poverty schools are roughly evenly represented throughout the school rankings. A notable feature of the flat-line figure is that there are still considerable differences in the growth measures within any vertical slice in the graph. In other words, even when schools are compared to other schools with similar student bodies, large differences in test-score growth are clearly visible.

As we would expect, schools serving disadvantaged students are ranked higher by the two-step approach than in either of the other models. For example, high-poverty schools (those with at least 80 percent FRL students) make up just 4 percent of schools in the top one-quarter of all schools based on the SGP approach and 10 percent of top-quarter schools based on the one-step value-added approach. Using the two-step approach, however, high-poverty schools represent 15 percent of the top one-quarter (see Figure 2).

Choosing an Approach

Using the SGP and one-step value-added approaches, low-SES schools are ranked lower, on average, than high-SES schools. If this is entirely the result of bias in these two approaches, then the two-step method is an attractive alternative. But what if high-SES schools truly are more effective, on average, than low-SES schools? Even if this is the case, we argue that the two-step approach is still the most appropriate for use in evaluation systems for three reasons.

First, the two-step approach is best equipped to encourage public school employees to work hard. Research on how individuals respond to incentives shows that different signals often need to be sent to competitors in different circumstances. These signals need not be direct measures of absolute performance; instead, they should be indicators of performance relative to peers in similar circumstances. The logic is that if advantaged competitors are competing directly with disadvantaged competitors, neither group will try as hard as they would if all competitors were evenly matched. The two-step method encourages optimal effort by leveling the playing field. In contrast, the SGP and one-step value-added approaches do not result in balanced comparisons across school types and, in fact, favor the advantaged group, which runs counter to the goal of eliciting maximum effort.

Second, the two-step approach creates the kind of information that is most likely to help schools improve instruction. Measures of student achievement growth can improve instruction in K–12 schools by reinforcing positive educational practices and discouraging negative ones. For example, a positive performance signal might encourage a school to continue to pursue and augment existing instructional strategies. Alternatively, a negative signal can provide a point of departure for instructional change or outside intervention.

Information signals throughout the system can also be used to identify productive learning opportunities. Low- and high-poverty schools differ along many dimensions that likely influence what constitutes effective educational practice, including curriculum choice and implementation, instructional methods, personnel policies, and all the other day-to-day decisions that combine to create the educational environment. The two-step approach sends signals to schools about how they are doing relative to other schools in similar circumstances, rather than relative to all schools, many of which operate in quite different contexts. By doing so, this approach can help school leaders to identify those peer institutions that are performing well and are most likely to be a source of relevant lessons. Even if the two-step results conceal differences in absolute performance across schools in different contexts, they still facilitate comparisons among schools in similar contexts, which is sufficient to give schools performance signals that can be used to improve instruction.

Third, the two-step approach best avoids degrading the already-weak ability of high-poverty schools to recruit and retain teachers. It is well known that schools serving disadvantaged students are at a competitive disadvantage in the labor market (see "The Revolving Door," *research*, Winter 2004). As stakes become attached to school rankings based on growth measures, systems that disproportionately identify high-poverty schools as "losers" will make positions at these schools even less desirable to prospective educators. Policymakers should proceed cautiously with implementing an evaluation system that could worsen the working conditions in challenging educational environments. An important benefit of the two-step method is that the "winners" and "losers" from the evaluation will be broadly representative of the system as a whole.

Setting expectations for individual students is not the purpose of an evaluation system.

The two-step approach is preferable for each of these reasons, but a remaining concern about leveling the playing field across schools is that it will "hide" inferior performance at high-poverty schools. In our view, the best way to address this concern is to report the results from the two-step approach along with information on test-score levels. In fact, state- and district-level evaluation systems that incorporate test-score growth also typically report test-score levels and include them in schools' overall ratings.

Reporting test-score levels will allow policymakers to clearly see absolute differences in achievement across schools, regardless of which growth measure is adopted. Reporting results from growth measures that level the playing field in conjunction with information about absolute achievement levels is desirable because it allows for the transmission of useful instructional signals. For example, a low-SES school that is performing well can be encouraged to continue to refine and improve an already-effective instructional strategy (in terms of raising test scores compared to similar schools) but still be reminded that the students are not scoring sufficiently high relative to an absolute benchmark. The latter information need not disappear in any evaluation system that includes information on achievement growth.

A related concern is that the two-step approach will lower expectations for students in high-poverty schools. However, it is important to recognize that setting expectations for individual students is not the purpose of an evaluation system. Philosophically, policymakers may not want to lower expectations for disadvantaged students. If this is the case, then the proper approach to student-level evaluation is to set fixed performance benchmarks for all students and evaluate progress toward those benchmarks. None of the three approaches to measuring student growth that we consider here is designed to achieve this objective. For example, even SGPs allow for different growth targets for different types of students by taking into account individual prior achievement. Leveling the playing field between schools is a desirable property of a student-growth measure couched within the context of an educational evaluation system. Metrics used for other purposes may need to be designed differently.

Conclusion

We examine three broad approaches to measuring student test-score growth: aggregated student growth percentiles, a one-step value-added model, and a two-step value-added model. These approaches reflect the spectrum of choices available to policymakers as they design evaluation systems for schools and teachers. All three approaches produce growth measures that are highly correlated, but the high correlations mask an important difference. Only the two-step approach levels the playing field across schools so that "winners" and "losers" are representative of the system as a whole. Although the other approaches to measuring student growth may have benefits in other contexts, when one considers the key policy objectives of evaluation systems in education, the "leveled playing field" property of the two-step approach is highly desirable.

Some states are considering using, or are already using, aggregated SGPs as part of their evaluation systems. Policymakers in these states may not have carefully considered the issues associated with applying the SGP approach, or more generally, any "sparse" growth model, in the context of an evaluation system. A likely consequence is that schools and teachers serving disadvantaged student populations will be disproportionately counted as underperforming. At a minimum, states using SGPs in their evaluation systems should consider setting up "league tables" so that performance in high-poverty schools is not compared against performance in low-poverty schools.

Mark Ehlert is research associate professor of economics at the University of Missouri, Columbia, where Cory Koedel is assistant professor of economics, Eric Parsons is research assistant professor of economics, and Michael Podgursky is professor of economics.