ILLUSTRATION / MICHAEL GLENWOOD

How Should States Design Their Accountability Systems?

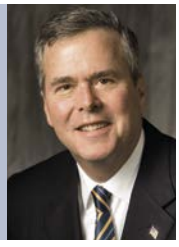
EDUCATION NEXT TALKS WITH JEB BUSH, HEATHER HOUGH, and MICHAEL KIRST

With the Every Student Succeeds Act (ESSA) replacing No Child Left Behind (NCLB) legislation, states have gained substantial new freedom to reshape their school accountability systems, including criteria for how to measure and communicate school performance to the public. One dominant model is the streamlined letter-grade system first adopted by Florida, which focuses on student achievement on annual statewide tests. By contrast, California is developing a dashboard-style system, which encompasses multiple measures, such as student attendance and school climate.

Below are two views on the merits of each model. Former Florida governor Jeb Bush, who pioneered education reforms in that state, including the A–F system, presents the case for summative ratings. From California, we hear from Heather J. Hough, executive director of the research partnership between the CORE Districts and Policy Analysis for California Education (PACE), and Michael W. Kirst, president of the California State Board of Education and professor emeritus of education and business administration at Stanford University, on the importance of multiple measures.

FLORIDA'S INTUITIVE LETTER GRADES PRODUCE RESULTS

by JEB BUSH



THE NO CHILD LEFT BEHIND ACT (NCLB) was a comprehensive, bipartisan response to a failing and inequitable public-education system, a system that held no one accountable for student learning, and as a result, consistently failed its most vulnerable charges. States were required to measure the academic achievement of all children, with schools accountable for results. Outcomes improved, particularly among minority and low-income students, according to data from the National Assessment *(continued on page 56)*

CALIFORNIA'S DASHBOARD DATA WILL GUIDE IMPROVEMENT

by HEATHER J. HOUGH and MICHAEL W. KIRST

AFTER MORE THAN A DECADE of strict federal mandates and measures of school success, a new education law is inviting policymakers across the country to rethink "accountability." The Every Student Succeeds Act (ESSA) takes a more comprehensive approach to assessing school quality than the No Child Left Behind Act (NCLB), moving beyond NCLB's focus on annual test performance to also consider factors like student academic growth, graduation rates, and rates of proficiency for English-language *(continued on page 57)*

BUSH
(CONTINUED FROM PAGE 55)

of Educational Progress (NAEP). But progress has not come fast enough, in part because NCLB came with an unintended consequence. The law's overly prescriptive approach created a perverse incentive for states to lower academic expectations in order to avoid federal sanctions.

Its successor, the similarly bipartisan Every Student Succeeds Act (ESSA), moves to correct some of those flaws by giving states more autonomy to fashion their own accountability systems and intervention policies. As states navigate implementation, I encourage them to use their expanded authority to strengthen accountability rather than retreat from it. This is in their students' interest, and their own self-interest. Look no further than research from Eric Hanushek, Jens Ruhose, and Ludger Woessmann about the strong correlation between achievement in a state's classrooms and growth in a state's economy to understand some of the compelling reasons to improve education (see "It Pays to Improve School Quality," *features,* Summer 2016).

A successful school-accountability system contains three basic elements: It gauges education quality and progress by measuring data that accurately reflect student achievement; it disseminates the results to parents and the public in a simple and transparent manner; and it rewards and incentivizes success and provides interventions to support low-performing schools and reverse failure. It is informative and focused on criteria that clearly support student success.

Lessons from Florida

Make no mistake: retreating from accountability is the easier path. In Florida, where I served as governor from 1999 to 2007, we know this from experience. Dating back to the 1970s, our state leaders attempted a series of ineffective initiatives to turn around one of the worst public-education systems in the country. At one time, almost half of our 4th graders did not qualify as even basic readers on NAEP.

A bold, new direction was required. And so in 1999, we overhauled our school system through accountability legislation that made student learning the focus of education. We adopted an accountability formula based on students' academic performance, requiring schools to focus resources on elevating achievement. Our letter-grade system gave parents a ready tool to assess school quality and make informed choices for their children. And even as the statewide Florida Education Association vehemently opposed these reforms, our students went on to become national leaders in making progress on NAEP (Figure 1).

That experience taught us that an accountability formula should reflect only objective measures of academic achievement. The focus should be on student performance on grade-level assessments in core subjects, and student growth on those assessments from year to year. At the high school level, other indicators such as four-year graduation rates and success in college- and career-ready coursework, including Advanced Placement, IB, or industry certification classes, should be added.

Data on inputs such as teacher training, disciplinary policies, attendance policies, and school resources may suggest important school-improvement strategies and should also be made available to parents. I don't disagree with a "dashboard" approach—it can provide important information to parents and inform intervention strategies.

Standardizing inputs such as disciplinary or attendance policies into an accountability formula diverts attention from student achievement, by micromanaging how districts, principals, and teachers run their classrooms.

But such data should not be included in an accountability formula. The bottom line must be student achievement. Standardizing inputs into an accountability formula diverts attention from student achievement, by micromanaging how districts, principals, and teachers run their classrooms. It bogs them down and reduces their flexibility in developing strategies that might work best for their individual situations.

As such, an effective accountability system requires rigorous assessments that accurately measure students' knowledge of state standards and preparedness for college or a career. Expectations for students and schools should be continuously evaluated and upgraded, with a realistic but constant raising of the bar.

In addition, teachers need to fully understand the goals and what is expected of them. This means state accountability systems must also be aligned to an individual teacher's classroom goals: Help all students meet proficient or higher performance; help all students make significant progress from wherever they were performing in the prior year; and pay laser-like attention to ensuring struggling students are on track to reach proficiency.

An effective formula includes both achievement and growth. This creates positive pressure for improvement, even in high-performing schools, and it recognizes the efforts of the extraordinary schools that have a disproportionate number of low-performing students but are making strong gains. The progress of the lowest-performing students should be included as well, regardless of what "subgroup" they're in, or the size of that subgroup. This ensures they receive the support they need to bring them up to grade level. *(continued on page 58)*

HOUGH & KIRST
(CONTINUED FROM PAGE 55)

learners. The law also requires at least one additional measure of "School Quality or Student Success" (SQSS), such as student engagement, college readiness, or school climate. And it empowers states to design their own accountability systems, leaving behind the one-size-fits-all mandates of NCLB.

In California, we've moved beyond assigning schools a single number score each year and are implementing a "dashboard" accountability system, to better capture and communicate multiple dimensions of school performance. Such a dashboard can provide rich information and support the many functions that accountability systems serve: providing guidance to parents and educators on district and school strengths and weaknesses; identifying struggling schools; and supporting the design and implementation of assistance strategies.

Yet, while ESSA requires states to consider multiple measures, current draft regulations then call on us to crunch them into a single, summative rating to identify struggling schools. This practice not only runs counter to the spirit of multiple measures, it is bound to create inaccurate ratings and should not be part of the final regulations adopted by the U.S. Department of Education later this year. While it may be true that moving to multiple measures will pose a new challenge for education stakeholders at all levels, trying to summarize all of these dimensions into a single number score or A–F letter grade will have misleading and negative consequences.

Simplicity, but at What Cost?

Single, summative ratings were in vogue over the past decade. The approach was pioneered in Florida, which began using letter grades for all its schools in 1999 under former governor Jeb Bush. Governor Bush played a role in spreading this idea to other states, and eventually 16 other states began to use the A–F grading system, with many others using something similar. The simplicity of such ratings meant it was easy for parents and the public to sort and rank schools by the supposed strength of their performance.

California embraced this approach as well. For more than a decade, we used the Academic Performance Index (API), which was based solely on test scores and established 800 as proficient on a scale of 200 to 1000. It is not clear why 800 was the magic number for a school to be judged doing well, but API became the coin of the realm. The scores, along with an accompanying number ranking schools across the state from 1–10, were promoted by web sites like GreatSchools.org, and became a marketing tool for real estate agents to sell houses in neighborhoods with "good" schools. Over time, it became clear that this kind of rating method punishes schools that serve disadvantaged communities; in California, the single score was so highly correlated with student demographics that it was sometimes referred to as the "Affluent Parent Index."

California relied on the overly simplistic API until 2013. While the public accepted it, a school's API only told parents, educators, and policymakers how students performed on English and math tests—an absurdly narrow view of school performance. Indeed, a primary impetus for the expanded measurement under ESSA was to move away from NCLB's narrow view of student success: by establishing test scores as the "bottom line," NCLB led many schools to focus exclusively on improving scores in tested subjects, which does not adequately prepare students to thrive in a competitive and complex global economy.

Most schools earn high scores in some areas and low scores in others, which means that a summative score, by definition, averages out this variation and conceals specific strengths and weaknesses.

New research shows that summative scores like API are not only uninformative, they are inaccurate when it comes to identifying low-performing schools. It's an important distinction, because ESSA requires that states designate their lowest-performing 5 percent of schools receiving Title I funds as in need of Comprehensive Support and Improvement (CSI), which triggers additional support and intervention. To determine which schools need CSI the most, it's important to understand which schools are struggling the most. And now with multiple measures, understanding which schools have low performance is not as straightforward as when we were only measuring test scores, as uninformative as they may have been.

The study, by Heather Hough, Emily Penner, and Joe Witte at Policy Analysis for California Education (PACE), examines the potential effects of using single measures in California's CORE Districts, where multiple measures of school performance are included in annual accountability reports. The six CORE Districts, which received a waiver under NCLB to develop their innovative, multiple-measures system, serve nearly one million students, almost three-quarters of whom come from low-income families. The unique, locally driven accountability system focuses on academic outcomes alongside nonacademic indicators, including rates of chronic absenteeism, suspensions, and expulsions, and measures of school climate, culture, and students' *(continued on page 59)*

BUSH
(CONTINUED FROM PAGE 56)

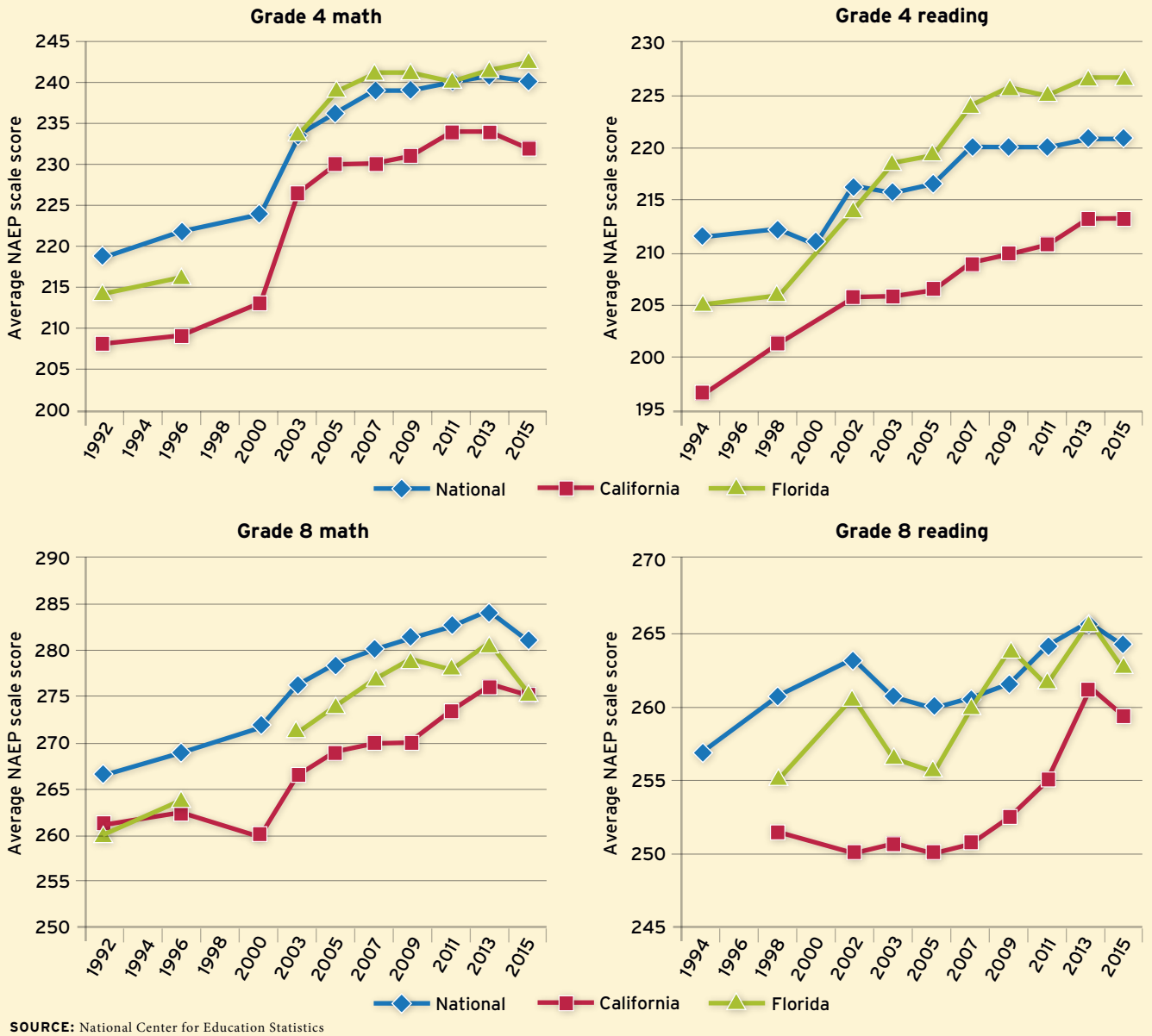
Simple and Transparent Reporting

An effective accountability system also requires that parents have a clear and concise measure of school performance. They should not have to struggle through confusing mazes of charts and spreadsheets to find out if their children are in a good learning environment. To get there, we begin with a simple, comprehensive, actionable score that captures the overall success of a school in advancing academic achievement. The most intuitive approach for parents is grading schools on an A–F scale.

School letter grades have a distinct advantage for educators as well: they are very effective at focusing educators on the goal of maximizing academic achievement. *(continued on page 60)*

Greater Grade 4 NAEP Gains in Florida (Figure 1)

In the last two decades, as NAEP scores have risen for the nation's public schools, Florida has made relatively greater strides, especially with grade 4 student performance.



SOURCE: National Center for Education Statistics

HOUGH & KIRST
(CONTINUED FROM PAGE 57)

social-emotional skills.

The PACE report demonstrates that used independently, different academic measures would identify different schools as the lowest performers. For all but the lowest-performing 1 percent of schools (which struggle across the board), a single number will inevitably produce arbitrary judgments about which schools are "better" and "worse," concealing the specific strengths and weaknesses of specific schools and depriving educators of the information that they need to improve.

The authors investigated the extent to which different academic measures—academic performance, academic growth, graduation, and English-language proficiency—would identify similar schools if used independently. They found that schools in the bottom 5 percent on any given indicator differed dramatically from measure to measure. In elementary and middle schools, for example, many schools with low academic performance also demonstrate high growth relative to similar schools (Figure 1). Just 13 percent of those schools are identified among the bottom 5 percent by both measures. Given how differently these measures distinguish among schools, summing them up in a single number or grade is a serious error.

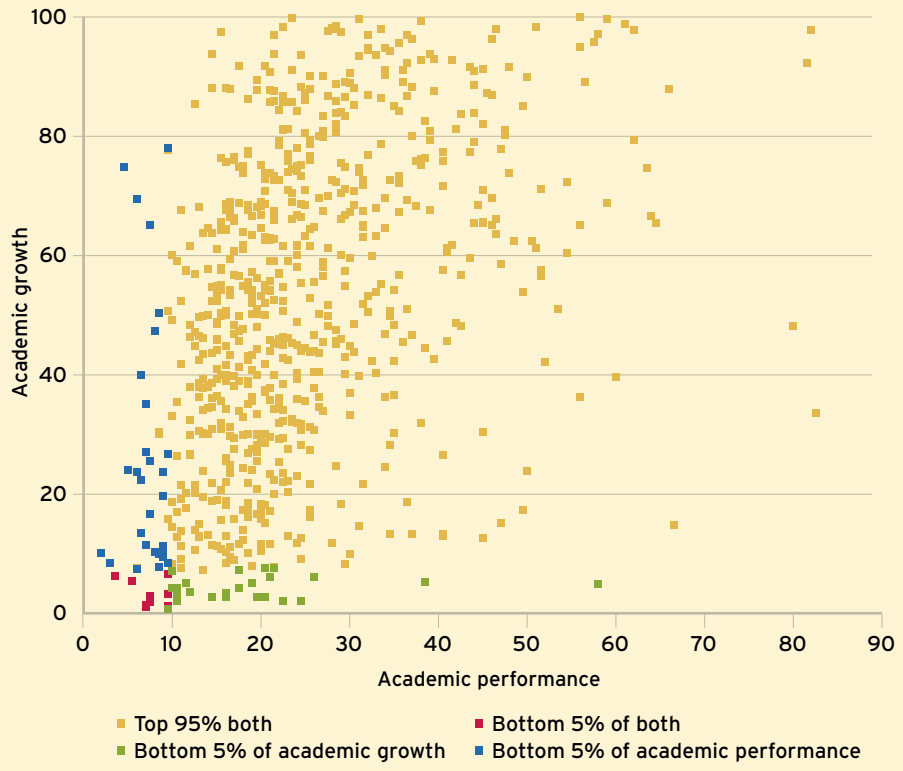
Most schools earn high scores in some areas and low scores in others, which means that a summative score, by definition, averages out this variation. PACE shows that an equally weighted summative score will identify schools that are low on all indicators, but will not identify many schools that are low on specific indicators. Among the studied schools, 2 percent are in the bottom 10 percent on all indicators, and all of them are identified using the summative measure. However, only 40 percent of schools in the bottom 5 percent for academic performance are identified for CSI using the summative measure (Figure 2). Similarly, only 45 percent of schools in the bottom 5 percent for academic growth are identified by the summative measure. For English language proficiency, it was 22 percent, and for graduation rates, it was 38 percent. By aggregating across measures that represent very different dimensions of performance, the summative score may not identify schools as low-performing if they are very low on just one measure, even if they are merely average, or even moderately low, on others.

Many Dimensions of Success

In addition to the academic indicators, ESSA specifies that states must include at least one indicator of "School Quality or Student Success" (SQSS). The CORE Districts have already begun collecting a wide range of such indicators locally, and there is much to be learned from their experience about how such measures can be integrated into state systems. As with the four academic measures, schools do not often demonstrate low performance on multiple measures *(continued on page 61)*

How a Single Measure Fails (Figure 1)

Schools in the bottom 5 percent on any given indicator differ dramatically from measure to measure. Schools where academic performance is very low are not necessarily the same schools where there is little academic growth.



N =749

SOURCE: Policy Analysis for California Education (PACE)

BUSH
(CONTINUED FROM PAGE 58)

After the implementation of Florida's letter-grade system, decades of failure were quickly reversed and our state became a national leader in advancing student achievement. Other states took notice and began implementing similar reforms, and today, there are currently 17 states using an A–F grading scale.

An analysis of the eight states with multiple years of implementation of the A–F grading system found they were making faster improvements on NAEP 4th- and 8th-grade reading and math tests than the nation as a whole. The analysis, by the Foundation for Excellence in Education (ExcelinEd), included Arizona, Florida, Indiana, Louisiana, Mississippi, New Mexico, Oklahoma, and Utah.

Letter grades are especially helpful in identifying schools that are struggling. Failure is perpetuated when it is hidden. But you can't hide from an "F."

An "F" is not a punishment. It is a distress signal. States and districts can respond with any number of strategies, including more resources, instructional coaches, a change in leadership, and more effective teachers. And in Florida, those that received such a signal changed their school policies and practices in meaningful ways and made long-term improvements, according to an exhaustive five-year study by Cecilia Elena Rouse, Jane Hannaway, Dan Goldhaber, and David Figlio.

Schools that received a grade of "F" not only improved test scores the following year, but those improvements "remained for the longer term," researchers wrote. "We also find that 'F'-graded schools engaged in systematically different changes in instructional policies and practices as a consequence of school accountability pressure, and that these policy changes may explain a significant share of the test score improvements (in some subject areas) associated with 'F'-grade receipt."

Concerns about California

A grade is a snapshot of school effectiveness designed to encourage parents to learn more. This is where states can do much better than they are doing now, by making relevant information accessible through a well-designed school report card that clearly and concisely lays out the calculations used to arrive at the school's letter grade.

ESSA calls on states to provide annual reports, which must include information such as disciplinary data, absenteeism, per-pupil spending, teacher evaluation results, or school surveys. Such a dashboard gives a more complete picture of a school, better preparing parents to ask questions of their child's principal, teachers, or school board members. It also may encourage more parental involvement, and can help them identify schools that best meet their children's needs. The challenge is to create accountability formulas and report cards that communicate these many data points clearly.

One way to impede and prevent accountability is to dilute the importance of academic achievement and cloud the data provided to parents. This is the path that California appears to be taking with its new accountability formula, which abandons a comprehensive, summative performance score in favor of ratings on nine different elements, many of which may or may not have much impact on student success. These include inputs such as parental involvement, school climate, whether instructional materials and school facilities are considered sufficient, and implementation of academic standards. A draft report card under consideration would use colored boxes to indicate school performance on these elements, an approach deemed "practically impossible" to understand by the *Los Angeles Times* in July 2016. "If you're a parent trying to figure out whether one school in your district is better than another, well, there's no clear way to do it."

> An effective accountability system requires that parents have a clear and concise measure of school performance.

This is not transparency. It is a fog machine. Parents will be confronted with a mishmash of confusing and unprioritized data that lead to no conclusion. Principals will spend valuable time trying to comply with criteria that may have little bearing on how their students perform, and may or may not boost student achievement. Hamstringing them with state-dictated criteria distracts them from what should be their primary focus.

One also wonders how California plans to comprehensively identify its lowest-performing schools, as is required by the new federal law. The school grading systems identify schools with grades of "F" for comprehensive support. While these very low-performing schools may not have every indicator in the bottom 5 percent, it is obvious when looking at their data that these are the schools in need of the highest level of support.

Those supporting California's approach mistakenly argue that using only academic indicators puts too much emphasis on test results. But whether a student will succeed after high school and move on to a meaningful career depends primarily on one thing: Is he or she academically prepared for college and a good career?

We still have far to go when it comes to transforming education in our country. But there is much to be learned from the student-centered systems enacted by Florida and similarly minded states. Strengthening and improving accountability systems have proven effective in achieving the results our students, parents, educators, and taxpayers deserve. ■

HOUGH & KIRST
(CONTINUED FROM PAGE 59)

simultaneously, and there is a wide range in how the measures identify schools in the bottom 5 percent compared to one another and compared to the summative academic score. For example, in looking at rates of chronic absenteeism vs. rates of suspensions and expulsions, 90 schools identified as being in the bottom 5 percent of all schools by either measure, yet only 16 percent of those schools are similarly identified by both measures.

ESSA regulations specify that nonacademic measures can-

California's multiple measure, dashboard-style accountability system is focused on providing schools and districts with a variety of data for a more comprehensive picture of a school's successes and challenges.

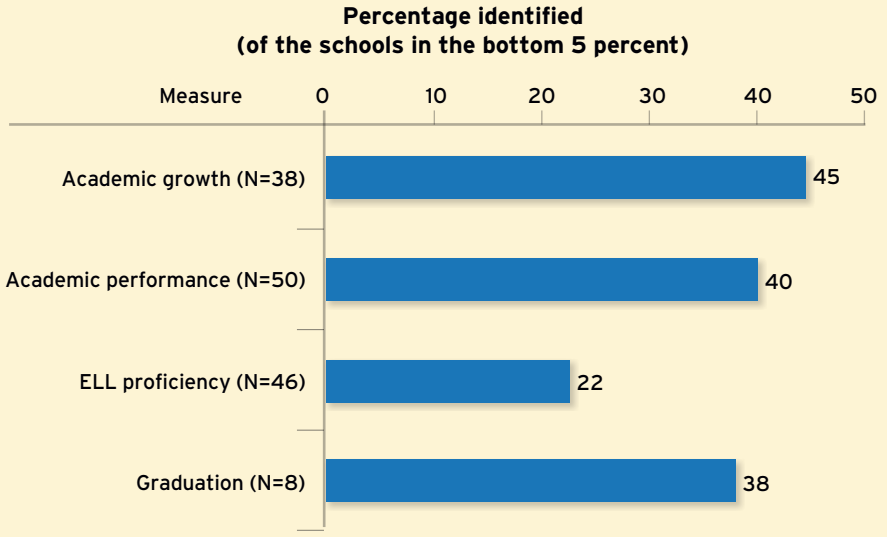
not prevent a school from receiving a CSI designation that would otherwise have been identified using the academic measures. Given the difference between the nonacademic and the academic measures, this effectively forces states to assign SQSS indicators very little weight in a summative score, such that they do not change the identification of schools using the academic measures. We found that an SQSS measure would have to account for less than 1 percent of the summative score to ensure it did not change which schools are identified for CSI. If the SQSS indicators are important signs of school performance, as the law suggests they are, they should be accorded a meaningful weight in the process of identifying schools for support and improvement. This suggests that a summative score is particularly problematic when considering the inclusion of SQSS measures in states' accountability systems.

As an alternative, the PACE report shows how states can identify schools for CSI using a method that considers each indicator rather than aggregating the indicators into a summative score. Using a dashboard of measures, states could use a tiered approach to make a series of decisions about school performance on particular indicators. This would enable states to make judgments about whether or not schools need CSI based on a comprehensive evaluation of all the data. For example, instead of averaging or differently weighting scores on academic performance and academic growth, a state could decide to identify for CSI only schools that have low academic outcomes and are not demonstrating growth. Similarly, of two schools with similar academic achievement, a state could choose to focus limited resources for CSI on a school with poor SQSS outcomes rather than a school with positive SQSS outcomes, since the latter school may be on a road to improvement while the former is not.

It is much like a school counselor trying to decide which students to support with limited resources. Should she focus intensive support to a student with all Ds, or to a student with mostly Cs and Ds, and one F? *(continued on page 62)*

Many Low-Improving Schools Missed with Summative Score (Figure 2)

Only 45 percent of the schools that would be identified for intervention based on a summative score that assigns equal weight to each academic growth indicator scored in the bottom 5 percent for academic growth.



SOURCE: Policy Analysis for California Education (PACE)

HOUGH & KIRST
(CONTINUED FROM PAGE 61)

Her decision reflects a value judgment, and may depend on other characteristics of the student. By the same token, the full information in the multiple measures is more informative than a single number.

The Strength of a Dashboard

There is growing agreement among policymakers, school and district leaders, and researchers that the most important use of school effectiveness measures should be in driving continuous improvement at both the local and state levels. To this end, California is developing a multiple-measures, dashboard-style accountability system that is focused on providing schools and districts with a variety of data, for a more comprehensive picture of a school's successes and challenges. Such detail captures multiple dimensions of school and district performance and can drive local improvement efforts tailored to each unique context. It can also help school and district officials better inform themselves and their communities about how specific programs and services are working to improve student outcomes. Sharing data in this way can simultaneously present a holistic view of how a school is doing and—with thoughtful visualization—highlight a limited number of metrics to focus attention and not be overwhelming to consumers.

In California, we believe parents, as educated consumers and advocates for their children, want to know more about how public schools are performing, and that policymakers should ensure the public has the necessary tools to make good use of multiple measures. Already, many parents intuitively understand that holding schools accountable for performance cannot be reduced to a single number, in the same way that they appreciate that their own students may be doing well in one subject but not another. Parents are familiar with multiple measures when they read their child's report card, and certainly do not want their child's school performance reduced to a single number.

In our view, a dashboard will give parents the information they need to make wise choices about the schools their children attend, rather than misleading them with arbitrary judgments about whether schools are "good" or "bad." Focusing the accountability system on practical and actionable tools for continuous improvement comes with trade-offs in the ability to rank schools, but the benefits more than justify the costs. ■