When COVID-19 forced a hiatus in federally mandated assessments in spring 2020, it interrupted a quarter century of effort to track, disaggregate, and publicize achievement levels at the school level. The aborted school year put a big data gap where 2020 scores should have been. Combined with the already-raging assaults on testing, state education leaders find themselves in a fraught and difficult place.

Arguably since Congress passed the Improving America’s Schools Act (IASA) in 1994 and most definitely since it enacted No Child Left Behind (NCLB) in 2002, test-based school accountability has played a big role in American public education. Although the latest revision of the Elementary and Secondary Education Act, known as ESSA (2015), lets states make more decisions regarding their school accountability, state leaders should stick with their assessments because they improve student learning and school performance.

Chester E. Finn Jr. and Eric A. Hanushek
The Evidence

Despite this inverted structure and the mounting public uproar over the heavy-handed federal role, NCLB actually produced improvements. In both the period of state actions before its enactment and after its passage, accountability systems led to greater student achievement. Moreover, systems that incorporated consequences for schools and personnel had the strongest effect.

International evidence indicates that countries with testing programs that allow for external comparisons have students who do better on international achievement tests. Alternative testing approaches, including school inspectors, do not lead to better overall student performance when they lack the ability to compare performance across students and schools.

The international evidence also highlights a generally overlooked part of assessments—that not all testing is about schools. Many countries develop testing for graduation and student placement into higher education or other institutions. These exit exams are generally designed to do two things: give students incentives to learn and provide direct comparisons of students across local grading systems. The use of such exams has generally shown beneficial effects across countries. Importantly, the value of local school grades rises when there is an external test that can be used to calibrate the meaning of those grades.

It is often said in arguments against NCLB-like reforms that top-ranked countries such as Finland do not use testing and accountability yet rank very highly on international assessments. But Finland has a national exit exam that is mandatory for all students and
that determines access to tertiary education. Performance on the Finnish Matriculation Examination has strong implications for students and schools.6

The United States has also used consequential student exams for purposes other than school accountability. Twenty-four states required exit exams for high school graduation in 2013, with most of their passing scores set at least at the 10th grade level.7 Moreover, a number of states require students to meet minimum test-based standards in order to be promoted to the next grade. Evidence suggests that both of these testing regimes have had somewhat positive effects on achievement, although there are some mixed findings.8 Both regimes have also faced considerable political pressure, leading to slow but discernible movement away from such test use.

The objections raised about test-based accountability do not erase the fact that students learn more when there are measures of performance and when schools pay attention to levels of achievement. Some have argued that it did not work because it did not make U.S. schools the best in the world,9 but that is not a legitimate criterion for evaluating accountability systems.10 Others contend that NCLB narrowed classroom instruction because it only focused on reading and math. Emphasizing basic academic skills was, of course, part of the design because it is hard to argue that basic reading and math skills are not key to most subsequent learning. At the same time, transcript studies at the high school level show that the curriculum of the typical student in fact became broader and more rigorous in academic coursework with the introduction of state and federal accountability.11 It’s a fact, however, that more time and focus on basic academic skills necessarily means less is available for other parts of schools such as the arts, a trade-off that comes from setting priorities in schools.

One common argument against test-based accountability is that it leads to teachers spending too much time teaching to the test. This would chiefly be a problem if tests fail to reflect what we want students to know. In programs such as Advanced Placement and International Baccalaureate, for example, which are highly regarded by most teachers, preparing students to ace the end-of-year exams is typically viewed as honorable and gratifying.

NCLB is often decried for being a "high-stakes" testing regime. Yet the state assessments used in the aggregate accountability systems of most states attach no rewards or penalties to student test takers and typically not to their teachers either. To be sure, exit exams and college admission tests carry high stakes for students, but there is generally less (though tangible) political pushback to high-stakes testing that involves students than to the kind that judge schools and educators.

It is true that the availability of state assessments can lead to larger stakes for teachers, because they facilitate linking the performance of students to specific educators. The possibility of evaluating teachers based on how much their students learn has been recognized for 50 years.12 While it was not part of the original NCLB, Secretary of Education Arne Duncan included teacher evaluation as part of the department’s Race to the Top initiative and in a 2012 "education flexibility" offer to states.13 Continuation of this aspect of accountability—and teacher resistance to it—appears to be the most important pressure for scaling back if not eliminating regular student testing.

The use of student performance to gauge teacher effectiveness requires that the teacher's influence be separated from other factors that bear on student achievement. This separation is generally accomplished through “value-added” analysis. There is a large literature on such analysis, and two generalizations can be made: First, there are significant differences in teacher effectiveness, differences that have huge impacts on students' labor-market prospects and on the U.S. economy.14 Second, nobody would argue for using value-added scores alone to evaluate teachers. Teachers have wider impacts than what can be measured on state math and reading tests. More important, only a minority of teachers—perhaps a quarter—can be evaluated at all in a value-added context because of the limited subject and grade-level testing in the schools. Thus the drumbeat against "evaluating teachers just on narrow tests" appears to be more a general pushback against any evaluation of teachers, not to the overuse of tests.
impossible under most circumstances. That missing data will invalidate key elements of most states’ ESSA accountability plans, remove valuable information from school report cards, and—if the absence of growth data places greater weight on simple achievement data—will give rise to new equity concerns.

Although ESSA is all but certain to remain the law of the land for some time, it is likely that the U.S. Department of Education will see many more requests for waivers from elements of it and that a number of states will consider amending their approved accountability plans. How that may play out depends, of course, on individual state circumstances, education-governance structures, and political dynamics, but some movements in this direction are all but certain. Meanwhile, there obviously had to be a hiatus in the consequential use of end-of-year test information for evaluating schools, teachers, and students, both because there is a big data hole and because almost no students had a full year of proper instruction.

With the resumption of school in fall 2020—in all the complex forms it is taking—states and districts would be well advised to deploy whatever assessments they can tap to gauge where individual children, groups of children, and entire schools and systems are at the year’s start. The cessation of in-person instruction in spring 2020 has obviously produced significant learning gaps but, perhaps as important, has led to much larger variations in the achievement that will be found in most schools during this next round of learning. Data on baseline performance are sorely needed to shape and adapt curriculum and instruction and to align instruction more accurately to needs.

To satisfy formal requirements under their ESSA plans but, more important, to be better able to plan for subsequent years, states and districts should plan to resume their familiar assessment regimen at the end of the coming year. End-of-year data from 2020–21, if available, can be compared with end-of-year data from 2018–19 to calculate two-year changes and to chart the course of what may be a series of interrupted years.

For growth or value-added calculations in particular, a single year of missing data on individual students causes many complications and more than a year will render such calculations impossible under most circumstances. That missing data will invalidate key elements of most states’ ESSA accountability plans, remove valuable information from school report cards, and—if the absence of growth data places greater weight on simple achievement data—will give rise to new equity concerns.

Although ESSA is all but certain to remain the law of the land for some time, it is likely that the U.S. Department of Education will see many more requests for waivers from elements of it and that a number of states will consider amending their approved accountability plans. How that may play out depends, of course, on individual state circumstances, education-governance structures, and political dynamics, but some movements in this direction are all but certain. Meanwhile, there obviously had to be a hiatus in the consequential use of end-of-year test information for evaluating schools, teachers, and students, both because there is a big data hole and because almost no students had a full year of proper instruction.

With the resumption of school in fall 2020—in all the complex forms it is taking—states and districts would be well advised to deploy whatever assessments they can tap to gauge where individual children, groups of children, and entire schools and systems are at the year’s start. The cessation of in-person instruction in spring 2020 has obviously produced significant learning gaps but, perhaps as important, has led to much larger variations in the achievement that will be found in most schools during this next round of learning. Data on baseline performance are sorely needed to shape and adapt curriculum and instruction and to align instruction more accurately to needs.

To satisfy formal requirements under their ESSA plans but, more important, to be better able to plan for subsequent years, states and districts should plan to resume their familiar assessment regimen at the end of the coming year. End-of-year data from 2020–21, if available, can be compared with end-of-year data from 2018–19 to calculate two-year changes and to chart the course of what may be a series of interrupted years.

Meanwhile, many of the familiar assessments, both formative and summative, may need to be adapted to a schooling environment at least some of which takes place outside
the school building for at least some students and at least some of which takes place outside traditional school hours. With careful planning, these testing adaptations can yield sturdy, valid, reliable gauges of student learning with minimal glitches, such as those that beset some who took at-home Advanced Placement exams in May 2020. (The College Board deserves kudos for improvising an assessment arrangement that apparently worked for the vast majority of AP students.)

Perhaps most important for state board leaders to understand and communicate to their constituents is this: Results-based accountability for schools and students is perhaps the most impactful education policy that we have. Few if any other strategies produce such broad improvements in achievement.

All should welcome the quest for additional sources for improving student learning and school performance, as do we. But until and unless such sources prove as stable, reliable, and revealing as testing, American education cannot stop testing. To do so means flying blind, uncertain whether schools are following a flight plan and getting close to the intended destination. Put differently, no enterprise can succeed without regular, reliable data on its own performance. Testing cannot furnish all the data that educators and policymakers need, but it is an essential source of indispensable information.

While having good assessment data cannot ensure gains in student achievement, not having good assessment data can virtually guarantee no improvement.

Testing cannot furnish all the data that educators and policymakers need, but it is an essential source of indispensable information.

—


17. Of course, this environment also calls for many changes in staffing, budgets, curricula, technology, and the delivery of instruction.

Chester E. Finn Jr. is Distinguished Senior Fellow at the Thomas B. Fordham Institute and a senior fellow at the Hoover Institution of Stanford University. Eric A. Hanushek is the Paul and Jean Senior Fellow at the Hoover Institution.