Test-Based Accountability in Distressed Times

State leaders should stick with their assessments because they improve student learning and school performance.

Chester E. Finn Jr. and Eric A. Hanushek

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

systems—particularly when it comes to sanctions, interventions, and consequences—the assessment requirements did not significantly change in 2015 nor did the obligation to disaggregate and publicize school-level results.

Pandemic-induced assessment waivers and a testing holiday will intensify the longstanding pushback against these requirements. They will embolden test critics and sundry education interest groups to declare, "See, we don't really need those damned tests. We can't even use them during unusual circumstances such as at-home study—and when we do use them they distort the curriculum, cheapen instruction, worry teachers, alarm parents, and scare kids. Given the learning gaps and uncertainties presented by the present plague, it would be cruel to go back to using them and whatever results they might show in spring 2021 will surely be misleading."

We disagree. We will offer state boards of education suggestions on how to proceed in the short and the longer run. But first, we review the history of test-based accountability as it has been practiced and the evidence of its impact.

The History

NCLB significantly enlarged the role of the federal government in education. Its focal point was evaluation of school outcomes, and it contained incentives for ensuring high levels of overall achievement and broad impacts across all subgroups defined by race, ethnicity, poverty, and language.

The NCLB feature garnering the most attention was the requirement that all students be proficient in reading and math by 2014. As time passed, it became clear that this extremely ambitious goal was not going to be met. School staff led a simultaneous, steady drumbeat aimed at ending the accountability regime altogether. In time, NCLB became a four-letter word to many. What is often not recognized, however, is that 43 states already had their own test-based accountability system in place at the time the NCLB became law in 2002.

In many ways, NCLB had it backward: States were charged with identifying the learning standards and testing regimes to describe student proficiency while the federal government dictated what actions should be taken if progress was insufficient. The federal government, without knowledge of the demands on or capacities of schools, clearly was unprepared to dictate how all schools should provide education. At the same time, individual states have little way to set standards for the national and international labor markets for which they were preparing students. Moreover, NCLB measured pupil performance by level of achievement without regard for the preparation and readiness with which students came to the schools.

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.[1] Moreover, systems that incorporated consequences for schools and personnel had the strongest effect.

These U.S. data are consistent with international evidence that indicates that countries with testing programs that allow for external comparisons have students who do better on international achievement tests.[2] Alternative testing approaches, including school inspectorates, do not lead to better overall student performance when they lack the ability to compare performance across students and schools.

International evidence indicates that countries with testing programs that allow for external comparisons have students who do better on international achievement tests.

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.[3] 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.[4]

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.[5] 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.

Students learn more when there are measures of performance and when schools pay attention to levels of achievement.

Currently, 34 states require some measure of growth in student achievement to be used in the evaluation of teachers, but this is down from 43 just five years earlier.[15] This decline, however, is not consistent with the research evidence on the effectiveness of personnel systems that use teacher value-added in their operations. Although understandably contentious, teacher value-added has proved to be useful in contributing to the overall evaluation of teachers when such information is available from student testing programs.

Washington, DC, offers the best example of the influence on student learning from a personnel system incorporating teacher value-added. Its IMPACT system uses value-added analysis for the quarter of the teachers where it is possible, but the largest portion of the evaluation of all teachers comes from an objective observational rating. Based on the overall evaluation, the District of Columbia gives large bonuses to effective instructors and dismisses grossly ineffective teachers, and these steps have contributed to the best achievement gains of a major U.S. city.[16] A similar system in Dallas, Texas, also appears to be significantly improving student outcomes, particularly when it is used to guide staffing in schools with concentrations of disadvantaged students. These are just two examples of how improved compensation policies can be usefully introduced into school policy.[17]

In fall 2020, states and districts would be well advised to deploy whatever assessments they can tap.

The Year Ahead

Important technical questions arise when an assessment regime built on an annual cycle—like the familiar school calendar—gets interrupted. Although a simple resumption of testing may still accurately display student achievement at whatever point in time the tests are again given, a year (or more) of missing data will affect trend lines, confound growth calculations, and complicate applications of data derived from them. Psychometricians and testing directors will have their work cut out trying to make the necessary adjustments, adding caveats, and double-checking reliability.

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.[18] 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. ■

[1]Eric A. Hanushek and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24, no. 2 (2005): 297–327; Thomas S. Dee and Brian A. Jacob, "The Impact of No Child Left Behind on Student Achievement," *Journal of Policy Analysis and Management* 30, no. 3 (2011): 418–46; David Figlio and Susanna Loeb, "School Accountability," in Stephen Machin, Eric A. Hanushek, and Ludger Woessmann, eds., *Handbook of the Economics of Education* (Amsterdam: Elsevier, 2011): 383–421.

[2]Annika B. Bergbauer, Eric A. Hanushek, and Ludger Woessmann, "Testing," *NBER Working Paper* 24836 (Cambridge, MA: National Bureau of Economic Research, July 2018).

[3]Eric A. Hanushek and Ludger Woessmann, "The Economics of International Differences in Educational Achievement," in Hanushek et al., *Handbook of the Economics of Education.*

[4]Guido Schwerdt and Ludger Woessmann, "The Information Value of Central School Exams," *Economics of Education Review* 56 (2017): 65–79; Margaret E. Raymond, "The Diploma Dilemma," Hoover Education Success Initiative (Stanford, CA: Hoover Institution, February 2020).

[5]E.g., Pasi Sahlberg, "Education Policies for Raising Student Learning: The Finnish Approach," *Journal of Education Policy* 22, no. 2 (2007/03/01): 147–71; Linda Darling-Hammond, *The Flat World and Education: How America's Commitment to Equity Will Determine Our Future* (New York: Teacher's College Press, 2010).

[6]Matriculation Examination Board (Finland), website, https://www.ylioppilastutkinto.fi/en/.

[7]U.S. Department of Education, *Digest of Education Statistics, 2017*, advanced release ed. (Washington, DC: National Center for Education Statistics, 2018).

[8]John H. Bishop et al., "The Role of End-of-Course Exams and Minimal Competency Exams in Standards-Based Reforms," in Diane Ravitch, ed., *Brookings Papers on Education Policy 2001* (Washington, DC: Brookings, 2001); Guido Schwerdt, Martin R. West, and Marcus A. Winters, "The Effects of Test-Based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida," *Journal of Public Economics* 152 (2017/08/01/): 154–69.

[9]Michael Hout and Stuart W. Elliott, eds., *Incentives and Test-Based Accountability in Education* (Washington, DC: National Academies Press, 2011).

[10]Eric A. Hanushek, "Grinding the Antitesting Ax: More Bias Than Evidence behind NRC Panel's Conclusions," *Education Next* 12, no. 2 (Spring 2012): 49–55.

[11]National Center for Education Statistics. 2011. *America's High School Graduates: Results of the 2009 NAEP high school transcript study.* Washington, DC: U.S. Department of Education.

[12]Eric A. Hanushek, "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," *American Economic Review* 60, no. 2 (May 1971): 280–88.

[13]U.S. Department of Education, "ESEA Flexibility," June 7, 2012, https://www2.ed.gov/policy/eseaflex/approved-requests/flexrequest.doc.

[14]Eric A. Hanushek and Steven G. Rivkin, "The Distribution of Teacher Quality and Implications for Policy," *Annual Review of Economics* 4 (2012): 131–57; Eric A. Hanushek, "The Economic Value of Higher Teacher Quality," *Economics of Education Review* 30, no. 3 (June 2011): 466–79; Raj Chetty, John N. Friedman, and Jonah Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review* 104, no. 9 (September 2014): 2633–79; Eric A. Hanushek and Ludger Woessmann, *The Knowledge Capital of Nations: Education and the Economics of Growth* (Cambridge, MA: MIT Press, 2015).

[15]Elizabeth Ross and Kate Walsh, *State of the States 2019: Teacher and Principal Evaluation Policy* (Washington, DC: National Council on Teacher Quality, 2019).

[16]Thomas S. Dee and James Wyckoff, "Incentives, Selection, and Teacher Performance: Evidence from IMPACT," *Journal of Policy Analysis and Management* 34, no. 2 (Spring 2015): 267–97; Thomas S. Dee and James Wyckoff, "A Lasting Impact: High-Stakes Teacher Evaluations Drive Student Success in Washington, D.C.," *Education Next* 17, no. 4 (Fall 2017): 58–66.

[17]Eric A. Hanushek, "The Unavoidable: Tomorrow's Teacher Compensation," Hoover Education Success Initiative (Stanford, CA: Hoover Institution, January 2020).

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

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

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.