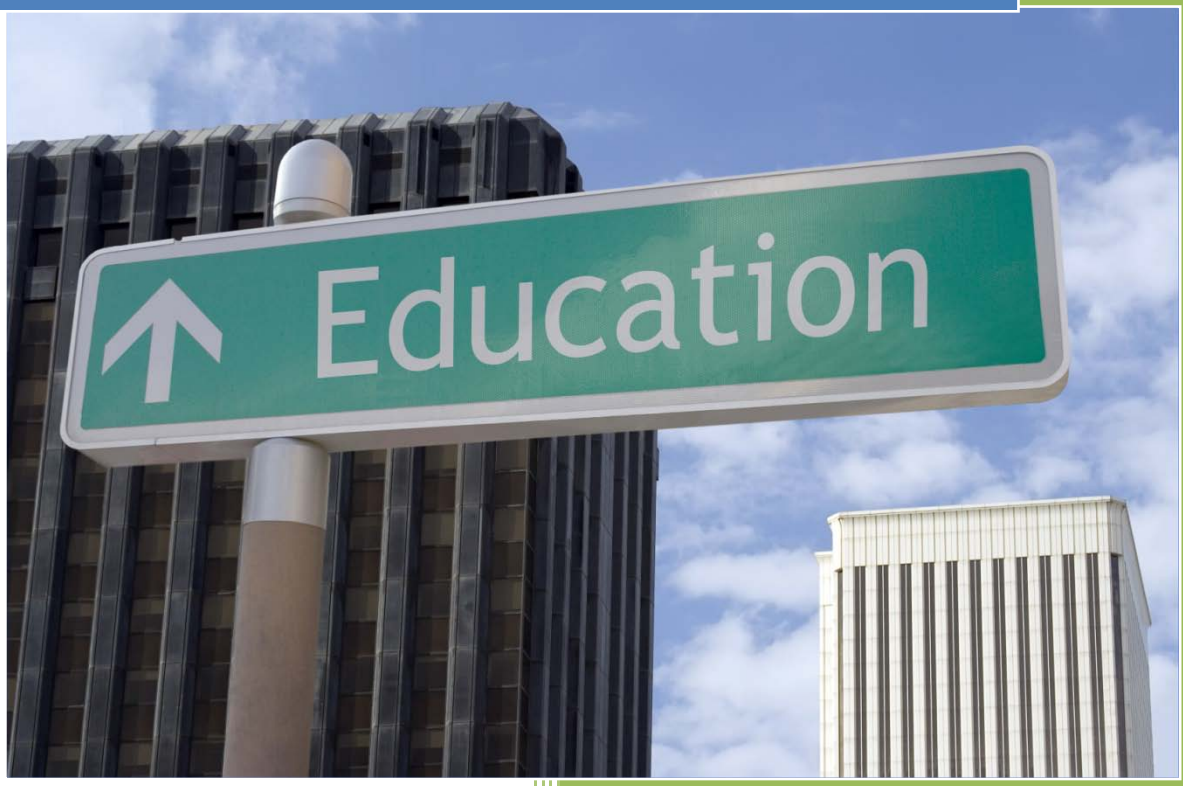
An Assessment by Hoover Institution's Koret Task Force on K–12 Education

# American Education in 2030




## An Evidence-Based World

Eric A. Hanushek

In 2010 there were many questions about testing students, including how the information would be used. Parallel questions asked whether performance on the existing tests even mattered. After all, the test were narrow and did not reflect either deeper thinking skills or other noncognitive facets that research was beginning to identify as important for job performance and participation in society.

Now, in 2030, these issues have been resolved. It has become clear that the performance of students matters. It is also clear that testing can now indicate how individual skills vary across the population. Perhaps most important, schools and teachers can and do now build their instructional programs around the observed results of students.

The 2030 system relies heavily on data–data available to and used by schools, teachers, and parents. One essential building block is systematic information on students' gains in learning as they progress through school. Parents find this tracking of their children's performance useful in working with their children and the schools and in deciding on which schools their children should attend. Teachers also benefit from the regular feedback throughout the school year in formulating learning plans for their students. They also have clear guidelines for what students should be learning from the learning standards and from the test diagnostics they receive. Schools also can use year-end and course assessments to help evaluate both programs and teacher performance.

This information has proved useful in raising student achievement. Performance has gone up both in absolute terms and in relative terms when compared to students in other nations–although such gaps still exist. Bringing about more equality in achievement by race and ethnicity or by income level, however, still remains a challenge.

Our current situation clearly differs dramatically from the No Child Left Behind (NCLB) era and the wars over testing and accountability. It pays to review what changed and why.

## The Nature of Testing

Businesses have long used various measures of their outcomes to decide what was working well and what was not. For instance, businesses use information on sales and revenues to identify which products are being demanded. Combined with information about production costs, the same data can indicate which activities are profitable and which are not. Combined with goals or targets, those data can also be used to evaluate managers and production staff.

The innovation in education near the end of the twentieth century was a shift to regular measurements of student performance as an element of school management, seemingly a natural move paralleling the operations of businesses. But it met with surprising resistance from a variety of quarters. Some of the resistance represented legitimate concerns; others did not.

To understand how those concerns were resolved by 2030, let us begin with some notion of the 2002 testing situation (the point of introduction of NCLB). There had been precedents for NCLB in that it was built on extensive state experience with test-based accountability. At the time of passage, all but a handful of states had already established their own accountability systems, albeit with considerable variation across states.

The proliferation of state accountability systems was based on the popularity of the "standards movement." That idea was simple and powerful: states should specify what they expected students to know (by subject and grade), should build instructional programs around those, should measure the accomplishment of those standards, and should hold schools and teachers accountable for meeting those objectives.

Although logical and appealing, each step presented problems and challenges that coalesced into a national debate once NCLB highlighted the commonality of issues previously defined as state-specific questions. The development of standards was particularly contentious, as various people introduced their ideas of what should be identified, how specific those ideas identified should be, and, relatedly, how they should be taught. Part of the discussion had an ideological component, and part was simply an uncertainty about how to define such goals.

The debates about testing, however, were not as much centered on the underlying standards as on the tests themselves (and the uses to which the tests were put, discussed in the next section). Before NCLB states had used a variety of tests, including off-the-shelf norm-referenced tests and specially constructed criterion-referenced tests. The norm-referenced tests did not link specifically to the standards of any state and provided only relative performance measures when students were compared to a representative sampling of students. NCLB was focused specifically on lower-achieving students, consistent with the federal government's historical role in educating the disadvantaged; its device for bringing about reform was holding schools responsible for ensuring that all students were "proficient" according to state learning standards. By using proficient as a measure (something that sounded like an absolute measure of achievement), the federal statutes surrounding NCLB assumed (or required) criterion-referenced tests that matched state standards because the standards defined what students should know. As a result of federal pressure, each state developed a separate set of tests for accountability purposes (because standards were themselves state-specific).

Several elements entered into developing the tests used for accountability, and the 2002 resolution was not optimal from an educational policy view. To begin with, tests are simply a set of assessment items that try to sample knowledge of the standards—but they do not necessarily perfectly measure all areas of the standards. Completely covering the areas of the standards in both breadth and depth would be enormously time consuming and expensive. To achieve efficiency in testing, therefore, tests are generally developed to be most discriminating at a particular level of performance and less at others as roughly indicated by the proportion of test questions. As a result, the immediate state reaction to the NCLB requirements was to develop tests that were "densest" near the level of proficiency that the state had set and thin at higher levels of performance. Such choices led to the charge that the standard tests were set at an inappropriately low level and thus should be dropped from use.

Another thing that entered into the general testing situation was that individual states were free to choose their own proficiency levels. Proficient has no simple definition, and the term was used very differently across different states, tests, and uses. The exact motivation for each state's choice has never been clear. Some states, such as Massachusetts, set high proficiency levels, perhaps to challenge all its schools. Other states, such as Texas, set low standards but then ratcheted up the requirements, perhaps encouraging schools progressively to improve. Yet others chose low levels and made no effort to adjust them. Still other states lowered their proficiency levels over time, perhaps evincing concern about too many schools "failing" under NCLB's accountability provisions. The varying levels of proficiency caused confusion if not discontent with the overall structure of test-based accountability.

Using standardized tests also bothered some, in that the tests focused on low-level facts whereas what was needed was the development of higher-order thinking and reasoning skills. These higher order skills, so the charge went, could not be assessed with multiple-choice, fill-in-the-bubble instruments.

## The Use of Tests

Nonetheless, the largest issue faced in 2002 and the subsequent years of debate over NCLB was not the tests themselves but the use to which they were put. NCLB, which set a goal of 100 percent proficiency within one schooling generation, set regular milestones for the level of performance of each school's students ("adequate yearly progress"). Schools consistently not meeting those goals faced increasingly stringent sanctions.

The structure of the accountability framework also led to situations in which suburban schools serving kids from well-off families looked systematically better than those serving disadvantaged populations. But it was never clear how much of the difference to attribute to schools and how much to attribute to the poorer preparation that disadvantaged students brought to schools. Comparing similar schools permitted

some better comparisons, but the fundamental uncertainty about interpreting the source of observed differences remained.

The target of having all children proficient by 2014 was also questioned. Specifically, many doubted whether all children, regardless of background or disabilities, could meet the goal of full proficiency in the state-specific standards. To assess intermediate progress, judgments about schools were made on a disaggregated basis for subgroups by race and ethnicity, by special education status, by economic disadvantage, and by English-language learner status. Thus, the goal of 100 percent proficiency appeared even more stringent because it implied equal progress for groups that traditionally had lagged behind.

The accountability information did, however, provide the opportunity to trace student learning over time by linking year-to-year performance for each student. It was also recognized that, in principle, this kind of information could be linked to school programs and to teachers, thereby providing feedback on value added. The primary issue to emerge was political: the teachers' unions did not want test information used to evaluate teachers and resisted developing the appropriate data linkages.

Using accountability tests for management purposes, as opposed to fine-tuning classroom instruction, was often cited as the final drawback. Because accountability testing was generally done near the end of the school year, and because results were frequently unavailable until the summer, such testing was useless to that year's classroom teacher. Although next year's teacher might use such test information to identify learning deficits at the beginning of the school year, it provided no information during the school year in which the test was conducted.

## The Path to Today

The current situation is dramatically different. A variety of forces, emerging around 2010 when initial versions of NCLB were being reconsidered, pushed toward scraping the whole idea. Even though some gains in student performance had been observed, the opposition suggested that the gains were not large enough, that instruction had been adversely affected, that the system was very costly, and so on. Nonetheless, scrapping test-based accountability was not chosen.

A variety of factors preserved the underlying system of testing and accountability, although in the somewhat altered form we see today. Perhaps none of those factors was more important than the recognition that U.S. students were not competitive worldwide. Although that information had been available even before NCLB, with the expansion of international testing around the turn of the century and with the attention testing received around the world, policy makers in the United States began to be increasingly concerned with U.S. of students' performance. For a time, it was argued that performance on those tests really did not matter, but public opinion

shifted against this position as research began to show that such tests were important indicators of economic development. Moreover, many other countries–both those already developed and those then developing–confronted their own testing issues and began increasing their performance on international tests, thus raising the concerns of U.S. policy makers.

No one should doubt, however, the importance of the revelation of student performance to parents that was engendered by NCLB. Those early accountability data, although not warmly welcomed by many school personnel, were embraced by parents who saw them as providing information they had not previously had.

As a result, instead of trying to eliminate testing and accountability, policy makers launched new efforts to improve the design (and impact) of the experiments begun in the states during the twentieth century and subsequently codified nationally in NLCB. Those efforts led to significant changes that, although not yet complete, have shown the power of information and accountability to improve school outcomes.

## The Altered Shape of Testing and Accountability

Central to much of the early "redesign" work was recognition that the testing regimes could be significantly improved. First, the testing was dramatically expanded to a wider range of performance levels through adaptive testing. By using computerized testing that initially sorted individuals into levels according to the difficulty of test questions and then provided in-depth questions at the right level, it became clear that more valid and reliable test assessments were possible. Those new test regimes provide today's high-quality information with less time spent on the testing itself. Second, as computerized testing became ubiquitous, the test items expanded in numbers and quality. No longer was it necessary to have a single test prepared for each year, grade, and subject. Individual students are now given a random selection of test items in each major section of a test, thus eliminating much of the ability and incentive to cheat on the tests along with relieving intense concerns about test security. But an even more important element was the realization that the test bank itself could be made publicly available. By having a sufficiently large and encompassing set of questions, teachers no longer "teach to the test" in the pejorative sense; they now teach to the range of items on the test, which are better vetted for content and accuracy by their public nature. (One by-product of this development was the realization that many attacks on standardized tests were actually quite confused. Standardized tests were now developed that reflected deeper learning and thinking as well as mastery of the basics, showing that standardized was not synonymous with low level or rote.)

Linked to the expanded quality and range of the tests was an expanded concept of accountability. The first versions of accountability focused almost exclusively on basic skills. But with improved testing it was easy to develop reporting and

accountability across a wider range–permitting a commensurate increase in incentives for performance at the top. The focus on proficient or not proficient led to a variety of distortions that raised the potential for teachers' concentrating most attention on kids close to the line as opposed to farther above or below the cutoff. By moving to a more continuous model of accountability, albeit one still weighted heavily on achieving minimal competency, these distortions in incentives have been removed.

It is surprising to many people who do not know its history that NCLB was not based on gains in achievement for individual students. As has been recognized for some time, the current system based on learning growth yields student performance data that are much more closely related to the actions and effectiveness of schools and teachers because differences in entering achievement are taken into account. Simultaneously, the term "proficient" has been dropped from the accountability vocabulary. Used instead is a standard of learning gains that yields high performance for all students (albeit at differing levels for each individual).

Some of the changes that we now take for granted actually occurred more gradually. As the system moved toward more-detailed and clear student performance measures, teachers and other school personnel realized that using those measures in evaluation and reward systems would be valuable to them and their students. Moreover, the pressures on schools from the public and from policy makers led the personnel to be more flexible. As a result, our current system–developed in close concert with school personnel–combines student gains on standardized tests with other direct evaluations including expanded peer evaluations. With everyone in the process having a clear objective, the development and subsequent improvement of personnel performance systems now works more cooperatively and smoothly.

An important cultural change also occurred in the period 2015-2020, when local school districts realized that the test-based accountability system was not the only management device available. As a result, today's local districts find it appropriate and useful to introduce other goals and objectives to ensure that the broader purposes of the schools–developing of students in more than just the tested areas–receive appropriate weight by principals and teachers.

A parallel development is related to accountability; for a while, efforts were actually made to expand the accountability system. Specifically, most observers at the beginning of the twenty-first century noted that regular information on student performance throughout a school year could be used to adjust instruction. By employing formative assessments that measured performance on various blocks of material, a teacher could quickly determine student's comprehension levels and adjust accordingly. Because such instructional programs would be built on assessments related to the same content standards as the accountability system, it seemed reasonable to combine the systems. After years of failed attempts, however, the schools stopped trying to combine into a single system the management of

achievement at the classroom level with the overall accountability system. Today's structure, which some still think is a compromise, uses ongoing feedback to students and teachers through a well-developed formative assessment system that is parallel to the system of annual accountability testing. The difference in the level of information, in the timing of informational needs, in the development of appropriate assessments for the different systems, and in the feedback mechanisms for the two purposes led people to see that developing a single system would be too cumbersome. Moreover, as is obvious in the educational marketplace today, developing instructional management systems has burgeoned into a competitive industry that has pushed development and innovation in instructionally useful ways.

The improved data on student performance, both within the school year (formative) and at the end of the year (summative), now provide schools and districts with the means to evaluate what is and is not working. Thus, the overall testing program introduced the basic data needed for a continuous improvement program in which programs, policies, and personnel are evaluated on the basis of performance. Large school districts have developed the capacity to modify what they are doing to improve performance. Smaller districts, however, have yet to do this effectively, relying instead on the general program evaluations of the larger districts.

## A Somewhat Uneasy Truce

One issue raised by the four decades' discussion of standards and accountability, but present since the forming of the Republic, remains on the table. The states by constitutional construction and by historical development have primacy in education, with the federal government serving a more limited role. This mix was in many ways challenged NCLB, which gave states the role of setting standards, testing, and proficiency levels and the federal government the role of specifying how any remedial actions should be accomplished–such as the use of choice programs or supplemental educational services.

After considerable debate, the historic role of the states in determining how to educate students and what to do if performance is unacceptable was restored. The majority of members of Congress came to realize that the 90,000 schools of the nation could not be effectively run out of Washington. As a result, we now see the federal government offering suggestions on how to improve schools, based on its research efforts, but having removed itself from telling individual schools and districts what actions to take if they are not performing well. (This nonetheless does not eliminate the distrust by some of the actions and capacities of state and local officials.)

Accepting the idea that states and localities should be responsible for developing remedies for deficiencies in student performance has contributed to today's model, in which the federal government concentrates on what student should know. Many recognized that the United States was effectively a single economic market and that the

economic health of the nation depended on the skills of its workers. This realization, present when NCLB was introduced, has fueled a continuing debate over whether there should be national performance standards. By subsidizing the development of consortia to devise standards and by paying for general test development, the federal government is today the de facto leader in defining the skills needed by all citizens. But the debate has not ended. Even though we do have more uniform standards, some states believe that this is too much federal intrusion and have refused to accept them. This position is reinforced by the ongoing disputes about the specifics of standards, the level of rigor, and the like–leading some to question the wisdom of putting all the weight on a single set of national standards.

## The Results

The results in 2030 are mixed. The improved accountability and use of data have improved overall test scores. Even though Finland, Hong Kong, and Canada remain ahead on international math and science tests, the gap for U.S. students has been reduced by half. Those gains represent a remarkable change relative to the stagnation that generally held between 1970 and 2010.

On the other hand, the distributional outcomes have only marginally improved. Although fewer minorities lack basic skills, the gaps with nondisadvantaged students have not closed, even as the schools have improved overall achievement across the spectrum. Thus the policy focus of the day remains on achievement gaps, particularly now that overall achievement has moved in a favorable direction. Those mixed outcomes for the more disadvantaged students also keep the federal government actively developing new educational programs.

---

Eric Hanushek is the Paul and Jean Hanna Senior Fellow at the Hoover Institution, Stanford University and a member of the Institution's Koret Task Force on K–12 Education. He is best known for introducing rigorous economic analysis into educational policy deliberations. He has produced some fifteen books and over 200 scholarly articles. He is chairman of the Executive Committee for the Texas Schools Project at the University of Texas at Dallas, a research associate of the National Bureau of Economic Research, and a member of the Koret Task Force on K-12 Education. He currently serves as chair of the Board of Directors of the National Board for Education Sciences. His newest book, *Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in America's Public Schools,* describes how improved school finance policies can be used to meet our achievement goals.

The preferred citation for this publication is
Eric A. Hanushek, " An Evidence-Based World," in *American Education in 2030* (2010), edited by Chester E. Finn Jr., www.americaneducation2030.com