NONE OF THE ABOVE

A NEW VISION FOR STATE STANDARDIZED TESTING

BY LYNN OLSON AND THOMAS TOCH

MAY 2024

FutureEd
Independent Analysis, Innovative Ideas

About the Authors

Lynn Olson is a FutureEd senior fellow. Thomas Toch is FutureEd's director.

About FutureEd

FutureEd is an independent, solution-oriented think tank at Georgetown University's McCourt School of Public Policy, committed to bringing fresh energy to the causes of excellence, equity, and efficiency in K-12 and higher education. Follow us on Twitter at @FutureEdGU

Usage

NONE OF THE ABOVE

A NEW VISION FOR STATE STANDARDIZED TESTING

TABLE OF CONTENTS

Foreword

Statewide standardized testing has played a central role in education policy for decades, as policymakers have sought to get a clearer picture of how schools are performing and spur them to improve. Yet state tests, required by federal law, have grown increasingly divisive. They've been attacked from many directions for many reasons.

At the heart of the controversy is the fact that stakeholders want the tests to serve two different, equally legitimate, and largely incompatible roles. They want the tests to provide policymakers information on student achievement that's comparable across schools and school districts to hold schools accountable for results. And they want them to give educators and families detailed information to improve instruction and track individual student progress.

The competing priorities have caused high-quality tests developed at substantial cost to be attacked by critics and abandoned by states, spawned new testing initiatives that have struggled to address both roles of testing simultaneously, and paralyzed the national discussion on ways to teach students to higher levels in a post-pandemic era when that work is critical.

The conflict is playing into the hands of opponents of all state testing, who would like to strip the testing provisions from federal law, putting the future of state testing at risk—and with it, testing's vital contributions to instructional improvement, school quality, research, and educational equity.

In this report, Senior Fellow Lynn Olson and I propose a way out of the testing morass. The key, in our view, is decoupling state testing from school accountability systems required by the federal government. That would allow changes to state testing that address many critics' concerns, while ensuring that policymakers continue to have a clear window into how students are performing. And it would allow several promising testing innovations to focus on the important task of helping teachers teach effectively—work that the innovations are best suited to support.

We describe in detail the rationale for a new blueprint for state standardized testing and the blueprint itself. We know, given the intensity of the testing debate, that the third way on testing that we're proposing will draw criticism. That's fine with us. Our goal is to spur a much-needed conversation about how best to resolve what has become a years-long stalemate on a critical issue.

We are grateful to the many colleagues in the education sector who shared their perspectives on state standardized testing for this report; we have listed them in an appendix. FutureEd team members Maureen Kelleher, Bella DiMarco, Molly Breen, and Merry Alderman contributed their editorial and design expertise to the project. And we are grateful to the Walton Family Foundation for funding the report and for its commitment to exploring challenging issues in education policy.

Thomas Toch
Director, FutureEd

Fifty-five years ago, there were no statewide academic standards or statewide testing programs. Public education was largely a black box. Aside from commercially created norm-referenced tests that were only able to report how students were performing against a national sample of their peers, policymakers and taxpayers had few ways to know whether students were learning or if their educational investments were paying off.

That started to change in 1965, with the passage of the federal Elementary and Secondary Education Act (ESEA) as part of the War on Poverty. The new law poured millions of dollars into schools serving students from low-income families and, importantly, required standardized tests to gauge the effects of those funds. Then in 1969, Congress mandated the first federally funded snapshot of student achievement, the National Assessment of Educational Progress, to track national trends in student learning in key subjects over time.

That same year, Michigan launched the first statewide testing program. In the 1970s, as states assumed an increasing share of school funding compared with local governments and policymakers continued to seek insight into what their investments were yielding, state testing continued to expand. The publication of "A Nation at Risk" and other education manifestos in the early 1980s led to a wave of state education reforms and further accelerated the growth in state testing. "A Nation at Risk," drafted by a commission established under the U.S. Department of Education, warned of a "rising tide of mediocrity" in American schools and urged states to adopt achievement tests to measure student performance. By the end of the 1980s, 47 states were operating at least one testing program, up from 39 in 1984.

At the federal level, requirements for state testing also mushroomed through successive reauthorizations of ESEA under both Democratic and Republican administrations. As governors, business leaders, and civil rights advocates called for higher standards and greater accountability in U.S. schools—especially those attended by disadvantaged students—lawmakers gradually increased both the volume of state testing and the consequences for schools that failed to make progress.

Today, statewide standardized testing has become ubiquitous in public education, with some 25 million students assessed annually in reading and math and less frequently in science. State testing helps policymakers know if students are proficient in key subjects, track trends, and allocate resources. Researchers heavily rely on statewide standardized tests to evaluate reforms. The tests let parents know how their children and local schools are performing and highlight longstanding disparities in educational opportunities and outcomes.

Yet by the early 2000s, the bipartisan coalition that once supported statewide testing began to erode. Educators protested during the George W. Bush Administration that the heavy focus on test-based accountability was narrowing the curriculum, lowering standards, centering instruction on test preparation, and incentivizing cheating. Opposition

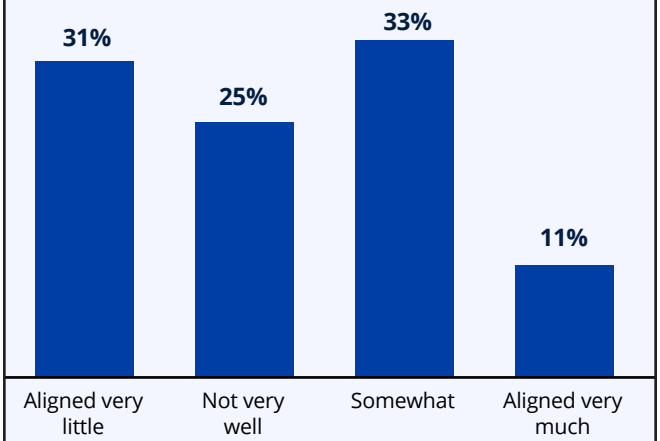
intensified during the Obama Administration, when both Tea Party conservatives opposed to a larger federal role in education and teacher unions wanting to protect their members resisted attempts to tie state tests to the demanding new Common Core State Standards and to teacher evaluations. The suspension of state testing in the first and second years of the pandemic further undermined its standing.

Some critics claim state tests take time away from teaching and learning without contributing enough to instruction and provide results too late in the school year to be useful to educators and families. Others charge the tests are biased against students from diverse racial and ethnic backgrounds and fail to measure important skills beyond academic knowledge.

But the heart of the problem is that different stakeholders want state tests to serve two different, equally legitimate, and largely incompatible roles. The first: providing policymakers information on student achievement that's comparable across schools and school districts to hold schools accountable for results. The second: providing detailed information to educators and families for instructional improvement and individual student progress.

Student Views of Statewide Assessments

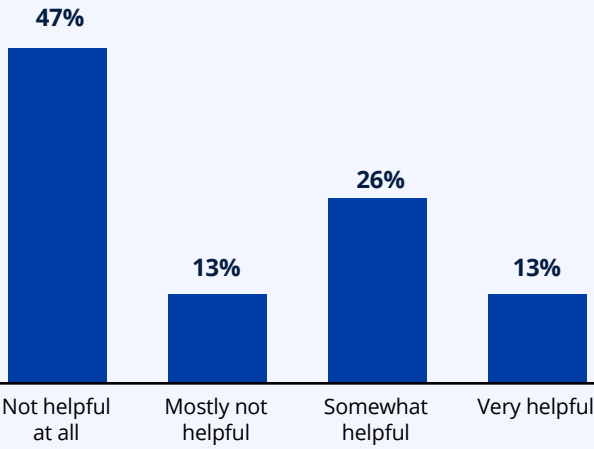
College students' perceptions of the alignment between statewide assessment results and other college-readiness indicators, such as grades and teacher feedback



Source: Education Trust survey of focus group participants, 2021-22

Parent Views of Statewide Assessments

Parents' perceptions of the helpfulness of statewide assessments in keeping their children on track for college



Source: Education Trust survey of focus group participants, 2021-22

In particular, the federal government's requirement to use annual state test scores to hold schools accountable for results, enshrined in the Every Student Succeeds Act of 2015, makes it hard for states to respond to parents' and educators' demands for deeper and more timely measures of students' understanding over the course of the year. That's because the outsize emphasis on state tests to make accountability decisions requires high levels of test security, comparability, and a focus on a single, year-end score. In contrast, assessments that prioritize teaching and learning need to be transparent and intimately connected to what students are studying in the classroom every day. For that reason, it's hard to dramatically improve today's statewide tests without changing the federal accountability requirements that drive their use.

Some states have gone beyond federal law and tied state test results to teacher evaluations, high school graduation, and student promotion decisions. The dominant role of assessments in accountability systems at the federal, state, and

local levels has led many parents and educators to conflate testing with accountability, further undermining support for state tests.

The competing priorities—the production of aggregate information for accountability versus detailed information for improvement—have caused high-quality tests developed at substantial cost to be attacked by critics and abandoned by states, spawned new testing initiatives that have struggled to address both roles of testing simultaneously, and paralyzed the national discussion on ways to teach students to higher levels in a post-pandemic era when that work is critical.

The conflict is playing into the hands of opponents of all state testing, who would like to strip the testing provisions from federal law. Should that happen, it would put the very future of state testing at risk—and with it, vital contributions to instructional improvement, school quality, research, and educational equity.

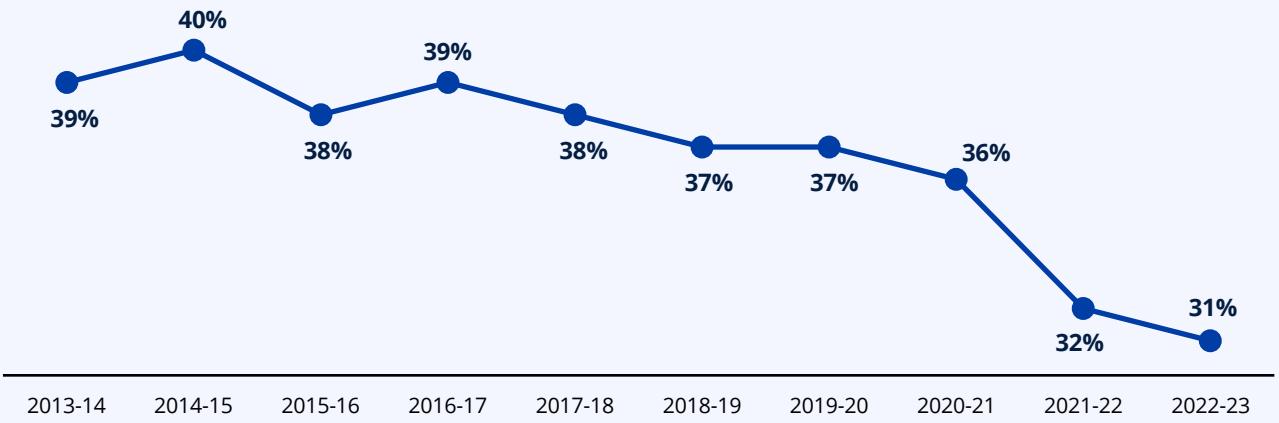
So, what is needed to save statewide testing and the many contributions it makes to education? The answer is a fundamentally different model of assessment that uses state testing to illuminate student and school performance (and drive improvement tied to high standards) but does not carry high-stakes consequences for students and schools. In practice, the Every Student Succeeds Act already began shifting in this direction by removing the specific sanctions for low-performing schools contained in previous law and leaving interventions up to the states, which have done little to address the performance of their weakest schools since ESSA's enactment.

A new, two-tiered testing model would respond to critics' concerns about the time and energy focused on state tests by allowing policymakers to reduce the scale and improve the quality of state tests. And it would permit a second tier of testing improvements designed to better support instruction because those tests would not have to simultaneously support school accountability decisions.

The net result would be less testing and a testing infrastructure with far greater capacity to serve the two core purposes of testing: providing insights into the performance of the public education system, and helping schools communicate individual student performance to parents and improve instruction.

Informed by FutureEd's extensive research and writing on standardized testing and by recent interviews with dozens of testing experts and stakeholders, this report explains why a new model

Percentage of Students Meeting All or Most of the ACT College-Readiness Benchmarks



This chart reflects the percentage of students who met at least three out of the four ACT college-readiness benchmarks (English, Math, Reading, and Science). The benchmarks are determined by the minimum score needed on an ACT subject-area test to indicate a 50% chance of obtaining a B or higher or 75% chance of obtaining a C or higher in the corresponding credit-bearing college courses. Sources: The ACT 2023 National Graduating Class Profile Report; The Condition of College and Career Readiness: National 2018

of state standardized testing is needed and how it would work. It examines how one of the key metrics for gauging national, state, and local academic progress became so controversial and describes testing innovations underway that attempt to address both accountability and instructional improvement, outlining past and potential barriers to their success. Finally, the report describes the changes needed to break out of the current policy paralysis on testing, laying out a blueprint for a two-tiered system of assessment that promotes high standards, greater transparency for parents and policymakers, and a central focus on teaching and learning.

The Rise of State Testing

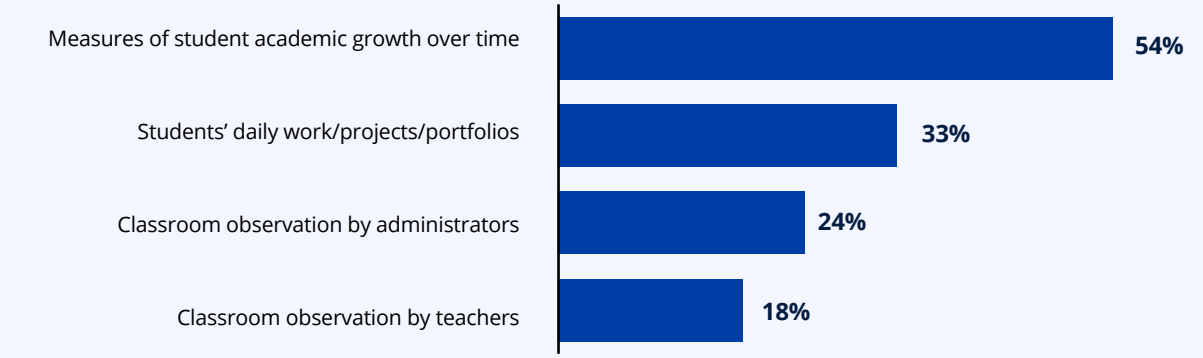
Historically, the federal government's primary role in education has been to provide resources and ensure equitable educational opportunities for underserved students. But in the past 30 years, as the push to raise expectations for American schoolchildren intensified and achievement gaps between student groups based on income, race, and ethnicity remained firmly in place, both Republican and Democratic administrations came to rely more heavily on state testing tied to academic standards as a cudgel to drive improvements rather than simply a monitoring tool.

This shift began with the Clinton Administration's 1994 reauthorization of ESEA, the Improving America's Schools Act. That reauthorization, for the first time, required states to adopt academic standards and develop tests against those standards in at least three grade levels. Significantly, the standards and tests were to be the same for all students, whether they received federal assistance targeted to low-income children or not. As a result of the law, all states (except Iowa) had moved forward in developing state standards and tests aligned with those standards by the year 2000.[1]

Yet the pursuit of equity and excellence in education remained uneven across states.[2] To address what he described as the "soft bigotry of low expectations," President George W. Bush placed significantly more emphasis on test-based accountability in the next reauthorization of ESEA. The No Child Left Behind (NCLB) Act of 2001, which passed with bipartisan support, required annual state testing in reading and math for every student in grades 3 through 8 and once in high school, as well as science tests at three grade spans.

NCLB required states, districts, and schools to publicly report test data by race and income, shining a bright light on long-standing disparities in education. And it set strict timelines for schools to get every student to state-set proficiency levels on the standardized tests. Failure to do so resulted

Percentage of Teachers who Value Test Scores and Other Educator-Evaluation Measures



Source: 2023 Voices from the Classroom: A Survey of America's Educators, Educators for Excellence

in an escalating series of interventions, including the possibility of replacing school staff, permitting families to send their children elsewhere, or closing the school. NCLB effectively tripled the size of the state testing market in the six years after it was passed.[3] It also significantly expanded district use of commercially developed interim and benchmark assessments to measure whether students were on track to do well on state tests.

While NCLB was designed to address educational inequities, states, districts, and schools frequently reacted to the law's pressure in unproductive ways. These included focusing instruction on test preparation, narrowing the curriculum to tested subjects, and even cheating to avoid some of the law's harshest sanctions—all of which generated anti-testing sentiment.[4] Studies found that many state tests focused too much on basic skills at the expense of higher-order thinking. Some states lowered their standards to make it easier for schools to avoid sanctions.[5] The law's focus on proficiency also created an unlevel playing field for schools educating large percentages of low-income students; the schools could be labeled "failing" even if their students showed substantial improvement but remained below the proficiency level. This was true even when students at these schools showed greater relative improvement than students at schools serving the affluent.

The Obama Administration sought to address concerns about the quality of state testing while maintaining the focus on high expectations for all students. Race to the Top, a competitive grant program created in 2009 as part of the American Recovery and Reinvestment Act, provided incentives for states to align their tests with more ambitious college- and career-ready standards. The federal government also funded two assessment consortia—Smarter Balanced and the Partnership for Assessment of Readiness for College and Careers (PARCC)—to encourage states to work together to develop better tests. But the administration also made the competitive grants contingent on states creating new teacher evaluation systems that judged teachers "in significant part" based on students' test results.

The administration's use of federal funding to incentivize states to adopt more rigorous standards, tests and test-based consequences for teachers led to an unlikely alliance between Tea Party conservatives opposed to the federal usurpation of local control and teacher unions opposed to tying job security to tests still under development. Both the National Education Association and the American Federation of Teachers (the latter historically a proponent of higher standards and better assessments) pumped millions of dollars into anti-testing campaigns at the state level and, along with conservatives, encouraged parents and their children to boycott state tests.[6]

This political opposition led many state legislatures to abandon their support of PARCC and Smarter Balanced tests despite studies showing that the tests were of higher quality than many existing state tests.[7] As importantly, it resulted in a watering down of accountability measures in the 2015 reauthorization of ESEA. The federal Every Student Succeeds Act (ESSA) maintained the frequency of state tests tied to college- and career-ready standards but gave states ample autonomy to determine the interventions associated with poor test results, essentially gutting NCLB's strongest improvement measures.

A federal Government Accountability Office report, published in February 2024, found that eight years after ESSA's enactment most states are not implementing even the broad accountability requirements remaining in the law. According to the report, only 42 percent of school-improvement plans for the lowest performing schools address all three elements required by ESSA: a needs assessment, addressing resource inequities, and ensuring interventions are evidence-based. The report also found "wide variation" in the plans.[8]

Yet opposition to state testing among certain groups remains strong, despite the watered-down accountability provisions in federal law. At the 2023 conference of the Network for Public Education, a

harsh critic of test-based accountability that has close ties with the national teacher unions, mentions of statewide testing elicited jeers from the audience.

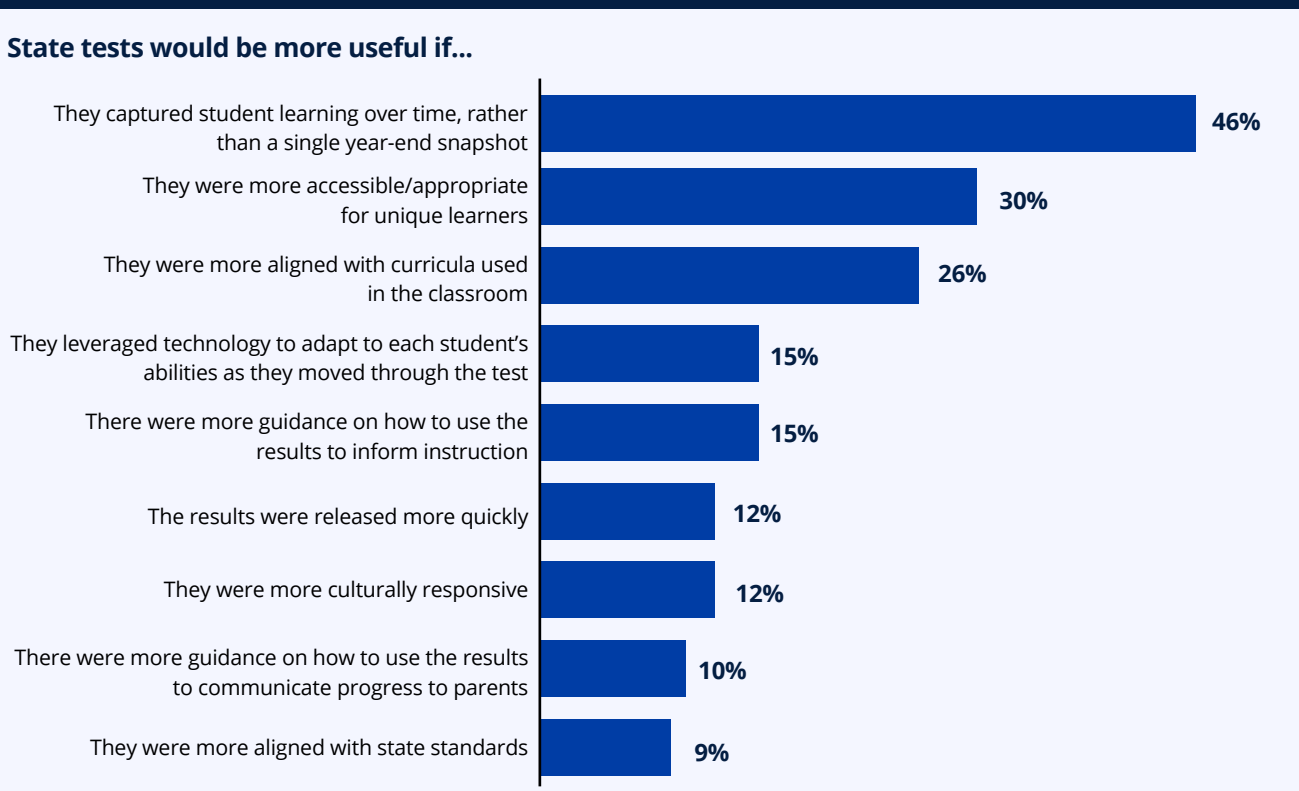
One Test, Many Missions

Some voices in the national testing debate are fundamentally opposed to standardized testing, and their calls for innovation are really code for doing away with state tests altogether. The organization FairTest, for example, has sought for decades to end the "misuse and flaws of standardized testing," including eliminating college admissions tests and test-based accountability systems.

That's not the majority viewpoint. Surveys of parents and teachers have found they value state tests as one measure of student performance. But they'd like the tests to be better in multiple ways, often placing too many expectations on a single instrument.

A national survey of teachers by Educators for Excellence found that 90 percent believe students should have summative measures of their learning from the beginning to the end of the year, and 83 percent believe that teachers should be in part responsible for their students' academic progress. But only 56 percent believe their state reading and math tests accurately measure student mastery of state content standards.[9] Asked what would most improve the usefulness of statewide standardized assessments, 46 percent called for tests to "capture student learning over time, rather than a single snapshot at the end of the year," and 30 percent said they should be "accessible/appropriate for unique learners," such as students with disabilities and those learning English.[10]

Teacher Perspectives on Improving State Standardized Assessments



This chart reflects the percentage of teachers who chose each improvement to state standardized assessments as the most useful. Respondents were asked to select their top two choices. Source: 2024 Voices from the Classroom: A Survey of America's Educators, Educators for Excellence.

National parent surveys by Learning Heroes, a nonprofit that works to increase family engagement in education, have repeatedly found that parents value state tests as one measure of achievement, though they prioritize report card grades and teacher feedback more.[11] National surveys by the National Parents Union have found that parental support for state tests, rather than declining since the pandemic, has increased. "We have more support now for statewide academic testing than we ever had," says Keri Rodrigues, president of the national advocacy group. In October 2020, the organization found that 69 percent of American families wanted to "continue to assess how well students are learning using statewide tests" rather than "take a break from statewide testing."[12] But, Rodrigues adds, at a time when people "expect the instantaneous delivery of results" in other areas of life "parents really don't understand why testing has not kept pace."

An April 2023 report by the National Urban League and UnidosUS—based on focus groups and interviews with 258 Black and Brown students, parents, teachers, out-of-school-time staff, civil rights leaders, policymakers, psychometricians, researchers, and others—illustrates the numerous and sometimes incompatible expectations people now have for state testing.

The report concluded that while statewide standardized tests have helped reveal disparities in academic opportunities by race, income, and language proficiency and have contributed to holding the system accountable for those outcomes, they have not resulted in closing opportunity gaps or targeting resources where they are most needed (perhaps an unrealistic expectation for state tests alone).[13]

The report found that stakeholders want measures that better engage students, emphasize real-world application, and tap into skills beyond academic knowledge, such as social-emotional learning. Such measures include performance assessments, projects, and portfolios of student work, though the report acknowledged that such measures do not yield scores that are comparable across districts and schools. Advocates of more personalized, competency-based instruction are similarly interested in more flexible and holistic measures of student learning that award credit based on demonstrated mastery rather than course completion.

Policy Paralysis

States and test developers have tried to reconcile these competing demands in a variety of ways. But none have found a way out of the paralysis that grips testing policy today.

Many states have made useful technical fixes to the standardized tests that the Every Student Succeeds Act requires them to administer every year. Most now administer their tests online, not with paper and pencil—enabling automated scoring of open-response items, quicker results, more innovative item types (such as computer simulations and drag-and-drop responses) and better accommodations for students who need them. Many have switched from fixed-form to computer-adaptive tests that adjust the difficulty of test questions for individual students based on their prior responses. This can both shorten testing time and better measure a student's actual skill level. There's also evidence that the quality and rigor of state tests have risen, in part thanks to the efforts of PARCC and Smarter Balanced.[14] More recently, standardized test results—from both state tests and NAEP—have shed light on student learning loss during the pandemic.

Other, more fundamental testing reforms have value, but all are challenged by the continued demand to have state tests serve both accountability and instructional purposes.

Through-year assessments, the most common approach that states are experimenting with, seek to meet parents' and educators' desire for more timely, useful information by giving tests multiple times throughout the school year that can both inform instruction and yield a final, summative score. The goal of spreading tests throughout the year is to

have individual tests be shorter, less burdensome, and more closely tied to instruction. But it's not yet clear that through-year assessments will lead to less testing time overall. And it's proving challenging to arrive at a summative score or to ensure that the tests given throughout the year are not viewed as high stakes. (*Read more about through-year assessments here.)*

Performance assessments that ask students to show what they know—for example, by completing an experiment or conducting an analysis rather than just answering multiple-choice questions—have a long history in education. Such tasks can increase student engagement; require them to apply knowledge to real-world problems; give students more choice over the texts, topics, and ways to demonstrate what they know; and better measure deeper learning. Despite continued interest in their use, states have moved away from replacing statewide standardized testing entirely with performance assessments because of their costs, the number of items required to estimate individual achievement, and challenges with reliably scoring student work. (*Read more about performance assessments here.)*

Competency- or skills-based assessments permit students to progress at their own pace based on demonstrated mastery, rather than focusing on end-of-year measures of grade-level knowledge. Skills-based systems also seek to capture a wider variety of skills, from career and technical skills to interpersonal skills, self-management, and digital problem-solving, that traditional state tests do not. Such skills are increasingly sought by employers and are associated with success in life and in the workplace.[15] In 2023, the Educational Testing Service (ETS) teamed up with the Carnegie Foundation for the Advancement of Teaching (CFAT) to build a new suite of skills-based assessments for secondary students over the next several years. These measures will be used only at the local level to help improve instruction initially and won't be used to rate school performance. *(Read more about competency-based assessments here.)*

Each of these measures addresses some of the most prominent critiques of current state tests. But scaling them will likely prove challenging if they are required to meet the same standards for reliability and comparability currently required of state tests used for accountability decisions.

Federal Constraints

If there's a symbol of the barrier federal accountability requirements have created to improved state standardized testing, it's the U.S. Department of Education's Innovative Assessment Demonstration Authority (IADA).

Created as part of ESSA, IADA was designed to provide flexibility for up to seven states to pilot more novel approaches to statewide standardized testing that could be used as part of accountability systems. But so few states applied for IADA—only five have received approval in the past nine years, and two subsequently withdrew—that last November U.S. Secretary of Education Miguel A. Cardona announced that he was relaxing the requirements of the assessment pilot to encourage more states to make use of it. "[W]e cannot expect innovation from the field of education while protecting the status quo from Washington, D.C.," Cardona wrote to the states.[16]

As Nicholas Munyan-Penney, who tracks IADA for the policy and advocacy organization Education Trust, observes, many states say they're reluctant to seek IADA support because the program's application requirements are cumbersome, approval is iffy, and there's no funding attached if they're successful. Some state testing officials go further, saying they doubt the federal government really wants them to innovate.

But the larger reason behind such scant innovation is that states and test developers are hemmed in both by ESSA accountability requirements and by stakeholders' competing demands of state testing.

ESSA's expectation that states test reading and math for every student in grades 3 through 8 every year,

for example, has contributed to both the length and volume of state testing. "Volume is a killer," says Scott Marion, executive director of the non-profit Center for Assessment and a former state testing director. "The number of tests that states have to administer constrains a lot of things because it costs so much money."

Equally challenging is ESSA's requirement that state tests generate sub-scores and diagnostic information for every student tested. It is not realistic to expect end-of-year state tests to yield diagnostic results at a grain size meaningful to teachers—notwithstanding the fact that teachers often receive the results long after the end of the school year. "It's a well-intended requirement to say, 'Let's get every bit of information out of this test that we could possibly get,'" notes Aneesha Badrinarayan, the director of state performance assessment initiatives for the Learning Policy Institute, a California-based research and policy center. But, she adds, "The requirement for sub-scores leads people to make silly decisions about their test design." ESSA's requirement, in other words, is simply not practical.

Lorrie Shepard, a testing expert at the University of Colorado at Boulder, goes so far as to say that states should "stop lying to parents" that state test scores are diagnostic. If the requirement were dropped, she argues, states could use a strategy called matrix sampling to test individual students on just a subset of grade-level standards in more depth while still yielding state, district, and school-level data. Matrix sampling would allow states to test a wider range of curriculum content, she adds, because every student would not have to answer every question.

Alison Timberlake, deputy director for assessment and accountability in the Georgia Department of Education, notes that given the widespread emergence of tests woven into instructional materials and designed to deliver diagnostics, ESSA's expectation that states yield diagnostics by "[testing] every kid every year on the full depth and breadth of the standards … isn't necessary anymore."

Yet, the law's heavy reliance on test scores to rate schools—an ESSA provision that accountability advocates strongly support—reinforces the need for annual testing of every student. In contrast to its predecessor, the No Child Left Behind Act, ESSA requires states to judge schools in part on how much they contribute to students' academic growth, not solely on test scores in a given year. Many rightly view this as a fairer way to judge schools, particularly those with many students who start the year far below grade level. But most states now measure growth using models that compare current and prior year test scores for similar students, which requires testing every student every year.

Uncertain Alternatives

In the absence of alternatives to ensure that students, especially historically underserved students, are not falling through the cracks, civil rights, equity, and parent advocacy groups have been reluctant to abandon federal requirements that states test every student in grades 3 through 8 in reading and math every year, report separate scores for various groups of students, and hold schools accountable for those results. "I don't want to give up on 3 to 8 testing until we have a roadmap or indicators that are going to produce the information that we need to be able to drive resources, or drive diagnostics, in a way that meets the needs of different levels of the system," says Education Trust President Denise Forte, a leading voice on school accountability who helped draft ESSA.

Annual state tests often provide the only standards-based information parents receive about whether their child is performing at grade level. They serve as an important check on classroom grades, which studies have found to be inflated and often poorly aligned with test score outcomes.[17] Accountability advocates also worry, perhaps ironically, that if states no longer provide individualized scores for parents, it will further diminish support for state testing.

All of this has made it difficult to break the stalemated debate on the future of state testing.

Not surprisingly, state testing directors and vendors have largely kept their heads down in the absence of a clear mandate for change. "The best thing for a state test director is quiet," says Scott Marion of the Center for Assessment, adding that when innovations fail—such as when the shift to online testing led to large-scale snafus in a handful of states—it threatens the credibility of the entire testing enterprise.

A Two-Tiered System

Today's fraught state testing landscape argues for a system that continues to use state tests to give policymakers, parents and the public a window into school and student performance, while decoupling test results from consequences for schools and school systems. Ironically, this system has effectively been in place since ESSA devolved accountability decisions to states and the states largely declined to respond decisively to low-performing schools.

Focusing state testing on more fully illuminating educational performance could diminish much of the current opposition to state tests. At the same time, creating a related system of innovative assessments that help teachers support students throughout the school year would begin to disentangle the multiple purposes of assessments and get us past today's testing gridlock.

Innovation starts with having a realistic national conversation about how much a single test can accomplish rather than searching for a unicorn assessment that can be all things to all people. Asking one measure to produce both fine-grained instructional information in real time and large-grained accountability results for schools and districts means that it will never do either effectively.

Moving away from current, test-heavy accountability systems would allow more consideration of other insights into student experiences and outcomes that many stakeholders in the testing debate want. For example, do schools create a sense of belonging and well-being among students? Do students have access to high-quality instructional materials in the hands of expert teachers? This shift could result in a system of assessments and accountability much more tightly focused on supporting high-quality teaching and learning. And it would permit the use of innovative measures—such as student surveys—that struggle to meet the technical requirements for reliability that high-stakes decisions demand.

A more balanced system of assessments also could minimize the footprint of statewide summative tests in exchange for better, standards-aligned formative assessments. In return for shorter, statewide tests that produce summative results—but not diagnostic information for individual students—states might be required to submit detailed plans for how they will support districts' use of standards-aligned formative and diagnostic assessments that meet key quality criteria. These plans could lay out how states promise to help districts produce more timely and instructionally useful data that do not count for high stakes, such as Indiana is trying to do.

Intentionally developing a system of state and local tests that share the same view of student learning could create a more coherent system of assessments than the current tangle of state and local measures. The premise is that local and school-level transparency would drive behavior and the allocation of resources rather than high-stakes consequences.

State tests would focus on aggregate data for policymakers and education leaders to monitor educational opportunities and indirectly support instruction by providing resources for high-quality curriculum and professional learning. Local and classroom-based assessments, i.e., second tier assessments, would provide the fine-grained, timely information needed by teachers, students, and families.

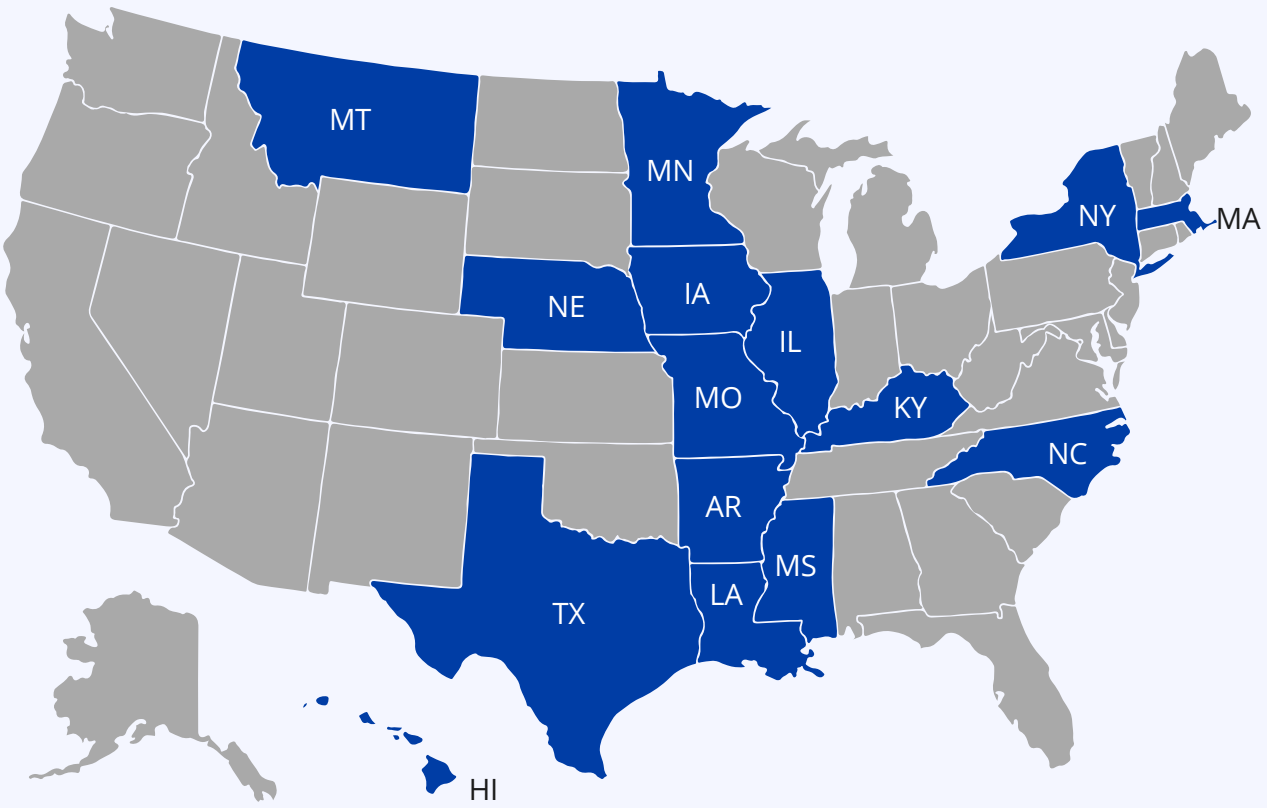
States that want the flexibility to design competency- or skills-based measures in place of existing state

tests could be required to show how they will build educators' capacity to shift to a competency-based model of instruction that allows students to progress based on mastery rather than seat time and how they will help students and families understand and use such measures.

In addition, states, districts, and schools could be required to make all interim, benchmark, and diagnostic test results routinely available to families in formats they can understand, including why and how such data should be used. Right now, teachers have reams of data at their fingertips that parents often know nothing about. This would include requirements that states provide all educators with the time and resources to meaningfully use such measures, aligned to high-quality standards and instructional materials, and to share those results with families.

The shift away from the existing use of state tests with their heavy focus on accountability does not eliminate the role of the states in signaling what high-quality teaching and learning looks like. By incorporating computer-based simulations into their science/technology assessments, for example, Massachusetts is modeling a focus on scientific practices and applied learning, not just factual knowledge. Lorrie Shepard of the University of Colorado at Boulder says states could create sample curriculum units with embedded assessments for districts to use for instructional purposes, assessments that align with state standards and mirror the performance tasks on state summative

States Receiving Funding Under the Federal Competitive Grants for State Assessments Program, 2019 to 2022



The U.S. Department of Education launched the Competitive Grants for State Assessments program in 2019 to promote innovation in state testing. The program replaced the Enhanced Assessment Grants program. Source: U.S. Department of Education

tests. The sample units might signal important trends in high-quality instruction, such as encouraging more student discourse in mathematics or closer analysis of complex texts.

The federal government also could revisit the idea of student or matrix sampling to reduce the footprint of state tests. This could include measures such as testing students on a sample or subset of grade-level content every year instead of testing all grade-level standards; census testing of students in key subjects in alternate grades or every other year; and matrix sampling of some state standards, with each student answering only a subset of test questions that together address all the standards.

According to Marianne Perie, director of assessment research and innovation at WestEd, a nonprofit research, development, and consulting company, when Obama Administration officials were designing the request for proposals for what would eventually become PARCC and Smarter Balanced, they asked psychometricians and researchers what changes were needed to improve the quality of state testing. "I think to a person we all said, 'stop testing every single student every single year'," Perie says. "Go to a sampling approach for school accountability and then, maybe, in key grades assess every kid every year. This over-testing, nobody wants that."

Many of these changes would require changes to federal law. And while the Elementary and Secondary Education Act is unlikely to be reauthorized any time soon, it's not too early to start such conversations and to prioritize both short- and long-term goals, particularly given the rapid advances in AI. (See sidebar on page 14.) Moving to a new, two-tier system of assessments now would set the stage for ESEA reauthorization and undermine the case of testing opponents to abandon federal testing requirements altogether.

The U.S. Department of Education's expansion of the federal assessment pilot program is a step in the right direction. The department should go further by supporting state testing experiments that rely on greater transparency for families to promote improvement. John Bailey, a senior fellow at the American Enterprise Institute, suggests that in exchange for letting some proportion of a state's schools—say 10 percent—try something significantly different, the federal government could double down on transparency. "Score reports are going to become more robust. I think states and districts would gladly take that tradeoff."

States will need more money to support innovation, along with a clear agenda for short- and long-term priorities. An important source of funds is the Competitive Grants for State Assessments program, a federal testing initiative that provides money to states to support the development of innovative measures of student achievement, including performance and technology-based assessments, computer-adaptive tests, projects, and extended performance tasks.

The federal government could also provide funding from the Institute for Education Sciences and the National Science Foundation, as well as encourage states to work together in consortia, as happened during the design of alternate assessments for students with special needs or those learning English. States' commitment to innovation would necessarily include a clearer set of strategies across national funders and a framework to evaluate new models (e.g., through-course assessments) before scaling them. "Could we have a joint agenda on assessment and measurement?" asks Bethany Little, managing principal at Education Counsel who formerly served in education policy roles in the Clinton White House and on Capitol Hill. "Could we agree that, over the next 10 years, these are the things we're going to go after?"

A new, coherently designed two-tier assessment system would give something to nearly everyone in the testing debate, providing the foundation for a testing reform agenda that many stakeholders could get behind and resolving the paralysis gripping standardized testing.

And it's not a new notion. More than 20 years ago, in a report titled "Knowing What Students Know," the National Academy of Sciences called for systems of "balanced assessments"—local, state, and national testing systems with shared expectations for students that provide multiple sources of evidence, systems rooted in a common model of teaching and learning supporting education decision-making at different levels.[18]

This year, the National Academy released a new report, "Reimagining Balanced Assessment Systems," that examines both the continued barriers to and potential for creating such systems. "To work synergistically," the academy argues, "assessments at different levels of the educational system must be compatible, although different in grain size and specificity."[19]

There has never been a large, vocal constituency for standardized tests, however valuable comparable measures of student performance in math, reading, and other essential skills may be to an educational enterprise as vast and decentralized at the American public education system. Today, state testing systems are newly vulnerable: as critics seek to dismantle them, these overburdened systems, expected to fill myriad functions and filling none of them well, are largely incapable of significant change. It's time to build a new system of assessments in public education envisioned by the experts at the National Academy, a system with a narrower, more manageable role for state tests. Liberating state tests from the responsibility to hold schools accountable for student results is the place to start.

CAN ARTIFICIAL INTELLIGENCE TRANSFORM STANDARDIZED TESTING?

Many in the education field hold out hope that generative artificial intelligence can dramatically improve large-scale standardized assessments, including enabling more efficient test development, more actionable and automated test scoring and reporting, and an improved student experience.

Duolingo, a popular online platform for learning languages, has been a leader in this regard. For years, the only way to generate test items was for teams of expert test developers to research and write them and then field test them with large, representative cohorts of students. The Duolingo English test, an online, computer-adaptive exam used by people in over 225 countries and regions to certify their English skills, is the first high-stakes test created using AI and machine learning.

Duolingo has used the technology to create thousands of test items automatically, assess the language ability required for each item, grade them, and produce a final score. Human experts edit and review the items for accuracy, fairness, and potential bias. Scores from the Duolingo English Test correlate highly with two other traditional English tests, including the Test of English as a Foreign Language and the International English Language Testing System, at a much lower cost. The most recent tests include open-ended speaking and writing exercises, including interactive items that simulate a real-time conversation with one of the Duolingo characters.[1] The company also published "Responsible AI Standards" for assessment to help ensure transparency, fairness, privacy, and security in AI-generated testing.[2]

States and test publishers are taking notice. New Meridian already has working proofs of concept to use AI to generate test items and forms, which would then go through its human review process. "We're anticipating at least 30 percent efficiencies leveraging AI for some of the item content and test development work," says Arthur VanderVeen, president of New Meridian.

The Hawaii Department of Education has issued a request for proposals to leverage AI technology to streamline the test development process by simulating responses to field test items from a representative student population. This innovative approach aims to reduce the reliance on traditional time-consuming and expensive human field testing. "We're still just learning about AI capabilities and the extent to which it's generating results that are reliable," says Brian Reiter, assessment administrator for the Hawaii Department of Education.

AI also has the potential to generate score reports that are more comprehensible and actionable for students, parents, and teachers. John Bailey, a senior fellow at the American Enterprise Institute, recently used GPT-4, a powerful machine-learning model, to rewrite New York's score report for parent in simpler language. "Everything about it was ten times better," he says. Bailey adds that AI also could generate evidence-based suggestions for teachers about what to do next based on a student's test score results. Imagine a world, he says, where teachers could get back their students' test scores and be prompted to ask what to try next based on guidance from the federal What Works Clearinghouse.

Lack of Imagination

"The biggest issue I see right now is a lack of imagination because people don't know what's possible," Bailey says. "I worry we're not thinking creatively enough because we haven't exposed

people to what's current, much less if you follow the innovation path of what's going to be available a year or two from now."

AI also could be used to improve the student experience. For example, AI might find ways to automatically score student projects or hand-written papers and provide students with feedback and questions to guide improvements. Voice technology already can analyze students' reading fluency scores. Image-based scoring could be used to assess the schematics that students produce in science classes.

But the potential must be balanced against caution, particularly when tests are used for accountability decisions. "These generative AI tools are probabilistic, so they are not going to give every student the same response every time" says Kristen DiCerbo, chief learning officer at Khan Academy, which offers AI-powered tutoring for students. "Standardization becomes quite difficult. If we're attaching a lot of stakes, it's not going to be clear whether one student got the same experiences as different students and whether those experiences are comparable to each other."

Also, because generative AI creates items using large data sets of prior information, says Tony Alpert, executive director of the Smarter Balanced system of assessments, "we have to be careful about the extent to which AI is trained across a diversity of information that best represents the diversity of student populations," so that it does not introduce bias.

The biggest barrier to AI in education may be human skepticism. Automated essay scoring has been used for years and has proven as reliable as human scorers, yet the National Council of Teachers of English still has a 2013 statement on its website "that basically says computers cannot score essays because they cannot do what the teacher does," notes DiCerbo. "The challenge there isn't technical as much as it's hearts and minds."

To build trust in AI, Smarter Balanced and IBM Consulting in January 2024 announced a new initiative to develop design principles and guidelines for the use of AI in educational assessments. As part of the initiative, the two organizations will create a multi-disciplinary advisory panel and design a proof-of-concept that uses generative AI to translate Smarter Balanced math assessments into other languages more efficiently and with better quality.

"What would it take for people to be confident that our use of AI was responsible?" asks Alpert. "We agreed that the first and most important step was to have a set of design principles to guide implementation and decision making about whether or not to use AI, and then designing and implementing an AI solution."

---

1 Sophie Wodzak, "The Duolingo English Test: A year in Review," 2024 and "Can a Standardized Test Actually Write Itself?" April 6, 2022, Pittsburgh: Duolingo. Also see Burr Settles, Geoffrey T. LaFlair and Masato Hagiwara, "Machine Learning-Driven Language Assessment," Transactions of the Association for Computational Linguistics, vol. 8, pp. 247-263, 2020, https://doi.org/10.1162/tacL_a_00310.

2 Jill Bernstein et al., "Responsible AI Standards for Assessment," 2022, Pittsburgh: Duolingo.

INTERVIEWS

Tony Alpert, executive director, Smarter Balanced

Elsie Arntzen, state superintendent, Montana Department of Public Instruction

John Bailey, senior fellow, American Enterprise Institute

Aneesha Badrinarayan, director of state performance assessment initiatives, Learning Policy Institute

Derek Briggs, professor in research and evaluation methodology, University of Colorado at Boulder

Nathan Dadey, senior associate, Center for Assessment

Kristen DiCerbo, chief learning officer, Khan Academy

Carla Evans, senior associate, Center for Assessment

Denise Forte, president and CEO, Education Trust

Cheryl Harmon, chief academic officer, CenterPoint Education Solutions

Andrew Ho, professor of education, Harvard Graduate School of Education

Bibb Hubbard, founder and president, Learning Heroes

Sue Krause, interim chief executive officer, CenterPoint Education Solutions

Bethany Little, managing principal, Education Counsel

Lei Liu, director, Educational Testing Service

Lydia Liu, associate vice president, Educational Testing Service

Scott Marion, executive director, Center for Assessment

Julie Murgel, chief operating officer, Montana Department of Public Instruction

Nicholas Munyan-Penney, assistant director of P-12 policy, Education Trust

Teresa M. Ober, research scientist, Educational Testing Service

Marianne Perie, senior research director of assessment and accountability, WestEd

Sarah Quesen, director of research and innovation, WestEd

Brian Reiter, assessment administrator, Hawaii Department of Education

Keri Rodrigues, president, National Parents Union

Lynn Schemel, chief academic officer, Indiana Department of Education

Lorrie Shepard, distinguished professor of research and evaluation methodology, University of Colorado at Boulder

Laura Slover, managing director Skills for the Future Initiative, Educational Testing Service

Ethan Stone, CEO and co-founder, Educators for Excellence

Allison Timberlake, deputy superintendent for assessment and accountability, Georgia Department of Education

Arthur VanderVeen, president, New Meridian

Ross Wiener, vice president and executive director education & society program, The Aspen Institute

ENDNOTES

1 Robert Rothman, Something in Common: The Common Core Standards and the Next Chapter in American Education, 2011, Cambridge, MA: Harvard Education Press.

2 ibid.

3 Thomas Toch, "Margins of Error: The Education Testing Industry in the No Child Left Behind Era," 2006, Washington, D.C.: Education Sector.

4 See, for example, Jane Hannaway and Laura Hamilton, "Performance-Based Accountability Policies: Implications for School and Classroom Practices," 2008, Washington DC: Urban Institute and RAND Corporation; Joseph J. Pedulla et al., "Perceived Effects of State-Mandated Testing Programs on Teaching and Learning: Findings from a National Survey of Teachers," 2003, Chestnut Hill, MA: National Board on Educational Testing and Public Policy; and Tamara Wilder, Rebecca Jacobsen, and Richard Rothstein, Grading Education: Getting Accountability Right, 2008, New York: Teachers College Press.

5 Victor Bandeira de Mello et al., "Mapping State Proficiency Standards on NAEP Scales: 2005-2007," October 2009, Washington DC: Institute for Education Sciences, National Center for Education Statistics.

6 Lynn Olson and Craig Jerald, "The Big Test: The Future of Statewide Standardized Assessments," April 2020, Washington DC: FutureEd, Georgetown University.

7 Lynn Olson, "The New Testing Landscape: How State Assessments Are Changing Under the Federal Every Student Succeeds Act," September 2019, Washington DC: FutureEd, Georgetown University.

8 United States Government Accountability Office, "K-12 Education: Education Could Enhance Oversight of School Improvement Activities," January 2024, Washington DC: US Government Accountability Office, https://www.gao.gov/assets/d24105648.pdf

9 Educators For Excellence, "Voices from the Classroom: A Survey of America's Educators," 2023, New York, NY: Educators for Excellence.

10 Educators for Excellence, "Survey of America's Educators, 2024, New York, NY: Educators for Excellence.

11 Gallup and Learning Heroes, "B-Flation: How Good Grades Can Sideline Parents," 2023, Washington DC: Gallup.

12 Echelon Insights and National Parents Union, "National Parents Union Survey of K-12 Public School Parents," October 2020, Boston, MA: National Parents Union.

13 National Urban League and UnidosUS, "Education Assessment, Accountability & Equity: 2022 Phase 1: The Final Report," 2023, Washington DC: National Urban League and UnidosUS.

14 Lynn Olson, "The New Testing Landscape: How State Assessments Are Changing Under the Federal Every Student Succeeds Act," September 2019, Washington DC: FutureEd, Georgetown University.

15 Ou Lydia Liu, et al., "A New Vision for Skills-Based Assessment," 2023, Princeton, NJ: Educational Testing Service.

16 Miguel A. Cardona, U.S. Secretary of Education, Letter to Chief State School Officers, November 20, 2023, Washington D.C.: U.S. Department of Education.

17 TNTP, EdNavigator, and Learning Heroes, "False Signals: How Pandemic-Era Grades Mislead Families and Threaten Student Learning," 2023, New York, NY: TNTP.

18 National Research Council, "Knowing What Students Know: The Science and Design of Educational Assessment, 2001, Committee on the Foundations of Assessment, James W. Pellegrino, Naomi W., Chudowsky, and Robert Glaser, editors, Washington, DC: National Academy Press.

19 National Research Council, "Reimaging Balanced Assessment Systems," 2024, Scott F. Marion, James W. Pellegrino, and Amy I. Berman, editors, Washington, DC: National Academy Press.

NONE OF THE ABOVE

A NEW VISION FOR STATE STANDARDIZED TESTING

FutureEd
Independent Analysis, Innovative Ideas