Which of the Following Approaches to State Testing Works for U.S. Schools?

Choose the answer that best addresses student learning loss

EDUCATORS AND POLICYMAKERS AGREE that state standardized testing needs improvement. Student scores had been slipping for nearly a decade even before the Covid-19 school closures generated unprecedented drops in student learning. Apart from a pandemic-induced pause in spring 2020, state testing systems remained in place throughout that stretch—but failed to halt the decline. Now policymakers and school leaders are wrestling not only with recovery efforts but also with new questions about the practicality and effectiveness of our current approach to testing. Is it time to relax the federal requirement that schools test each student annually in grades 3–8 and once in high school? Or are there other ways to ensure state testing systems support gains in student performance?

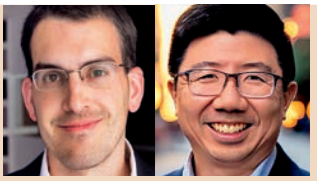
In this forum, Lynn Olson and Thomas Toch from Georgetown University's FutureEd advocate for a sampling approach to testing that would give states valuable insight on aggregate performance without overburdening teachers and students. Chad Aldeman, founder of the phonics program Read Not Guess, and Dale Chu, senior visiting fellow at the Fordham Institute, argue for innovation in state testing to enhance speed and reduce costs while still reporting on the performance of individual students.

Improve the Quality, Preserve the Mission of Statewide Testing

BY LYNN OLSON AND THOMAS TOCH

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 to promote improvement. But support for state testing has been steadily eroding. If testing advocates hope to preserve state testing and its many benefits, it's time for policymakers to rethink the role of the tests, including the possibility of abandoning the federal requirement that every state use test results to identify schools for improvement.

CONTINUED ON PAGE 64



Testing All Kids Serves Students, Parents, and Teachers

BY CHAD ALDEMAN AND DALE CHU

IF YOU FOLLOWED THE 2024 ELECTIONS, you were probably inundated with polling data. Experts ask a random sample of individuals who they plan to vote for, and if the sample is truly random, it doesn't take all that many respondents to get a fairly accurate representation of public opinion.

Of course, as we saw in 2024, polls can be wrong. They are susceptible to sampling error, and results can vary based on *who* exactly is answering the questions. We also wouldn't

CONTINUED ON PAGE 65

MICHAEL WARAKSA

OLSON & TOCH CONTINUED FROM PAGE 62

State tests have been attacked from many directions and for many reasons. They're a time sink, critics charge. They encourage schools to prioritize low-level skills, test prep, and tested subjects at the expense of a richer curriculum, and they cause teachers to focus disproportionately on "bubble students"—those who are close to testing at proficiency. They prompt school districts to clog the calendar with additional tests to gauge students' readiness, and they fail to help teachers in their classrooms.

At the heart of these and other indictments lies the fact that different stakeholders want state tests to serve two distinct, equally legitimate, and largely incompatible roles. Some want the tests to provide policymakers with information on student achievement that's comparable across schools and school districts, with the goal of holding schools accountable for results. Others want the tests to give educators and families detailed information to improve instruction and monitor individual student progress. States and test developers have tried to reconcile these competing demands, but it has proven impossible to achieve both goals.

A government study found that, eight years after passage of the Every Student Succeeds Act, states hadn't produced complete improvement plans for even half of their lowest-performing schools.

Solving the Stalemate

As a result, many states have abandoned the high-quality state tests developed at substantial public cost by the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced consortia. The competing priorities have stymied nascent testing innovations and paralyzed the national discussion on how to do the critical work of helping students acquire the academic skills, knowledge, and habits of mind necessary to pursue meaningful postsecondary options. And they have played into the hands of those who would strip all state testing provisions from federal law—putting at risk state testing's contributions to research, school improvement, and educational equity. Recent moves by Wisconsin and other states to lower the proficiency bar on their tests suggests the importance of public scrutiny of student performance.

The best way out of the testing stalemate is to reduce the demands on state testing by revamping federal testing provisions designed to identify low-performing schools for improvement and then lean into understanding individual student performance at the local level to inform parents and strengthen instruction. This two-part strategy would weaken the case of testing abolitionists by improving the quality and shrinking the scale of state testing while preserving its core mission: helping policymakers, parents, and taxpayers understand public school performance against state standards.

Using state tests to compare schools' performance presents a challenge because it is predicated on high levels of test security, testing students on equally demanding grade-level content, and standardized rules for test administration and scoring.

That means teachers, parents, and students cannot see test questions or answers without the costly process of crafting new items every year. It means testing students in every school and school district on comparable content under the same conditions. And it means presenting a student's answers as a single, year-end score that's aligned to the state's standards. Only with these features can states confidently—and legitimately—use test results to target schools for improvement or otherwise hold them accountable for their students' performance.

But these requirements conflict with the demand, enshrined in the federal Every Student Succeeds Act of 2015, or ESSA, and echoed by many testing advocates, that state tests yield "diagnostic" information that teachers and parents can use to help individual students improve. While that's a worthy goal, it's virtually impossible for large-scale, end-of-year state tests to capture individual performance in sufficient detail to guide teachers' work with each student—not to mention that teachers typically receive state test results long after the school year has ended.

A better strategy would shift the focus of state testing to giving policymakers, parents, and the public an annual window into student and school performance, while stopping short of tying test results to consequences for schools and expecting them to yield a teaching plan for every student.

This shift would allow states to scale back testing—ESSA currently requires that they test every student every year in reading and math in grades 3 through 8 and once in high school and in science once per grade span (elementary, middle, and high school). States could reduce the amount of testing by borrowing the sampling approach used by pollsters. States could test a representative sample of students in key grades, or they could adopt what psychometricians call matrix sampling, in which each student is tested in greater depth on only a sample of the relevant standards. Matrix sampling could allow states to improve test quality and test a wider range of curriculum content, because not every student would have to answer every question.

CONTINUED ON PAGE 66

ALDEMAN & CHU CONTINUED FROM PAGE 62

use a poll to determine the actual outcome of an election. For that, setting aside the ultimate power of the Electoral College to select the president, we use "one person, one vote."

Still, there are some use cases in which a sample is sufficient to get a pretty good idea of what's going on. During the pandemic, for example, scientists discovered they could get a reasonably accurate understanding of Covid-19's prevalence in a community by testing wastewater samples.

But there are other use cases that call for a more pointed and precise approach. Wastewater samples can't tell whether a given individual is contagious. For that, a person must be tested.

This distinction is also helpful in analyzing approaches to other forms of testing. In K–12 education, do states need to test all kids on their reading and math skills, or could a sample do the job?

Lynn Olson and Tom Toch advocate for the latter: "testing a representative sample of students in key grades or testing students in greater depth on only a sample of state standards." While there's an understandable appeal to this idea—namely the promise of less testing overall—the downsides are great. And there's no guarantee that the overall amount of testing would in fact decrease.

Most important, in our eyes, a sampling approach would strip parents of the one credible source of objective information about how their own children are performing. State tests are meant to serve as a check on grade inflation and on teachers and schools that have a mixed record of delivering honest evaluations to parents. It's not enough to communicate how a child's school is faring on average—parents deserve to know how *their* child is doing.

It's not enough to communicate how a child's school is faring on average—parents deserve to know how *their* child is doing.

A sampling approach would also compromise our ability to understand trends by school and for subgroups of students and to measure growth in individual students' achievement from one year to the next. These are not just wonky and technical uses of state testing, lest we forget how easy it once was for states to mask the performance of their lowest achievers, often kids from low-income families and in the racial and ethnic minority, by sweeping academic inequities under the rug. What's more, such an approach would mean forgoing the power of test scores to predict the later life outcomes we want for our children.

Finally, we doubt that accepting these downsides would sufficiently address the political concerns that Olson and Toch rightly assert now threaten state testing regimes. For better or worse, testing is inextricably linked to evaluations of school performance and accountability systems.

Olson and Toch clearly mean well in their calls to scale back testing, but most critiques of state tests are related more to their use in holding schools and educators accountable for results and the perception that tests are punitive. And to be clear, testing abolitionists want nothing less than the elimination of *all* standardized testing—state, local, or otherwise. Slightly reducing the total number of tests administered does little to satisfy their concerns. Olson and Toch assure us that locally administered tests could serve many of the same purposes as the current state tests do, but count us as skeptical. Many school districts eschew local tests, and those tests that are administered are not consistently aligned to state standards. What's more, local tests are subject to the same political headwinds as state tests, as demonstrated by a recent proposal in Los Angeles to begin exempting schools from local assessments.

Here it's worth flagging the important yet underappreciated role student-level achievement data can play in informing the strategic allocation of resources. Districts should use aggregated subgroup data to target resources (for example, funding and support staff) and evaluate the efficacy of improvement initiatives and interventions. Unfortunately, districts often fail to use state tests in this manner and, as a result, students—especially those from the most marginalized communities—don't receive the services and support they require. The work of school and district leaders should be grounded in the analysis and application of this achievement data.

All of this does not mean that we are satisfied with the status quo—far from it. We have spoken out adamantly about the need for states to share the results of their tests with parents and educators much faster than they now do. We also agree with Olson and Toch that state tests are not the best vehicle to provide detailed instructional roadmaps for educators.

If anything, annual state tests should be more like quick checks to make sure kids are keeping up with state standards. To return to the Covid metaphor, states might view their annual tests less like the Covid laboratory tests that were highly accurate but faced long processing delays, and more like the rapid, at-home tests that provide actionable, on-the-spot information to individuals.

When and Where to Use Sampling

Sampling approaches make sense when policymakers are trying to get a broad understanding of trends and patterns. In the business world, the Bureau of Labor Statistics surveys a sample

CONTINUED ON PAGE 67

OLSON & TOCH CONTINUED FROM PAGE 64

The NAEP Model

The highly regarded, federally funded National Assessment of Educational Progress uses matrix sampling to capture student performance at the national, state, and local levels, as do national testing systems in other countries. An alternative to sampling, proposed by Scott Marion, director of the nonprofit Center for Assessment, would test every student every other year or every other grade. Both the matrix and Marion models make sense to us.

Less testing would free up time to gauge other student experiences and outcomes that many stakeholders in the testing debate want measured, including, for example, whether schools are creating a sense of belonging among students.

As a practical matter, the move wouldn't have much impact on school accountability, which in most states has been substantially weakened under ESSA.

The No Child Left Behind Act of 2002 required consequences for schools if students weren't performing up to state standards, as measured largely by state testing. ESSA maintained the frequency of state tests but defined accountability very differently, requiring only that bottom-performing schools be identified and that improvement plans be drawn up for them. States would decide which steps, if any, schools should take to improve. In other words, the new law eliminated NCLB's strongest improvement measures and devolved accountability decisions to states—and many states have declined to act decisively on low-performing schools. A 2024 federal Government Accountability Office study found that, eight years after ESSA's enactment, states hadn't produced complete improvement plans for even half of their lowest-performing schools—serving 2.5 million students—identified under ESSA. And improvement efforts are underway in far fewer schools than that.

Nor, given education politics today, are more stringent federal accountability mandates likely to return any time soon.

But states that want to use state test data to target schools for improvement could do so under the modified regime we're proposing. Data on demographic subgroups within schools would be more limited under either matrix sampling or Marion's approach, making it difficult to use the data for measuring school performance because sample sizes would be too small. But states could compensate for that by reporting scores for the bottom 25 percent of students in the school. The reality is that there's not much evidence that disaggregating scores by race and socioeconomic status has made a significant difference in closing achievement gaps.

Local Diagnostics

The focus on diagnostics, meanwhile, could shift to local testing, where tests would be tied more closely than their state counterparts to instruction and teachers would get results in time to help their students, rather than receiving what amounts to autopsy reports after schools close at the end of the year.

Many school districts already use this approach. Alison Timberlake, deputy director for assessment and accountability in the Georgia Department of Education, told us that given the widespread emergence of locally adopted 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 [state] standards . . . isn't necessary anymore." States could establish review panels of psychometricians, curriculum specialists, and local educators to ensure that the local standardized tests are of high quality and are aligned to state standards.

Reducing the demands on state testing would yield another benefit: enabling the implementation of testing innovations that have struggled to meet the technical requirements for validity, comparability, and reliability demanded of state tests by ESSA. These include "performance assessments" that probe deeper levels of learning by asking students to show what they know by completing an experiment or conducting an analysis rather than merely answering multiple-choice questions; student surveys of school climate; and "competency- or skills-based assessments" that provide students immediate results and permit them to progress at their own pace based on demonstrated mastery.

There's not much evidence that disaggregating scores by race and socioeconomic status has made a significant difference in closing achievement gaps.

Compared to current state tests, these new forms of measurement are able to gauge a wider range of student competencies, from career and technical skills to interpersonal skills to digital problem solving. Federal policymakers could require that school districts measure students against the same standards and learning progressions used in state tests and that parents receive clear score reports so they understand exactly how their children are performing.

NCLB roughly tripled the amount of state testing in the nation's schools. It also led to significantly greater school-district use of commercially developed interim and benchmark tests to

CONTINUED ON PAGE 68

ALDEMAN & CHU CONTINUED FROM PAGE 65

of individuals and employers each month to get a reasonably accurate picture of labor market conditions. Similarly, the National Assessment of Educational Progress (NAEP) tests a sample of students at regular intervals to understand achievement levels in each state.

The results of these surveys inform policymakers and provide clues about where to begin looking for problems and solutions. However, the labor-market surveys aren't precise enough to be useful to individual employees or employers, let alone to researchers trying to do causal research. If an employer wanted to understand trends within their own company, they would need to look at the size of their own workforce and turnover rates among their own employees. In education, we have a derogatory term ("misNAEPery") for policymakers who merely eyeball the NAEP trends and try to argue for or against certain policy changes.

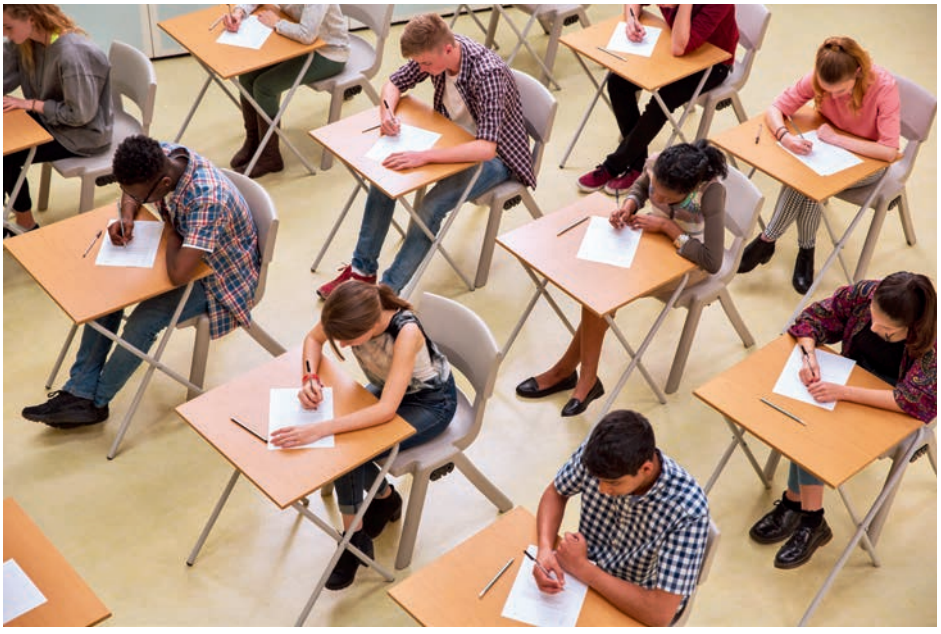
More-detailed use cases require more-detailed data. As parents of school-age children, we want to know how our kids are doing. And, while we generally trust teachers and principals (one of us is a former principal), we still appreciate seeing how our own kids are doing on objective, standardized tests. We want that common benchmark. If states switched to a sampling approach, in which only some kids were tested each year, the parents of untested students would miss out on receiving objective, comparable, and individualized results.

Policymakers also need detailed data on student-level performance. Research on student performance in Florida and North Carolina found that both schools and districts have a meaningful influence on student learning. That was especially true during the pandemic, when researchers found that the specific school a student attended accounted for about three-quarters of the widening gap between low- and high-achieving students in math and about one-third of the gap in reading.

Sampling would make it much harder to evaluate the performance of schools and districts, especially for discrete student groups. Olson and Toch downplay this problem, but, because of sample-size issues, it simply wouldn't be possible to look at school-level results for different student subgroups.

For a concrete example, imagine an elementary school with eight Black students in each of grades 3, 4, 5, and 6. To determine if this school should be held accountable for a given student group, a state would combine performance results across the grades and then see if the group met a minimum sample size. According to a recent analysis from Education Commission of the States, most states apply a minimum subgroup size of 10 to 20 students, with some as high as 30 students. With a total of 32 Black students, this school would just barely meet the minimum sample size, and it would be responsible for the performance of those students.

Testing abolitionists want nothing less than the elimination of *all* standardized testing—state, local, or otherwise. Reducing the total number of tests administered does little to satisfy their concerns.



ISTOCK

Administering state tests to all students every year still provides the most accurate measures of progress and accountability, but the results often arrive too late for schools to take action.

But if the state tested only a *sample* of students, the number of Black students tested in this hypothetical school would likely fall below the threshold. The sample sizes start to get very small very quickly. When one of us (Chad Aldeman) ran a sampling model for Washington, D.C., he found that about half of the

CONTINUED ON PAGE 69

OLSON & TOCH CONTINUED FROM PAGE 66

measure whether students were on track to do well on state tests. And these assessments were frequently layered on top of existing local testing. By reducing the state testing footprint and incentives for school districts to test students' readiness for state tests, the new model we're proposing would likely lessen standardized testing significantly in many schools and allow more time for instruction.

The Federal Role

These changes would require a revision of federal testing requirements or, in the short term, a willingness by federal officials to let states adopt the model under the U.S. Department of Education's Innovative Assessment Demonstration Authority. Designed to encourage novel approaches to state testing, IADA has attracted few takers since its creation nine years ago because it still requires states to meet federal accountability mandates. However, Education Secretary Miguel Cardona announced late in 2023 that he was relaxing the program's requirements to encourage more states to take part, an approach the new administration could be inclined to continue given its stated commitment to reducing federal oversight.

We know that the new testing paradigm we're proposing would spark controversy because it would require tradeoffs. Even though our approach would provide a significant level of transparency—and transparency itself serves as a form of accountability—many accountability advocates insist on the need for consequences for schools whose students perform poorly, even if accountability under ESSA has fallen short of that expectation.

Also, without annual testing of every student, measuring growth in student achievement over time poses difficulties. Federal law now rightly encourages that metric in order to more fairly evaluate the work of schools that serve large percentages of vulnerable students who start school behind their more privileged peers. What's more, a good deal of education research depends on the data derived from measuring student performance year after year. State testing of students every other year would permit policymakers to continue to measure student growth with a meaningful degree of confidence while providing an audit on local reporting. States and school districts could administer performance assessments and other innovative measures in the alternate grades or years.

We're encouraged that Chad Aldeman and Dale Chu agree with us that standardized testing is excessive, that state tests are not the best vehicle to provide detailed instructional roadmaps for educators or diagnostics on individual students, and that local educators and parents need test results more quickly.

But they sidestep the dilemma at the crux of our essay: that the current configuration of state testing is playing into the hands of testing opponents even as the school accountability system it powers has stalled in many states.

Nor in the course of critiquing matrix sampling do Aldeman and Chu acknowledge that testing every student every other year or every other grade—a second path to the more manageable testing system that we propose—preserves the capacity to measure student growth.

And their singular state testing fix—quicker "pulse tests" throughout the school year—runs up against the federal accountability-driven requirement that state tests produce a single, year-end score tied to state standards. That's why we propose faster, more frequent testing at the local level.

The new model we're proposing would likely lessen standardized testing significantly in many schools and allow more time for instruction.

ALAMY

Outgoing Education Secretary Miguel Cardona relaxed accountability mandates to encourage more innovation in state testing.

We support testing, but we also try to think realistically. Accountability advocates have not been inclined to compromise. More broadly, stakeholders have yet to engage in a clear-eyed national conversation about how much a single test can accomplish; policymakers continue to search for a unicorn assessment that can be all things to all people. Until they begin to explore alternatives such as what we have proposed here, the stalemate on standardized testing will continue—and the likelihood of losing state testing altogether under the next reauthorization of the federal elementary and secondary education law will increase.

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

ALDEMAN & CHU CONTINUED FROM PAGE 67

city's elementary schools would not be held accountable for low-income or Black students, less than 10 percent of schools would be responsible for Hispanic students or English language learners, and not a single elementary school would be accountable for the progress of students with disabilities.

The same math applies to school *districts* as well. Across the country, there are almost 9,000 school districts that serve between 100 and 1,000 students each. Collectively, those smaller districts educate more than four million students, but shifting to a sampling approach wouldn't tell us much about the performance of those students.

Note that it would be technically possible to "over-sample" student groups or students in small schools or districts, but that would defeat the purpose of sampling in the first place. It would also mean that the testing burden would fall disproportionately on the traditionally underserved student groups that policymakers are the most concerned about.

But perhaps the biggest drawback with the sampling approach is that it might accomplish neither its political nor its technical goals. Opponents of "high-stakes testing" often worry more about the perceived stakes than the tests themselves. Standardized tests are frequently scapegoated for school closures or teacher layoffs, but real sanctions resulting from them are few and far between. The truth is that the *threat* of accountability has always been greater than any actual consequences, and that's even truer today.

Moreover, the purported goal behind sampling is to reduce the amount of time kids spend taking tests, potentially freeing up more time for classroom instruction. This is a worthy aim, but the federally required state tests are not the main problem here. In fact, these exams account for only a tiny fraction of the time typically devoted to assessments each year. The real culprits are the layers upon layers of other tests adopted by states and local districts. There are potential solutions such as testing audits to reduce redundancy, but we're not holding our breath for Congress to develop some sort of *maximum* testing rule, so it would behoove individual states and districts to determine which tests deliver the greatest value.

Simply put, in our view, a sampling approach would have significant downsides without tangible benefits. Rather than backing away from the principle of testing all kids, we think there's room for innovation on what those tests look like and how states use them.

Faster, Cheaper Tests

Although we don't agree with Olson and Toch's proposed solution, we concur with their diagnosis of the problems with current state assessments. Their results come too late to be of much use to parents and educators, and the data they do produce are not detailed enough to inform instruction.

Our preferred vision for state tests is to make them more akin to rapid Covid tests or new pregnancy tests that promise ultra-fast results. Six states—Florida, Tennessee, Texas, Indiana, Louisiana, and Georgia—have been consistently fast at releasing results to the public in recent years, mainly because they use the results to inform some meaningful decisions—such as 3rd-grade reading requirements—and as a result have streamlined bureaucratic processes around verifying which students attended which schools. Another example comes from Ohio, where the state legislature mandated that parents receive their child's results no later than June 30 each year. More states should follow their lead and accelerate their timelines, especially as AI tools further speed up the scoring process.

Moving in this direction would also return state tests to their ultimate purpose—to serve as an honest check on schools and districts, rather than as detailed blueprints for instruction for every student (a function that state tests were never meant to perform). States could also complement these regular pulse checks with more open-ended formats and intensive assessment designs at key stages, such as 3rd-grade reading, middle school math, or specific subject areas in high school (as is done now with high school end-of-course exams and AP and IB exams).

Standardized tests are frequently scapegoated for school closures or teacher layoffs, but real sanctions resulting from them are few and far between.

We believe this approach is a better fit for the current moment. There's a troubling "perception gap" in which 90 percent of parents think their child is performing at or above grade level in reading and math, even as objective data put the number much lower. If anything, states need to do a better job of getting testing data to parents quickly so they have time to act on the information—to say nothing of presenting the results in a more compelling way.

Both of us support additional innovation in assessments. But policymakers need to be clear-eyed on the policy tradeoffs that come with different approaches. In the case of sampling, the perceived benefits need to be weighed against the very real costs. Instead of retreating from the principle of delivering standardized, objective results for every child, state policymakers and advocates would do well to focus on the needs of parents and families by improving the transparency and usability of state tests.

Chad Aldeman is an education writer and the founder of Read Not Guess. Dale Chu is a senior visiting fellow at the Thomas B. Fordham Institute.