THE CASE FOR STATE TESTING

National Center for the Improvement of Educational Assessment

March 2025

State leaders have long recognized the value of statewide tests. Many had well-developed testing systems long before the No Child Left Behind Act, signed in 2002, required annual assessments in grades 3-8. Another sign of that recognition is the recent plethora of new state testing requirements for K-3 students.

But would states continue to test if the U.S. Department of Education was disbanded or stopped enforcing the testing requirements of the Every Student Succeeds Act (ESSA, the successor to NCLB)? After all, state tests come in for their share of criticism.[1] State leaders often respond by citing federal testing requirements. But if enforcement diminishes or disappears, as part of the Trump administration's promise to downsize its role in K-12 schools and "send education back to the states," would states keep testing? Should they?

IT'S ALL ABOUT PURPOSE AND USE

It's a regular mantra in educational measurement: assessments are designed and validated for specific purposes and uses. As much as we may like our Swiss Army knives, we all know that they don't do everything well. Have you ever tried to use the scissor attachment? Similarly, assessments work best when designed to fulfill a limited set of purposes and uses.

State achievement tests are not useful for informing classroom instruction. We have other tests and processes for that. Hence, the call for balanced assessment systems, where a coherently organized set of assessments is designed and implemented to support different purposes and uses. Statewide tests are designed and validated to support the four major purposes and uses below. We contend that these are so important that states should continue to require students to complete statewide assessments regardless of whether the federal law is enforced.

1. Monitoring statewide educational growth and achievement
2. Evaluation and continuous improvement
3. Transparency and public engagement
4. Signaling rich learning expectations

[1] We agree that state tests need to be improved so they can better fulfill their intended purpose: improving and supporting student learning. We've written extensively about this. In response to the current political landscape, we're focusing this paper on the importance of state testing, but we want to emphasize that this focus doesn't imply that we think no changes are necessary to state testing systems.

Eliminating state tests would not eliminate the need for—and importance of—these functions; it would only eliminate crucial information needed to make good decisions about supporting schools.

Monitoring

Statewide standards-based achievement tests are critical tools for helping state education leaders and policymakers monitor students' educational opportunities and outcomes. High-quality statewide achievement tests are well-suited to this vital work because they are designed and administered to yield comparable scores across students, schools, and districts. As a result, these tests provide a means for state education leaders and policymakers to monitor the achievement and growth of all students in the state.

You might ask: Can any standardized test be used to monitor achievement and growth? In short, no. State summative assessments developed to meet federal requirements can monitor student learning in ways that off-the-shelf commercial interim assessments cannot. As we address in subsequent sections, state tests must meet rigid requirements for alignment, accessibility, and many other standards and undergo extensive independent review.

Having a means for comparability when monitoring educational performance is crucial for evaluating the degree to which school districts provide equitable and appropriate opportunities for all students to learn and develop. Monitoring efforts—perhaps as part of accountability systems—should help state leaders identify where students are receiving these opportunities and where they are not. In cases where they are not, leaders and policymakers should provide resources and other supports to rectify the shortcomings.

Advocates pushed for universal testing requirements because they were concerned that districts were "hiding" low-performing students.

Remember, advocates pushed for the universal testing requirements in NCLB's predecessor, the Improving America's Schools Act of 1994, and subsequent reauthorizations because they were concerned that districts were "hiding" low-performing students. There is no question that NCLB's testing requirements helped shine a light on students who had been left behind, especially students with disabilities and English learners. We've done a good job of shining that light. But moving forward, we need to up our game in providing support to improve the equality of learning opportunities.

When thinking about monitoring educational performance, most people default to the image of point-in-time achievement scores. Following the lead of our first board chair, Dale Carlson, the Center for Assessment has a long history of advocating for multiple perspectives of student and school performance, particularly student longitudinal growth. The shift from grade-span testing under IASA to testing annually in grades 3-8 under NCLB and ESSA has opened up noticeable advantages for measuring student longitudinal growth. When considering the importance of statewide achievement tests, we must recognize the value of documenting student growth and achievement; such measures provide insights into how individual and groups of students progress over time, which point-in-time scores cannot provide.

Evaluation and Continuous Improvement

The results of high-quality state assessments are a key outcome variable in curriculum, program, and policy evaluations. State tests are not designed to improve instruction in real time—that's a job best handled by classroom formative assessments—but when they're used as part of a well-conceived evaluation study, they can play an essential role in helping education leaders and policymakers understand which programs and policies are working well and which are not. Importantly, this use is not limited to summative evaluation. When used as part of a continuous improvement (formative evaluation) system, high-quality state assessments can provide information that allows leaders to adjust the implementation of multi-year programs and policies.

Evaluation and continuous improvement should occur at both local and state levels, because each has different program and policy initiatives. Local district leaders often pilot new curricular programs to determine which ones will be fully implemented. Documenting different student growth rates for those experiencing the new curriculum compared with those using the legacy materials can be an effective evaluation design. State assessment results are an appropriate outcome variable because of the test's superior technical quality, especially its alignment with the state's learning standards.

State tests can play an essential role in helping education leaders and policymakers understand which programs and policies are working well and which are not.

Unfortunately, the consequences associated with state accountability systems often obscure state tests' potential to serve evaluative purposes. Educators and local education leaders, particularly from lower-performing schools, focus on avoiding the consequences rather than on the insights they might gain from the assessment results. When used as part of thoughtful research and evaluation programs, state and local leaders can demonstrate the utility of high-quality state assessments to help study and improve their educational systems.

Transparency and Public Engagement

Depending on how education is funded in each state, public education is one of the largest budget items in state or local government budgets (sometimes both). We believe that public education is a foundation of democracy, and we support such expenditures. However, we also believe that the public should clearly understand how these funds are used and whether school districts provide their students with meaningful learning opportunities. Comparable statewide test scores are an essential source of information to support efforts to build public trust and increase this type of transparency. Of course, many other indicators of schooling should also be publicly reported, but student academic outcomes should be part of the mix.

Together with local assessment results, state assessment scores can be used to engage parents and other members of the public in conversations about their local schools and their goals for their students' futures. For example, released state test items, along with high-quality classroom assessments, can help parents understand what students are expected to know and be able to do. Every generation seems to think that "kids today" have it easier than when they were young. Engaging the parents and community members in fun activities such as quiz shows using released test items could help the adults see that students today are learning complex and meaningful ideas and skills.

Signaling

High-quality state tests that embody the state content standards can meaningfully represent the intended learning goals and provide explicit depictions of the content standards for teachers and students. Serving this signaling function well requires tests beyond simple, selected-response questions and similar types of items that draw on relatively low levels of cognition. Instead, tests should include extended writing tasks and performance assessments in mathematics and science that require students to apply their knowledge and skills to solve complex problems. Therefore, if states want to use the state assessment to signal the rich learning experiences that leaders hope to see in classrooms, they must ensure they are sending the right signals.

States can use the state assessment to signal the rich learning experiences that leaders hope to see in classrooms.

We are not saying that all state tests meet these four purposes optimally, but we believe that they have the greatest promise of doing this, even if they currently fall short in some areas.

TECHNICAL QUALITY

State tests must meet the following rigorous quality criteria if they are to serve these important purposes and uses:

- **Validity** is the overarching technical quality criterion. It is an evaluation of the degree to which the evidence supports the claims about the meaning of the test scores. In other words, if the designer claims that students who score proficient on the Grade 5 math test, for example, can use the required knowledge and skills to solve grade-level math problems, they must provide evidence to support such statements.

- **Alignment** was popularized as a technical criterion as part of the theory of action undergirding the standards-based education reform movement. Assessments must be designed to measure students' learning of the grade-level or grade-span content standards, and all content standards must be represented on the assessment. Alignment is evaluated by having content experts (often grade-level teachers) identify the knowledge and skills necessary to answer each test question and then matching this information against the knowledge and skills in the state content standards. The test items should only measure learning expectations in the grade-level content standards. Further, all of the content standards must be represented in the assessment, if not every year, there must be a systematic plan to assess all of the standards on a reasonable schedule.

- **Reliability** is generally conceptualized as stability over both time and items. All measures contain error (uncertainty). Reliability is a quantification of that error. For example, if a test was administered to a sample of students and an identical test was administered the next day, we would expect each group of students to produce similar performances.

- **Fairness and accessibility**, taken together, help ensure that all students, including students with disabilities and English learners, can access the assessment and show what they know and can do without experiencing any barriers to their performance. Universal Design (UD) and

Web Content Accessibility Guidelines (WCAG) are comprehensive frameworks that help guide fair and accessible learning and assessment experiences. States require test developers to adhere to digital accessibility standards and Universal Design frameworks to guide fair and accessible learning and assessment experiences. These requirements have become incorporated into the design and administration of essentially all state assessment programs.

- **Comparability** is essential for establishing valid inferences about scores across individuals, schools, or districts. Like validity, comparability is always evaluated in the context of specific purposes and uses. Large-scale statewide assessments must meet strict comparability expectations for student- and school- level comparability. For example, users should be able to draw similar inferences about two students receiving the same score on the same Grade 5 math assessment. Similarly, comparable assessments should allow users to support similar inferences about schools with the same growth and achievement results.

These criteria are elaborated in the "bible" for testing professionals, *The Standards for Educational and Psychological Testing*, and further specified for state assessments in the U.S. Department of Education's Standards and Assessment Peer Review Guidance. These documents provide a shared understanding of quality and guide the work of state testing professionals and their assessment company partners.

Educator Involvement

Technical measurement experts and assessment peers evaluate the quality of state tests, but they are not the only reviewers. Statewide achievement tests are unique in their level of educator, parent, and community involvement to ensure the technical quality of the assessments. Educators are typically involved throughout the entire assessment cycle. Educators in many states help write test questions and score open-ended questions such as essays.

Statewide achievement tests are unique in their level of educator, parent, and community involvement.

In all states, educators serve on content and bias/sensitivity review committees, reviewing every test question before the items even make it to a field test. Following the test administration, educator committees review the results of the field test to ensure that every test question is suitable to serve as an operational test question. Finally, when cut scores are set to define achievement levels for score reporting, educators comprise the majority of standard-setting committee members.

To summarize, state summative exams have been held to the highest quality standards of any tests administered to K-12 students in the country. These tests are thoroughly evaluated by local educators, state technical advisory committee members—independent national measurement experts—and peers through the U.S. Department of Education's peer-review process. This level of **independent** review and evaluation and the resulting **transparency** is far beyond what we see for any other assessment, including the National Assessment of Educational Progress (NAEP), the SAT, and the ACT.

DON'T FALL FOR SEEMINGLY SIMPLE SOLUTIONS

Some of us are old enough to remember the Lake Wobegon effect brought to light by Dr. John Jacob Cannell in 1988. In those days, many states used nationally norm-referenced tests as their statewide achievement test. Dr. Cannell showed that all states were performing above the national average. Bob Linn and colleagues helped explain the phenomenon, which was due in part to the use of old norms. Using these "off-the-shelf" tests allowed all states to feel good about their performance, even if many didn't deserve it. State standards-based achievement tests are designed to present an honest picture of student achievement and growth.

While norm-referenced tests are used much less frequently nowadays, commercial interim assessments are ubiquitous. District leaders and others have pressured state assessment leaders to replace the state test with their favorite interim assessment. The rationale is simple and somewhat compelling: "We already use Assessment X three times each year, and we like the results. Why should we add time with a state assessment?"

The answer is straightforward. Commercial interim assessments are not designed to support the summative purposes described throughout this document. Further, no commercial interim assessment has met the technical requirements that state assessments must meet and, therefore, cannot serve the purposes and uses outlined above. Interim assessments can be useful for school and district leaders to monitor students' performance throughout the school year, but that is a different use case than for statewide summative assessments.

No commercial interim assessment has met the technical requirements that state assessments must meet.

EdReports and the Center for Assessment recently recruited commercial interim assessment providers to participate in an independent quality review. Only one company agreed to participate, and we did not feel it right to publish the results with only one volunteer. The point is that companies can choose whether or not to open their systems to public scrutiny. State assessments do not have that choice. On the few occasions when one of these interim assessments was used as the state assessment, it failed to meet the U.S. Department of Education peer-review requirements.

Similarly, the explosion of early-elementary assessments, often to screen for reading difficulties, has left many measurement experts wondering how these tests can be used for high-stakes purposes, such as grade retention, without anyone knowing if the tests are any good. We've recently started inviting representatives from the companies responsible for these early reading (and occasionally math) assessments to present their technical documentation to state technical advisory committees. The technical documentation from these companies does not come close to the quality and rigor of the documentation for statewide reading, math, and science assessments.

SUMMARY

We want state assessments to serve critical purposes. They are crucial tools for monitoring the achievement and growth of all students in the state, evaluating programs, providing a way to report transparently about schooling in the state, and signaling to teachers and leaders important information about the knowledge and skills students are expected to learn. Meeting these needs requires very high-quality assessments with solid documentation. State assessments are the highest-quality assessments administered to our K-12 students. We can't give up on them even if the federal government takes its foot off the pedal.

ACKNOWLEDGEMENTS

We are grateful to members of the Council of Chief State School Officers' Technical Issues in Large-Scale Assessment (TILSA) state collaborative for their thoughtful review and suggestions.

The National Center for the Improvement of Educational Assessment, Inc. (the Center for Assessment) is a New Hampshire based not-for-profit (501(c)(3)) corporation. Founded in September 1998, the Center's mission is to improve student learning by partnering with educational leaders to advance effective practices and policies in support of high-quality assessment and accountability systems. The Center for Assessment does this by providing services directly to states, school districts, and partner organizations to support state and district assessment and accountability systems.

National Center for the Improvement
of Educational Assessment
Dover, New Hampshire

www.nciea.org