



# Developing State Assessment Systems That Support Teaching and Learning

## What Can the Federal Government Do?

Aneesha Badrinarayan and Linda Darling-Hammond

In collaboration with Michael DiNapoli, Tara Kini, Tiffany Miller, and Julie Woods

# **Developing State Assessment Systems That Support Teaching and Learning: What Can the Federal Government Do?**

**Aneesha Badrinarayan and Linda Darling-Hammond**

In collaboration with Michael DiNapoli, Tara Kini, Tiffany Miller,  
and Julie Woods

# Acknowledgments

We are grateful to the many policymakers, education leaders, researchers, and educators who contributed to our thinking by sharing their experiences with past and current assessments as well as their hopes for the future with us. In particular, we thank Randy Bennett, Edward Haertel, and James Pellegrino for their time and technical expertise in developing previous versions of technical recommendations that we draw upon here. Although all content, views, and recommendations made here are those of the Learning Policy Institute (LPI) alone, the following people helped make this report a reality by sharing their experiences, successes, and challenges with efforts to create more learning-centered assessment systems: Daniel Alcazar-Roman, Matt Blomstedt, Julianna Charles Brown, Sara Cooper, Nathan Dadey, Paul Dumas, Ellen Ebert, Cory Epler, Denise Forte, Jeff Greig, Jeremy Heneger, Ellen Hume-Howard, Alissa Kilpatrick, John King, Angela Landrum, James Lane, Paul Leather, Bob Lenz, Susan Lyons, Scott Marion, Mike McGee, Lillian Pace, Susan Patrick, Raymond Pecheone, William Penuel, Stephen Pruitt, Sam Ribnick, Jeffrey Riley, Breigh Rhodes, David Ruff, Lorrie Shepard, Michele Snyder, Corrine Steever, Dianne Tavenner, Rhonda True, Gene Wilhoit, Audrey Webb, Justin Wells, Katie Van Horne, and members of the State Performance Assessment Learning Community and the Interstate Learning Community.

We also thank our LPI colleagues Monica Martinez, Patrick Shields, Charlie Thompson, and Larkin Willis for their thought partnership around issues of assessment policy. We are indebted to the members of the LPI Communications team for their invaluable support in editing, designing, and disseminating this report. Without their generosity of time and spirit, this work would not have been possible.

This research was supported by the Carnegie Corporation of New York, Chan Zuckerberg Initiative, and Walton Family Foundation. Core operating support for LPI is provided by the Heising-Simons Foundation, William and Flora Hewlett Foundation, Raikes Foundation, Sandler Foundation, and MacKenzie Scott. We are grateful to them for their generous support. The ideas voiced here are those of the authors and not those of our funders.

## External Reviewers

This report benefited from the insights and expertise of two external reviewers: Randy Bennett, Educational Testing Service, and James Pellegrino, University of Illinois Chicago. We thank them for the care and attention they gave the report.

Suggested citation: Badrinarayan, A., & Darling-Hammond, L. (with DiNapoli, M., Kini, T., Miller, T., & Woods, J.). (2023). *Developing state assessment systems that support teaching and learning: What can the federal government do?* Learning Policy Institute. <https://doi.org/10.54300/885.821>

This report can be found online at <https://learningpolicyinstitute.org/product/developing-assessment-systems-federal-support>

This work is licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/4.0/>.



Document last revised April 17, 2023

# Table of Contents

Executive Summary .....	v
Introduction .....	1
History of Federal Testing Guidance .....	3
The Promise of the Every Student Succeeds Act for Advancing High-Quality Assessment ...	4
ESSA’s Innovative Assessment and Demonstration Authority: Opportunity and Barrier ....	7
Looking to the Future: Calls for Assessment Systems That Better Support Teaching and Learning .....	8
Current State Efforts.....	9
Challenges to Address .....	14
Supporting Next-Generation Assessment Systems: Recommendations for Federal Executive Action.....	16
Align Technical Requirements and Peer Review Processes With ESSA’s Assessment Allowances and Requirements .....	17
Enable IADA to Better Support Innovation in Assessment .....	30
Create Additional Pathways to Innovation.....	38
Conclusion.....	42
Endnotes.....	43
About the Authors .....	46

## List of Figures and Tables

Figure 1	Advanced Placement Computer Science Curriculum-Embedded Performance Task.....	27
Table 1	Summary of Assessment Requirements Under Section 1111 of the Every Student Succeeds Act .....	5



## Executive Summary

There is a growing call to reconsider current approaches to national and state assessment system policies and practices. State and local education agency leaders, educators, community leaders, and advocates have voiced concerns that our current state assessment systems—defined primarily by end-of-year multiple-choice tests—are unable to meet contemporary needs for information that supports teaching and learning.

More than 20 states are involved in efforts to transform one or more aspects of their assessment systems; however, the process of securing federal assistance and approval to make transitions to substantially improved systems poses numerous challenges. Among them are the costs and time required to change systems, the management of trend disruptions when new assessments are introduced, and interpretations of how to meet federal approval criteria under business rules that often keep new tests looking very much like old ones.

This report synthesizes policy analyses and findings from legal and research analyses, as well as consultations with national, state, and local leaders, to (1) outline the history of federal testing guidance and state responses that have shaped the current context, (2) describe strategies states and districts are pursuing to evolve their assessment programs into high-quality systems that both signal and support better teaching and learning processes for all students, and (3) identify key ways that the federal government could support assessment reforms that enable thoughtful assessment of meaningful skills in ways that also better support teaching and learning.

### **The Every Student Succeeds Act: Opportunities and Barriers for Meaningful Assessment Systems**

In 2015, the Every Student Succeeds Act (ESSA) opened new possibilities, relative to the prior decade under No Child Left Behind, for how student and school success are defined and supported in U.S. public education. The law deepened the concept of student learning to be more consistent with what students need to be successful in 21st-century society and careers, calling for measurement of “higher-order thinking skills and understanding” as part of “high-quality student academic assessments in mathematics, reading or language arts, and science.” ESSA intentionally created opportunities for assessment innovation by explicitly allowing the use of multiple types of assessments, including “portfolios, projects, or extended-performance tasks,” as part of state systems.

In addition to its statewide assessment provisions, which encourage all states to advance more innovative assessments that better support teaching and learning, ESSA also explicitly allows a subset of states to pursue innovation through the Innovative Assessment Demonstration Authority (IADA). This provision invited up to seven

states to apply for an innovative assessment pilot to implement new approaches to assessment and gradually scale them statewide. IADA defines innovation flexibly, allowing for state systems that may include competency-based assessments, curriculum-embedded performance assessments, and through-year assessment approaches. The primary promise of IADA is that it provides states a means to pilot new assessments by allowing a subset of districts to use the new assessments rather than the old ones, without double testing students. This is an important feature of reform: States are not simply substituting one commercially administered standardized test for another at the end of the school year.

While many states were initially pleased to have the opportunity to explore the flexibilities in the law through IADA, applying for and complying with the terms of the waiver have proved to be so onerous and constraining that few states have yet been able to use IADA to explore or implement innovative assessment designs. Fewer still have been able to develop systems that provide insights into student learning in ways that are particularly meaningful to teaching, as originally envisioned by ESSA.

## Calls for Assessment Systems That Better Support Teaching and Learning

Many state leaders see the opportunities ESSA creates and want to transform their state assessment systems to take advantage of these affordances—with or without IADA. Through a series of conversations with state and local leaders as well as teachers and partners in the education space, a common set of goals for assessments that can inform and improve teaching and learning in schools is emerging. These common goals are:

- **Assessment tasks should encourage applied learning and higher-order skills.** Statewide assessments should prioritize engaging, realistic tasks that promote and support better teaching and learning—and, ultimately, provide better information about student progress.
- **Assessments should be integrated into a system that supports high-quality teaching and learning.** To support meaningful, deeper learning for students, assessment systems should be both designed and used as part of a coherent, well-integrated system of curriculum, instruction, and professional learning for teachers.
- **Assessments should be part of accountability systems that support student access and success.** Assessments should be part of improved accountability systems designed to encourage behaviors and actions that lead to a more informed focus on school improvement, more equitable access to learning opportunities, and greater student success.

States are pursuing several promising approaches to address many of these needs. Some of these approaches strengthen the capacity for high-quality instruction informed by formative assessment in schools within integrated curriculum frameworks that also inform summative assessments. Other approaches position more innovative assessments to become part of the state summative assessment process itself, where sufficient comparability safeguards are in place. In nearly all efforts, states seek to address the vision and challenges described above in ways that position the state assessment system to signal and incentivize what high-quality teaching, learning, and student performance should look like, while allowing for appropriate flexibility for local decisions.

## **Possibilities for Federal Executive Action**

Many of the challenges that state and local leaders identify have to do with enabling conditions. Many states and districts have clear and compelling ideas about what would position assessments to better support teaching and learning, but they need time, support, and permission to innovate in those ways. Several possible federal executive actions could encourage innovative assessment systems that better support teaching and learning. Some actions can strengthen statewide assessment systems as a whole as well as those innovative systems developed under IADA, while other recommendations focus solely on strengthening IADA implementation. All actions discussed here are permissible under current federal law.

### **Align technical expectations and peer review processes with ESSA's assessment allowances and requirements**

The U.S. Department of Education's approval process for all state assessment systems is guided by an internally developed and moderated peer review process, used to render judgments about state systems. While ESSA explicitly encourages more instructionally relevant assessment approaches, the peer review process is often unnecessarily constraining and disincentivizes the very kinds of assessments that ESSA encourages. The Department of Education's interpretation of ESSA's assessment provisions in the peer review guidance reifies traditional standardized measures, privileging assessments that are administered once to all students, have a large number of relatively superficial items measuring many standards, and can be rapidly machine scored without needing expertise to evaluate.

If the federal peer review process were updated to truly focus on high-quality assessments and data without privileging any one particular approach, state assessment programs could then be designed in ways that are technically strong (potentially stronger than existing assessments) and still support a range of (allowable) instructionally useful innovations that states are considering. The recommendations highlighted below suggest ways to bring the peer review into alignment with the opportunities for instructionally relevant assessment allowable within ESSA.



- **Recommendation 1: Highlight opportunities and update processes to support more instructionally relevant assessments that reflect student performance in relation to both grade-level standards and multiyear learning progressions.** Rather than requiring that nearly all items be aligned only to grade-level standards, peer review guidance can emphasize ways to leverage items that sample along multiyear learning progressions to yield results that provide more precise information about what students know and can do, while still providing robust information about grade-level achievement to comply with ESSA's requirements.
- **Recommendation 2: Highlight technically sound approaches to meeting federal peer review requirements that allow state assessments to assess the depth of state standards while ensuring sufficient coverage.** Rather than prioritizing coverage of the easily tested standards with many superficial items, guidance can encourage states to sample strategically, allowing space for more holistic performance tasks that evaluate the complex forms of thinking, disciplinary practices, and performance intended by the standards.
- **Recommendation 3: Update peer review guidance to emphasize requirements for test security that are appropriate to the design of the assessment.** Rather than assuming that the only way to achieve valid scores is to keep tests secret, guidance could explicitly recognize that some test designs—such as those that use more authentic tasks—are not as sensitive to prior knowledge of what is being asked and should be subject to different expectations for test security than a multiple-choice test.
- **Recommendation 4: Revitalize foundational elements of the peer review process—peer reviewer selection and moderation of the process—to ensure that states can take full advantage of the opportunities provided by federal law and the Department of Education's technical guidance.** Rather than relying on peer reviewers and processes that are tied to traditional assessments, the Department of Education could ensure that peer reviewers include experts in innovative assessment and that they have the right guidance—including updated ideas about appropriate psychometric evidence—to meaningfully evaluate innovative systems.

### **Enable IADA to better support innovation in assessment**

While the inclusion of IADA within ESSA was first met with excitement by states, this optimism has waned. IADA does not currently offer states enough opportunity and flexibility to make the tremendous effort needed to create new assessment systems worth it. In fact, many of IADA's requirements are viewed as onerous and may actually limit efforts to develop innovative systems. The recommendations below highlight ways executive action could shift the cost-benefit trade-offs to open opportunities for innovation and remove barriers to state participation.

- **Recommendation 5: Update the interpretation of comparability of results within current IADA regulations to better enable high-quality innovative assessment approaches.** It is essential to ensure that a given assessment provides comparable standards-aligned tasks that generate comparable student scores across students, schools, and districts. However, IADA constrains innovation by requiring comparability of results across the innovative and traditional tests, limiting how much an innovative assessment can differ from the current test, even when the new assessment seeks to better surface student understanding of state standards (e.g., measuring more complex skills and abilities, addressing standards that are not well represented on current tests). Rather than requiring that new assessments produce the same scores as existing tests, guidance could encourage states to submit compelling evidence that their innovative tests are of equal or higher quality than the existing assessment and that they produce comparable scores among students taking the innovative assessments. Doing so will enable states to develop assessments that can better support teachers, leaders, and families in supporting students and their learning.
- **Recommendation 6: Utilize existing flexibilities and promulgate new regulations to allow for additional time to scale innovative assessment systems statewide.** The Department of Education could clarify and update regulations to provide states with additional time for planning, implementing, and scaling innovative systems.
- **Recommendation 7: Lift the cap on the number of states able to participate in IADA and allow for states to collaborate on assessment designs.** Should IADA become more attractive to states, only three additional states could currently participate. The Department of Education could prioritize completing the required Institute of Education Sciences (IES) report and eliminating the seven-state cap, allowing more states to take advantage of the opportunity.

## Create additional pathways to innovation

While IADA represents one major effort to create opportunities for assessment innovation, there are other ways the Department of Education can signal, incentivize, and support change. For example, the Competitive Grants for State Assessments (CGSA) program has been used to support states and multistate collaboratives in improving their state assessment systems. This grant program both provides funding and has fewer constraints than IADA, and it may be an effective avenue to support innovative state assessment efforts.

- **Recommendation 8: Use the CGSA program to stimulate individual or multistate efforts to develop and pilot new approaches that are instructionally useful and responsive to the broader view of assessment in ESSA.** The Department of Education could consider further leveraging CGSA—through both larger funding requests and strategic allocation of funding—to support innovations that specifically target assessment designs that seek to advance better teaching and learning.

A variety of federal executive action strategies could be implemented in the short term to encourage more innovative state assessment systems that better support teaching and learning, particularly as states work to support learning recovery related to the COVID-19 pandemic. Some strategies can help to strengthen state assessment systems in all 50 states under ESSA Section 1111(b). Other strategies can help to foster innovative assessments in the subset of states participating in the Innovative Assessment Demonstration Authority. In particular, those strategies include revising the options available for assessing comparability of new assessments in relation to standards for high-quality assessments generally and adjusting time frames to allow for design and scaling of new assessments. Additional headway can be made through an expanded CGSA program. A strategic approach could enable significant advances at this time, as the field is focused on dramatic improvements needed to support learning recovery, which require assessments more tightly tied to curriculum and instruction.

## Introduction

There is a growing call to reconsider current approaches to national and state assessment system policies and practices. The COVID-19 pandemic has heightened concerns among state and local education agencies, educators, and advocates that our current state assessment systems—defined primarily by end-of-year multiple-choice tests—are not sufficiently able to meet contemporary needs for information that supports teaching and learning. Concerns raised frequently by educators are that current state tests do not:

- offer actionable information when it is relevant to student learning;
- measure progress well for many students, since they are focused only on grade-level standards;
- measure learning across a coherent continuum, so as to better pinpoint what students know and are ready to learn next; and
- measure learning in ways that are culturally responsive and reflective of how students learn.

Furthermore, many state, district, and school-level leaders who are seeking to develop higher-order thinking skills associated with college, career, and civic readiness in the 21st century believe that current assessments poorly measure these competencies and are disconnected from a curriculum that seeks to develop these competencies, and thus do not incentivize aligned, high-quality teaching and learning. As one chief state school officer recently noted:

We've gone as far as we can go with our current large-scale assessment. We need new instruments that are more indicative of what kids know, and that also push our kids in the direction of better learning. ... If you have poor assessments, people will try to game them. If you have assessments that are high quality, people will focus on quality teaching and learning.

More than 20 states are involved in efforts to transform one or more aspects of their assessment systems;<sup>1</sup> however, the process of securing federal assistance and approval to make transitions to substantially improved systems poses numerous challenges. Among them are the costs and time required to change systems, the management of trend disruptions when new assessments are introduced, and interpretations of how to meet federal approval criteria under business rules that often keep new tests looking very much like old ones.

This report synthesizes policy analyses and findings from consultations with national, state, and local leaders to:

- outline the history of federal testing guidance and state responses that have shaped the current context;

- describe strategies states and districts are pursuing to develop high-quality assessment systems that both signal and support better teaching and learning processes for all students; and
- identify key ways the federal government could support assessment reforms that both enable thoughtful assessment of meaningful skills and better support teaching and learning.

## History of Federal Testing Guidance

Ever since the Elementary and Secondary Education Act was passed in 1965, each reauthorization of the federal education law has sought to move the schools closer to educational equity, with increasing emphasis in each reauthorization on educational assessments as a vehicle for this goal. In the 1990s, under the Improving America's Schools Act, states were challenged to establish standards for student learning and develop assessments that could begin to measure progress toward those standards. Many states (including Connecticut, Delaware, Kentucky, Maine, Maryland, Montana, Nebraska, New York, Vermont, and Wyoming) developed innovative approaches—much like those found in high-achieving jurisdictions around the world—that combine selected response and open-ended questions on sit-down tests with standardized performance tasks conducted in the classroom during the year, graded by trained raters (teachers) with standardized rubrics and back-reads for reliability so the performance-task results were part of the reported scores. These strategies were found to promote higher-order thinking skills, curriculum-connected assessment, and information about students' mastery of the standards for teachers to refer to during the year.<sup>2</sup> States generally constructed and managed their own assessments in-house. Testing time and costs were made more manageable by grade-span testing and matrix sampling on large-scale tests.

In 2002, the No Child Left Behind Act established the requirement that all students be tested annually and results be reported by student groups—defined by race, income, language background, and special education status. Although the law did not prohibit performance assessments or use of assessments during the course of the year, the George W. Bush administration's management of the law denied approvals to states using these methods (over the vehement objections of some, like Connecticut, which sued the federal government, and Kentucky, which returned each year with proposed strategies to maintain its well-functioning system) and encouraged states to use outside vendors offering primarily multiple-choice tests.

The shift in approach reduced the degree to which tests focused on critical thinking and performance skills. The RAND Corporation studied the tests that emerged during this era and found that, even in the 17 states with the highest standards, only 2% of items on mathematics tests and 21% of items on English language arts tests measured higher-level skills.<sup>3</sup>

The Obama administration sought to improve the quality of assessments by sponsoring the grant competition that ultimately produced the multistate consortia that developed the Partnership for Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced assessments, both of which expanded the use of more open-ended items and some modest performance tasks. Of these, Smarter Balanced survived as a consortium; it now serves 13 states and territories, plus the Bureau of Indian Education.

Although progress on assessment quality was made relative to the prior decade, the grant competition's requirements and the Obama administration's interpretation of those requirements limited the consortia's capacity for innovation. For example, the requirement that the assessments provide growth scores for use in teacher evaluations pushed the tests to the end of the year and limited through-course options. Additionally, the interpretations of how to measure the breadth and depth of the standards limited most items to quick responses, and the interpretations of how to measure grade-level standards prevented adequate assessment along the continuum of achievement that could both more meaningfully measure student growth and allow for more precise understanding of how to tailor instruction to student needs.

## **The Promise of the Every Student Succeeds Act for Advancing High-Quality Assessment**

In 2015, the Every Student Succeeds Act (ESSA) opened new possibilities, relative to the previous decade, for how student and school success are defined and supported in U.S. public education. The law deepened the concept of student learning to be more consistent with what students need to be successful in 21st-century society and careers, calling for measurement of "higher-order thinking skills and understanding" as part of "high-quality student academic assessments in mathematics, reading or language arts, and science."<sup>4</sup> ESSA intentionally creates opportunities for assessment innovation by explicitly allowing the use of multiple types of assessments, including "portfolios, projects, or extended-performance tasks"<sup>5</sup> as part of state systems. To support states in developing these more sophisticated types of assessments, ESSA also funds state innovation in assessment through the competitive state assessment grant program.

This explicit language in the law (see Table 1) opened the door for states to develop meaningful state assessment systems that are intentionally designed to improve teaching, learning, and, ultimately, student outcomes—and many states have been eager to take advantage of these opportunities. For example, some states have sought to develop state assessments that include performance assessment as a substantial component of their state systems, a strategy that research finds both creates greater curriculum equity—more students get access to high-quality teaching and learning experiences, as signaled by the state through the inclusion of rich performance tasks—and improves educational outcomes.<sup>6</sup>

**Table 1**  
**Summary of Assessment Requirements Under Section 1111 of the Every Student Succeeds Act**

SEC. 1111. STATE PLANS. 1111(b)(2)(B) Assessment Requirements	
<b>Assessments generally</b>	<p>(i) except as provided in subparagraph (D), be—</p> <p>(I) the same academic assessments used to measure the achievement of all public elementary school and secondary school students in the State; and</p> <p>(II) administered to all public elementary school and secondary school students in the State;</p>
<b>Information produced</b>	<p>(ii) be aligned with the challenging State academic standards, and provide coherent and timely information about student attainment of such standards and whether the student is performing at the student’s grade level;</p> <p>(xi) enable results to be disaggregated within each State, LEA, and school by—</p> <p>(I) each major racial and ethnic group;</p> <p>(II) economically disadvantaged students as compared to students who are not economically disadvantaged;</p> <p>(III) children with disabilities as compared to children without disabilities;</p> <p>(IV) English proficiency status;</p> <p>(V) gender; and</p> <p>(VI) migrant status...</p> <p>(xii) enable itemized score analyses to be produced and reported, consistent with clause (iii), to local educational agencies and schools, so that parents, teachers, principals, other school leaders, and administrators can interpret and address the specific academic needs of students as indicated by the students’ achievement on assessment items;</p>
<b>Technical requirements</b>	<p>(iii) be used for purposes for which such assessments are valid and reliable, consistent with relevant, nationally recognized professional and technical testing standards, objectively measure academic achievement, knowledge, and skills, and be tests that do not evaluate or assess personal or family beliefs and attitudes, or publicly disclose personally identifiable information;</p> <p>(iv) be of adequate technical quality for each purpose required under this Act and consistent with the requirements of this section, the evidence of which shall be made public, including on the website of the State educational agency;</p> <p>(xiii) be developed, to the extent practicable, using the principles of universal design for learning.</p>



**SEC. 1111. STATE PLANS. 1111(b)(2)(B) Assessment Requirements**

<b>Measurement methods</b>	(vi) involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding, which may include measures of student academic growth and may be partially delivered in the form of portfolios, projects, or extended-performance tasks;
<b>Administration flexibility</b>	(viii) at the State’s discretion— (I) be administered through a single summative assessment; or (II) be administered through multiple statewide interim assessments during the course of the academic year that result in a single summative score that provides valid, reliable, and transparent information on student achievement or growth;
<b>Assessment flexibility</b>	Locally selected, nationally recognized option for high school (H)
<b>Assessment design flexibility</b>	Computer adaptive assessments: (I) subparagraph (B)(i) shall not be interpreted to require that all students taking the computer adaptive assessment be administered the same assessment items; and (II) such assessment—“(aa) shall measure, at a minimum, each student’s academic proficiency based on the challenging State academic standards for the student’s grade level and growth toward such standards”; and “(bb) may measure the student’s level of academic proficiency and growth using items above or below the student’s grade level, including for use as part of a State’s accountability system under subsection (c).”
<b>Standards</b>	(A) IN GENERAL.—Each State, in the plan it files under subsection (a), shall provide an assurance that the State has adopted challenging academic content standards and aligned academic achievement standards (referred to in this Act as “challenging State academic standards”), which achievement standards shall include not less than 3 levels of achievement, that will be used by the State, its local educational agencies, and its schools to carry out this part.

Source: Every Student Succeeds Act. (2015).

## ESSA's Innovative Assessment and Demonstration Authority: Opportunity and Barrier

While ESSA's statewide assessment provisions (see Table 1) encourage all states to advance more innovative assessments that better support teaching and learning, the law also explicitly allows a subset of states to pursue innovation through the Innovative Assessment Demonstration Authority (IADA), Section 1204. This provision invites up to seven states to apply for an innovative assessment pilot to implement new approaches to assessment and gradually scale them statewide. An innovative assessment system within IADA is defined broadly as “a system of assessments that may include—(1) competency-based assessments, instructionally embedded assessments, interim assessments, cumulative year-end assessments, or performance-based assessments that combine into an annual summative determination for a student, which may be administered through computer adaptive assessments; and (2) assessments that validate when students are ready to demonstrate mastery or proficiency and allow for differentiated student support based on individual learning needs.”<sup>7</sup>

The primary promise of IADA is that it provides states a means to pilot new assessments without double testing students by allowing a subset of districts to use the new assessments instead of the old ones. This is an important feature of reform: States are not simply substituting one commercially administered standardized test for another at the end of the year (which many states have done outside of IADA).

While many states were initially pleased to have the opportunity to explore the flexibilities in the law through IADA, the technical complexity of doing so in ways that comply with IADA's requirements—in part due to the rules adopted by the Department of Education—has prevented states from realizing the promise of better assessment systems that ESSA envisions and IADA seems intended to enable. Few states have yet been able to develop systems that provide incentives and opportunities for teaching and assessing the more complex thinking, problem-solving, communication, and design skills students increasingly need to succeed in the rapidly evolving U.S. society and economy. Fewer still have been able to develop systems that provide insights into student learning throughout the year in ways that (1) are connected to the curriculum students are learning, and (2) provide support for educators to implement curriculum in ways that advance student progress.

As of early 2023, only five states—Georgia, Louisiana, Massachusetts, New Hampshire, and North Carolina—had applied for and received approval to pilot new systems under IADA. Those partners working with IADA states have learned that many of IADA's requirements are experienced as impediments to states' efforts to develop innovative systems. As one state education chief recently said, “IADA keeps us tied to the past” with respect to how assessments must be developed and administered. A seven-state limit on grantees could make it difficult for more states to apply for IADA. Yet it is unclear that many states would even attempt to do so, given the substantial burden to apply and constraints on innovation posed.

## Looking to the Future: Calls for Assessment Systems That Better Support Teaching and Learning

Many state leaders see the opportunities ESSA creates and want to transform their state assessment systems to take advantage of these affordances—with or without IADA. Through a series of conversations with state and local leaders as well as teachers and partners in the education space, a common set of goals for assessments that can inform and improve teaching and learning in schools is emerging. These common goals are:

- **Assessment tasks should encourage applied learning and higher-order skills.** Several state commissioners have noted that, in an effort to ensure that learners are covering all the standards efficiently, many statewide assessments have prioritized measuring discrete, disconnected pieces of information over investing in engaging, realistic tasks that would promote and support better teaching and learning—and, ultimately, provide better information about student progress.
- **Assessments should be integrated into a system that supports high-quality teaching and learning.** To support meaningful, deeper learning for students, assessment systems should be both designed and used as part of a coherent, well-integrated system of curriculum, instruction, and professional learning for teachers. Within such a system, assessments can intentionally support a vision for student learning and offer both incentives for that type of learning and information to support it during the year. A number of states are currently pursuing efforts to better integrate assessment with curriculum in service of learning, as described below.
- **Assessments should be part of accountability systems that support student access and success.** Current state assessment systems are embedded within decision-making and accountability frameworks that often do not meaningfully include factors beyond test scores that influence or reflect student success, such as indicators of opportunities to learn, the quality of teaching and learning experiences, students' contexts and school climate, and how well they are prepared for postsecondary success. State leaders and experts often posit that assessments should be part of improved accountability systems designed to encourage behaviors and actions that lead to a more informed focus on school improvement, more equitable access to learning opportunities, and greater student success. In addition, since federal accountability systems often operate separately from state and local systems for instructional improvement and with little coherence, accountability systems should be redesigned to actually help communities improve access to high-quality instruction by offering information to accompany assessment data about what students experience and with what results.

## Current State Efforts

States are pursuing several promising and innovative approaches to address many of these needs. Some of these approaches strengthen the capacity for high-quality instruction informed by formative assessment in schools within integrated curriculum frameworks that also inform summative assessments. Other approaches may position more innovative assessments to become part of the state summative assessment process itself, where sufficient comparability and safeguards are in place. In nearly all efforts, states seek to address the vision described above in ways that position the state assessment system to signal and incentivize what high-quality teaching, learning, and student performance should look like, while allowing for appropriate flexibility for local decisions. Strategies that states are pursuing include: (1) incorporating curriculum-embedded, performance-based approaches in statewide summative assessments; (2) designing statewide models for local performance-based assessments; (3) pursuing new assessment models that leverage existing data from student work; (4) connecting statewide assessment development efforts with capacity-building strategies for educators; and (5) exploring assessments as part of college and career pathways.

### **Incorporating curriculum-embedded, performance-based approaches in statewide summative assessments**

An important strategy for making assessments more meaningful and useful is to include curriculum-embedded performance assessments as part of the system. Performance assessments allow students to demonstrate what they know and can do by completing an authentic task that requires the use of targeted knowledge and skills. When performance assessments are designed to be embedded in the instructional process, these tasks can be administered when most appropriate to both student learning and teachers' curriculum plans.

Evidence suggests that curriculum-embedded assessments can support and incentivize high-quality teaching and learning as they:

- assess higher-order thinking and problem-solving skills;
- offer timely and instructionally relevant samples of student learning;
- provide empowering learning opportunities for students in and of themselves, enhancing curriculum equity and better preparing students for college and careers;
- can be used to support teaching and learning at the classroom level and guide decisions about curriculum and teacher development at the school and district levels;

- can provide professional learning opportunities for teachers through the development, implementation, and scoring of the tasks, along with better information for parents and families; and
- can offer comparable data about clearly defined areas of knowledge and skill when designed according to assessment development and validation standards.

Assessment programs in a number of countries (e.g., Australia, England, Ireland, New Zealand, Scotland, and Singapore, among others), as well as in the International Baccalaureate program and several Advanced Placement programs, demonstrate that performance-based assessments that are engaging, meaningful, and focused on sophisticated knowledge and practice can improve the quality of classroom learning.<sup>8</sup> Performance-based assessments can also create greater curriculum equity: Because all students must engage in the investigations and demonstrations of learning these assessments

Assessment programs in a number of countries, as well as in the International Baccalaureate program and several Advanced Placement programs, demonstrate that performance-based assessments that are engaging, meaningful, and focused on sophisticated knowledge and practice can improve the quality of classroom learning.

require (papers, presentations, design specifications, and other products), this kind of curriculum-embedded assessment (1) ensures that all students experience higher-level learning opportunities that prepare them for college and careers, and (2) encourages educators to develop teaching practices to ensure that students have routine access to this kind of instruction.

Using such tasks—often in conjunction with “on-demand,” “sit-down” tests that also require open-ended responses—also deepens educator understanding of teaching and learning when teachers are part of supporting students’ engagement with the tasks and are engaged in scoring efforts (typically evaluating the work of students outside their own classrooms). The tasks can be designed to provide instructionally useful feedback on student progress that can inform teaching and learning. They can also guide professional learning about high-quality curriculum, instruction, and assessment across classrooms.

The systems noted above—which offer centrally designed tasks and rubrics, training for administration and scoring, and audits where needed—also illustrate how curriculum-embedded performance tasks can be offered in ways that provide comparable data at scale.<sup>9</sup>

In the past, similar state-level examples in the United States have included the writing portfolios that have used common tasks and rubrics in Kentucky and Vermont; the common curriculum-embedded performance tasks that were part of the New Hampshire pace system; and the performance tasks used in Connecticut, Delaware, and Maryland that involved students in collaborative activities, such as science investigations, that they conducted together but wrote individually, allowing for individual assessment and reporting. Current and emerging examples of systemic use of performance tasks in states that contribute directly to summative scores include:

- **Smarter Balanced Assessment Consortium.** End-of-year assessments in mathematics and English language arts include performance tasks that are centrally scored and used as part of students' summative assessment scores. Additionally, the consortium is planning pilots that would separate the tasks from the end-of-year test and allow them to be integrated into the curriculum during the school year. This would allow the tasks to be more instructionally useful while still contributing to the overall score in appropriate ways.
- **Massachusetts's innovative science assessment.** Leveraging IADA, Massachusetts is developing a science assessment that uses a combination of more traditional assessment items and simulation-based performance tasks to engage students in activities that more closely approximate "doing science" as they might in the classroom, while still meeting the requirements, expectations, and constraints of on-demand, large-scale summative assessments. The state is also considering ways to include classroom-embedded tasks that happen throughout the year as part of the system.
- **New state assessments in science.** Like Massachusetts, many states are leveraging innovative item types and coherent item clusters to create science assessments that routinely engage students in closer approximations of "doing science," such as asking students to develop models of scientific phenomena, construct arguments from data, and pose solutions to real-world problems using scientific ideas and reasoning.
- **Maine English language arts/writing assessments.** In Maine, the state department of education is working with educators to reimagine the writing tasks on the statewide assessments, including (1) determining priorities among the writing standards that align with best practices in the classroom, and (2) developing appropriate writing tasks and prompts that align with those priorities and what teachers find more valuable for students.

### **Designing statewide models for local performance-based assessments**

States are also pursuing systematic and statewide use of local performance tasks—either locally developed and vetted or centrally developed for flexible local use—for formative and summative purposes. For example:

- **Colorado** has developed guidance and is supporting networks of educators in developing locally designed performance assessments that can be vetted by the state and used to meet the state’s graduation requirements.
- **Kentucky** developed parameters for local science assessments grounded in the Next Generation Science Standards that are designed locally and collected by the state for feedback to local educators and to inform the state summative assessment design.
- **Louisiana** has developed guidance for districts to use curriculum-embedded assessments for measuring “student learning targets”; some of the assessments are performance-based tools approved by the state, and others are constructed locally within the parameters of state guidance and curriculum.
- **Oregon** has established an annual local performance assessment requirement for grades 3–8 and high school in mathematics, scientific inquiry, speaking, and writing. The state provides some resources districts may use, such as sample tasks and general scoring rubrics, but districts have considerable flexibility in determining the nature of the performance assessments administered.
- **Vermont** provides educators with extensive supports for developing and using performance assessments as part of local assessment systems, including criteria, supports for culturally relevant performance tasks, and assessments that center learner agency.
- **Virginia** has rolled out a statewide performance assessment option in writing and history, enabling school divisions to use locally selected tasks in these areas.
- **Washington** has developed and made widely available several performance assessment resources for local use, including interdisciplinary tasks through the statewide ClimeTime initiative and curriculum-anchored assessment tasks in secondary science. Washington also requires the use of a state-developed curriculum-embedded civics performance task at least once in elementary, middle, and high school, leaving decisions about which task and when they are administered to local expertise.

### **Pursuing new assessment models that leverage existing data from student work**

Some school networks have begun to create systems to harvest the wealth of information gleaned through classroom-based instructional and assessment tasks to provide insight into student achievement. When data systems are in place to systematically elevate this kind of information directly from classroom practice, it can be used to inform instruction across classrooms, to support teacher professional learning, and as part of a multiple measures system for informing better programmatic decisions at the school, district, and network levels. For example, in

a competency-based assessment system like that employed by the Summit schools network, assessment is built directly into the model through a powerful technology platform that enables transparency and common interpretation of student work through common parameters and rubrics for tasks.<sup>10</sup>

One possibility system leaders may consider is how similar models might be used across schools and districts to surface indicators of student competence, as well as to provide information on the opportunity to learn across classrooms, schools, and districts.

### **Connecting statewide assessment development efforts with capacity-building strategies for educators**

Education leaders agree that a critical element of any effective assessment system moving forward is a dedicated capacity-building effort for educators. An added benefit seen in contexts where authentic performance assessments are used is that teachers become significant interest holders invested in the process of supporting students' engagement with the tasks and the resulting information and feedback that result from these new systems. As one state leader noted:

If we want to transform education for kids—to have students engaged, have their experiential learning stoked—it has to be about *engaging teachers*. You cannot have teachers be technicians who are just implementing; they need to be professionals who are sparking and creating the learning to light the fire in kids.

In New Hampshire, capacity building has included:

- developing processes, tools, and protocols for supporting districts and schools in creating and validating high-quality common and local performance tasks, along with guidance for teachers in how to use these tasks to enhance curriculum and instruction;
- assembling both the common and locally developed tasks into a web-based bank of validated performance tasks to be used for formative as well as summative assessments;
- organizing professional development institutes for cohorts of schools to support task design, validation, and reliable scoring, as well as data analysis to track student progress and inform instruction;
- building cohorts of expert teacher leaders in each content area to support this work; and
- creating regional support networks led by practitioner assessment experts to help build capacity in schools and to support regional task validation and calibration scoring sessions to achieve inter-rater reliability on locally scored tasks.



Similarly, in Maine educators are being engaged as co-developers of writing assessments, embedding capacity building directly into the assessment development process. In Virginia, the statewide performance assessment option in writing and history was rolled out in tandem with a capacity-building plan that will reach every school over a 5-year period.

## Exploring assessments as part of college and career pathways

Curriculum-embedded assessments are also emerging in many states (e.g., California, Hawaii, New Hampshire, New York, North Dakota, Ohio, Pennsylvania) and networks of schools (e.g., Linked Learning, Envision, New Tech, Asia Society) as part of specific high school pathways evaluating graduate competencies that are connected to college and career goals. These often feature shared assessment protocols and rubrics used to evaluate the accomplishment of competencies associated with specific disciplines and skill sets. The connection to college and career outcomes provides both an avenue for increasing authentic engagement and an additional validation measure for the assessment itself. Some of these systems have become sufficiently standardized to be used in higher education admissions. For example, a study<sup>11</sup> of the New York Performance Standards Consortium found that students in Consortium schools begin high school more educationally and economically disadvantaged than their peers and yet are more likely to graduate from high school, attend college, and persist in college than demographically similar peers. The study also found that, on average, students admitted to City University of New York (CUNY) as part of a CUNY–Consortium admissions pilot program that uses a holistic review process for admissions achieve higher first-semester college GPAs, earn more initial credits, and persist in college after the first year at higher rates than peers from other New York City schools, who, on average, have higher SAT scores. New systems for recording student learning, such as the Mastery Transcript, are emerging to capture these data.<sup>12</sup>

## Challenges to Address

While a common vision is emerging regarding how large-scale assessment systems can be positioned as useful to instruction, those seeking to enact assessment systems that reflect these priorities face several barriers. States and districts routinely identify the following challenges as areas for needed policy and resource support:

- **Time and resources.** Creating new systems of assessment that are high quality requires time for development, capacity building, stakeholder engagement, and implementation. Creating space and opportunities for states to pursue these efforts will require resources and support as new systems are being built. It should be noted that the need for resources is not always about substantially more spending; rather, what states and districts often need is support for more flexible use of existing funding, with some modest additional resources. For example, while the costs of scoring open-ended tasks are generally greater than the costs of scoring multiple-choice tests, they are no more than the costs

many districts are currently spending on interim tests and test prep focused on current lower-quality end-of-year multiple-choice tests.<sup>13</sup> One can argue that if that money is to be spent on testing, it could be more profitably spent on higher-quality assessments that stimulate deeper learning and contribute to professional learning (which has its own associated resources and costs).

Costs may also be mitigated by developing computer-based scoring of open-ended tasks as used in the Smarter Balanced Assessment Consortium's end-of-year assessments; Educational Testing Service's Cognitively Based Assessment of, for, and as Learning (CBAL) system of through-course assessment; and Council for Aid to Education's College and Work Ready Assessment. Similarly, in terms of timing, curriculum-embedded assessments take more time to administer (even though it is experienced as instructional time rather than "testing time") and require more resources to implement, given the needs for training and compensation of scorers, as well as audits of scoring processes if they are to produce comparable data across schools and districts. However, these activities associated with more time are also often directly useful for teachers: Scoring common assessments is powerful professional learning,<sup>14</sup> and figuring out how to implement curriculum-embedded assessments often leads to an evaluation of current curriculum and efforts to adapt instruction to better align with or build on the task demands.

- **Appropriate guardrails.** In scaling more innovative statewide systems, it will be important to create appropriate guardrails to ensure that the best approaches can be used to support all learners and that equity concerns for high-quality, comparable data, objective scoring of assessments, and equitable supports for student success are addressed.
- **Capacity building.** Curriculum-anchored assessment systems that improve instruction rely heavily on the capacity of educators to support learning of the kind measured by such tasks, as well as supporting development, implementation, and scoring of the tasks themselves—and developing capacity to learn from what they reveal. Moving toward scalable, instructionally relevant systems requires coherent statewide efforts to build teacher and leader capacity through professional learning tied to curriculum and assessment systems. Developing local teacher and leader capacity also requires investing in state education agency capacity to enable, develop, and sustain these systems.

## Supporting Next-Generation Assessment Systems: Recommendations for Federal Executive Action

Many of the challenges state and local leaders identify have to do with enabling conditions: States and districts have clear and compelling ideas about what would position assessments to better support teaching and learning, but they need time, support, and permission to innovate in those ways. This is an ideal role for the U.S. Department of Education. Several possible federal executive actions could encourage instructionally relevant innovative assessment systems. Some actions can strengthen all statewide assessment system efforts, while other recommendations focus solely on strengthening IADA implementation. All actions discussed here are permissible under current federal law.

For each action, this section outlines current requirements in law and regulation; describes why these requirements matter, lists challenges in implementation that are serving as a barrier to innovation; and offers possibilities for executive action (see Recommendations for Federal Executive Action).

### Recommendations for Federal Executive Action

Possible high-leverage actions the U.S. Department of Education could pursue, along with associated recommendations, include:

#### **Align technical requirements and peer review processes with ESSA's assessment allowances and requirements.**

- **Recommendation 1:** Highlight opportunities and update processes to support more instructionally relevant assessments that reflect student performance in relation to both grade-level standards and multiyear learning progressions.
- **Recommendation 2:** Highlight technically sound approaches to meeting federal peer review requirements that allow state assessments to assess the depth of state standards while ensuring sufficient coverage.
- **Recommendation 3:** Update peer review guidance to emphasize requirements for test security that are appropriate to the design of the assessment.
- **Recommendation 4:** Revitalize foundational elements of the peer review process—peer reviewer selection and moderation of the process—to ensure that states can take full advantage of the opportunities provided by federal law and the Department of Education's technical guidance.

#### **Enable IADA to better support innovation in assessment.**

- **Recommendation 5:** Update the interpretation of comparability of results within current IADA regulations to better enable high-quality innovative assessment approaches.

- **Recommendation 6:** Utilize existing flexibilities and promulgate new regulations to allow for additional time to scale innovative assessment systems statewide.
- **Recommendation 7:** Lift the cap on the number of states able to participate in IADA and allow for states to collaborate on assessment designs.

#### **Create additional pathways to innovation.**

- **Recommendation 8:** Use the Competitive Grants for State Assessments program to stimulate individual or multistate efforts to develop and pilot new approaches that are instructionally useful and responsive to the broader view of assessment in ESSA.

## **Align Technical Requirements and Peer Review Processes With ESSA’s Assessment Allowances and Requirements**

The Department of Education’s approval process for all state assessment systems is guided by interpretations of the law that have evolved over a number of years and a peer review process that relies on those interpretations to render judgments about state systems.<sup>15</sup> The peer review process includes guidance in the form of (1) critical elements of state assessment programs representing the technical requirements state assessment programs must meet, and (2) examples of evidence that detail the kinds of evidence states need to submit to show that their assessment programs comply with the critical elements.

The intent of ESSA is clear: It encourages high-quality assessments that (1) are designed to tell us how well students are meeting states’ adopted standards; and (2) yield data that can be trusted and compared across students, schools, and districts within states to help surface how students are being served. Assessments could be designed in several ways to meet these goals, and ESSA leaves most implementation of this intent to the Department of Education. For example, the law explicitly allows for:

- “multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding, which may include measures of student academic growth and may be partially delivered in the form of portfolios, projects, or extended-performance tasks”;
- “multiple statewide interim assessments during the course of the academic year that result in a single summative score that provides valid, reliable, and transparent information on student achievement or growth”; and
- measures of a “student’s level of academic proficiency and growth using items above or below the student’s grade level, including for use as part of a State’s accountability system.”

While the law encourages these more instructionally relevant assessment approaches, the peer review process often is unnecessarily constraining and disincentivizes the kinds of assessments ESSA encourages. The Department of Education’s interpretation of ESSA’s assessment provisions in the peer review guidance<sup>16</sup> reifies traditional standardized measures, privileging assessments that are administered once to all students, have a large number of relatively superficial items measuring many standards, are easily quantifiable, and can be rapidly machine scored without needing expertise to evaluate. In some cases, this is due to the language of the guidance itself; in other cases, states, technical partners, peers involved in reviewing state-submitted evidence, and staff within the Department of Education itself interpret what could be enabling opportunities in very narrow and limited ways.

More innovative approaches—such as those that sample student performance throughout the year and/or use performance assessments to provide evidence of student progress toward the more sophisticated expectations in most states’ standards—can meet and even exceed both ESSA requirements and professional standards for quality. However, they prioritize different elements of assessment design (for example, task validity and utility for understanding performance over speed of securing results; instructionally useful information from richer tasks that measure mathematical or scientific practices over coverage of a larger number of superficial standards from selected response items).<sup>17</sup> Because current interpretations of technical requirements are designed to work with more traditional assessments, states are often discouraged from pursuing what might be more useful assessment designs—designs that are used successfully in high-achieving countries around the world that better support higher-order thinking skills.<sup>18</sup>

For example, states that are interested in leveraging student portfolios or extended, curriculum-embedded performance tasks as a measure of students’ academic achievement cannot do so while meeting the peer review criteria for test security, which requires intense measures to ensure that teachers and students do not have knowledge of assessment content prior to administration. Similarly, states seeking to provide more useful information about student performance—either by assessing the depth of state standards or providing more precise information about student performance along multiyear learning progressions—cannot easily do so and still meet the Department of Education’s internal guidelines for grade-level standards coverage and subscores without developing an unreasonably long test. The functional result is that federal peer review guidance inadvertently limits opportunities that the federal law intentionally creates.

For states exploring innovation through IADA, the expectation that new assessments meet the same technical requirements as existing state assessments can be particularly limiting. State leaders have noted that, while they know what kinds of assessment changes are needed to better support learning, they are skeptical about investing the capacity and resources needed to pursue even modest innovations because of their experience or concern that their assessments (1) will not pass narrow

peer review interpretations, and/or (2) will not meet the Department of Education staff's interpretation of technical requirements for state assessment systems. In subsequent sections of this report, we discuss specific technical requirements that particularly limit innovative assessment activity (e.g., requirements around comparability and grade-level standards); however, these specific issues reflect a broader set of challenges with current technical expectations.

If the federal peer review process were updated to truly focus on high-quality assessments and data without privileging any one particular approach, state assessment programs could then be designed in ways that are technically strong (potentially stronger than existing assessments) and still support a range of (allowable) instructionally useful innovations states are considering. The recommendations highlighted below suggest ways to bring the peer review into alignment with the opportunities for instructionally relevant assessment allowable within ESSA.

**Recommendation 1: Highlight opportunities and update processes to support more instructionally relevant assessments that reflect student performance in relation to both grade-level standards and multiyear learning progressions.**

***Current ESSA requirements and federal technical guidance***

Currently, ESSA's assessment provisions—for both statewide assessments and assessments under IADA—allow for assessments that “may measure the student's level of academic proficiency and growth using items above or below the student's grade level,”<sup>19</sup> while requiring that assessments provide information “about whether the student is performing at the student's grade level.”<sup>20</sup> These requirements would allow the kinds of tests that extend along a learning continuum that were frequently used by states and districts prior to No Child Left Behind and that are still used in other nationally available tests today (e.g., NWEA MAP, I-Ready, Star).

Peer review guidance (Critical Element 2.1) requires that states submit evidence of how the items on a state assessment reflect the state's grade-level academic content standards, including test blueprints that align with the depth and breadth of a state's grade-level content standards. It also requires that states using computer adaptive assessments make proficiency determinations with respect to the grade in which the student is enrolled and use those determinations for all reporting. While this language does not explicitly require a particular percentage of items on a state's assessment be at grade level, the way this requirement is interpreted—for example, by technical partners conducting alignment studies—is that nearly all items must be aligned to specific grade-level standards and not include above- or below-grade content. For example, when the Smarter Balanced Assessment Consortium (SBAC) was created, technical guidance provided to SBAC leaders required that 80% of initial items students would encounter had to be aligned to grade-level standards. This was despite the computer adaptive capacity of the test to measure more precisely where students

are along a much longer learning continuum, while also reflecting where they are in relation to grade-level standards. In many instances, states are encouraged to ensure that close to 100% of items on their assessments are aligned to grade-level standards.

The operationalization of the guidance seems at odds with ESSA, which provides opportunities for states to assess student thinking through the course of the academic year, when students would still be developing grade-level understanding. This guidance also precludes (1) providing meaningful information about students who are performing near the bottom or top of the scale, and (2) the appropriate measurement of growth for most students, since there are insufficient items above or below grade level to allow for such measurement.

### ***Why it matters***

One feature of high-quality instructionally relevant assessments is that, when designed to do so, they can reflect student performance in relation to both grade-level standards and the multiyear learning progressions that underly most states' current standards in mathematics, English language arts, and science. For example, consider a group of 5th-grade students who are not yet performing at grade level. In current state assessments with an overemphasis on items aligned to grade-level standards, a test would be able to reveal simply that all of these students are not yet proficient—but would provide little useful information to students, their current or next year's teachers, or their families about which concepts and skills each of them has mastered and what next steps might be pursued to help students rapidly accelerate their learning by addressing their specific needs. In a system designed to surface student thinking along learning progressions (e.g., through an innovative through-year assessment), assessment data could tell teachers that student A has mastered some of the 2nd-grade standards and is ready to work on particular next steps on the learning progression, while student B has mastered most of the 4th-grade standards and could move on to particular content elements. In this case, the assessment scores would convey that students are not performing at grade level, as required by law, but would go beyond current measures in terms of the utility of those scores in informing meaningful instruction. Students A and B would benefit from different kinds of support, which this assessment could help teachers pinpoint—without losing sight of progress and proficiency relative to the grade-level standards.

While ESSA and current regulations clarify that individual assessment instruments—for both statewide and IADA assessments—can measure above and below grade-level standards (e.g., through computer adaptive assessments), the emphasis on generating annual summative proficiency determinations relative to grade-level standards discourages many states from designing systems that reflect where students are along multiyear learning progressions.

Understanding student learning across multiple years is particularly important now because of significant pandemic-induced learning disruption. Assessment design anchored in multiyear learning progressions allows for timely and relevant information that can guide learning acceleration. In contrast, knowing whether or not a student meets a cut score identified as “proficient” offers little information about how to support their learning.

Some research suggests how assessment systems that measure learning and knowledge gains along learning progressions rather than only grade-level standards can inform more useful teaching. For example, in a 2019 study of the use of individualized mathematics teaching tools,<sup>21</sup> schools that were able to measure student learning along a multiyear continuum and identify—and therefore teach to—the specific skills students had yet to develop saw an average gain in math achievement of 38 percentile points over 3 years. In contrast, those that were restricted by their school or district policy to focusing on grade-level standards alone saw a gain of only 7 percentile points over that same period of time, presumably because they were unable to help students solidify and then build on their understanding of skills and concepts they had previously missed.

It is worth emphasizing that current state standards in mathematics, English language arts/literacy, and science have been intentionally designed along multiyear learning progressions. This feature of standards design means that assessments can also be designed to reveal student progress along progressions and use that determination, along with strategic items reflecting grade-level standards, to reveal proficiency relative to grade-level standards while revealing the more precise performance information that could help guide instructional planning. This approach meets both the spirit of the law and classroom teaching and learning needs. Unfortunately, states have not been supported in using this design feature of the standards to use more precise instruments that provide useful information about student progress and proficiency.

### ***Possible actions***

- **Clarify that systems that generate information about both student progress along multiyear learning progressions and proficiency relative to grade-level standards are permitted and encouraged under ESSA.** Updated technical guidance could clarify the value of assessments having developmental scales that extend across grade levels and use of scale scores that allow for better measurement of growth in addition to proficiency cut scores. Federal guidance can highlight this interpretation of the appropriate standards goals for state assessments by explicitly describing what taking advantage of this opportunity could look like. Functionally, this might be achieved by (1) highlighting how sampling strategies could be leveraged to allow for assessments that tap a



broader range of content across grade levels and (2) including examples of systems that incorporate reporting against both grade-level standards and multiyear descriptions of student proficiency.

- **Update guidance for internal quality processes, such as peer review, to allow for more precise approaches to determining student proficiency that pinpoint student understanding along multiyear learning progressions.** As noted above, we have learned that states are reluctant to explore assessments designed against multiyear learning progressions because of concerns about failing quality control processes, such as peer review. Updated guidance and technical requirements, both those written for public use and internal rules as interpreted and used by Department of Education staff, can ensure that states are able to develop the most instructionally useful assessments possible while still providing data on proficiency relative to grade-level standards.

**Recommendation 2: Highlight technically sound approaches to meeting federal peer review requirements that allow state assessments to emphasize assessing the depth of state standards while ensuring sufficient coverage.**

*Current federal technical guidance and interpretation*

The expectation for alignment to the depth and breadth of content standards appears in the guidance for peer review in the following critical elements and associated examples of evidence:

- Critical Element 2.1: “The State’s test design and test development process is well-suited for the content, is technically sound, aligns the assessments to (1) the depth and breadth of the State’s academic content standards for the grade that is being assessed; or (2) the depth and breadth of the State’s ELP [English language proficiency] standards.”
- Critical Element 3.1: “The State has documented adequate overall validity evidence for its assessments consistent with nationally recognized professional and technical testing standards. The State’s validity evidence includes evidence that: The State’s academic assessments measure the knowledge and skills specified in the State’s academic content standards, including:
  - a. Documentation of adequate alignment between the State’s assessments and the academic content standards. The assessments are designed to measure in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity.
  - b. Documentation that the assessments address the depth and breadth of the content standards.”

The language in these peer review elements, as well as the associated examples of evidence, is largely reasonable. Taken at face value, it allows for strategic decisions about how content is represented on a state's test, and indeed emphasizes that states should submit evidence to demonstrate that the state uses approaches that "include challenging content and complex demonstrations or applications of knowledge and skills (i.e., items that assess higher-order thinking skills, such as item types ... that require synthesizing and evaluating information and analytical text-based writing or multiple steps and student explanations of their work)."

While the peer review guidance itself is not inherently limiting, the ways that this guidance has been interpreted by state education agencies, technical partners, and staff within the Department of Education often lead to an overemphasis on broad coverage and assessments that typically fail to measure the depth of standards appropriately, such as those elements of standards associated with mathematical and scientific practices, as well as research, writing, and analysis. For example, most states' test blueprints emphasize having enough 1- or 2-point items that cover as many standards within the discipline as possible to show sufficient representation of the domain to pass peer review.

This consistent interpretation is due, at least in part, to the following factors:

- **Subscores:** Peer Review Critical Element 3.3 states, "The State has documented adequate validity evidence that the scoring and reporting structures of its assessments are consistent with the sub-domain structures of the State's (1) academic content standards." By requiring subscores that reflect the organization of standards (for example, geometry and algebraic reasoning in math standards) the peer review guidelines have tended to result in states believing they should prioritize superficial coverage of breadth such that they have sufficient score points to generate each subscore, rather than being able to make principled decisions about lighter sampling in some areas to enable deeper assessments in others, as appropriate to the intended use of the assessment.<sup>22</sup>
- **Balance of content:** Peer review guidance for Critical Element 2.1 suggests that the evidence states submit to show that their tests align with the depth and breadth of standards include test blueprints that reflect the range of knowledge, cognitive process, and cognitive complexity expected by standards. This language is often interpreted to mean that the full range of a state's standards within a grade level must be represented on a state assessment, rather than a principled and appropriate sample.
- **Alignment methodologies:** States must submit an independent alignment study as part of the peer review process. Many of the commonly used alignment methodologies (e.g., Webb's methodology) prioritize standards coverage as the primary indicator of alignment, rather than a meaningful look at whether the test, in aggregate, surfaces trustworthy measures of student understanding

of the depth and breadth of the state's standards. Evaluation of alignment typically ignores how evidence is aggregated for purposes of drawing inferences and reporting.

- **Development practices:** Common large-scale assessment development practices are both implicitly and explicitly grounded in approaches that focus on low-level items and standards coverage over depth, such as an emphasis on what can be readily assessed using selected response items, and on using item difficulty statistics to exclude many items that students find challenging, often regardless of whether those items are meaningful representations of expected learning. This focus is often exacerbated by the people involved—many assessment experts providing technical assistance to states have worked largely under policies from the No Child Left Behind era, and the assumptions and trade-offs made under those assessment programs often play an implicit role in how expectations for assessments are interpreted and how assessments are designed to meet those expectations.
- **Tacit approval by the Department of Education:** Many state education agencies look to what has been approved by the Department of Education as an implicit guide to passing peer review. Without clear examples of a range of ways to meet peer review criteria, certain approaches become the standard operating procedure.

### ***Why it matters***

The peer review process, however inadvertently, yields assessments that are a mile wide and an inch deep; because there are so many items assessing isolated knowledge and skills, it is nearly impossible for this approach to appropriately measure the sophisticated understanding and abilities reflected in state standards (e.g., developing complex arguments, solving realistic problems). Indeed, it is common practice in state assessment design to either accept superficial measures of depth (e.g., Depth of Knowledge analysis) or simply to not assess those standards (e.g., disciplinary practices, content application expectations) that require higher-order thinking and deep application. Even though these features are at the center of current math, English language arts, and science standards, tasks requiring disciplinary reasoning and problem-solving are often left out of state assessments because states, assessment developers, and technical partners know that an assessment without measures of those features will be easier to develop and sufficient—or even better able—to pass peer review.

While many states are seeking to develop innovative assessments that prioritize opportunities for students to engage in higher-order thinking, problem-solving, and disciplinary reasoning, they do not feel empowered to do so with so many factors driving assessments toward traditional, breadth-focused designs.

### *Possible actions*

- **Update the critical element language to reflect more purposeful and strategic expectations for alignment.** One approach might be to update the critical element language within the peer review guidance to acknowledge that no assessment—particularly a single on-demand end-of-year assessment—is able to assess the full depth and breadth of the standards within reasonable time allotments. Instead, the critical element could require documentation that the assessments *strategically* sample the depth and breadth of the content standards, as needed to represent standards and, importantly, to meet the purposes and intended interpretive uses of the assessment. This could be further supported by removing the expectation of subscores that mirror the structure of the discipline, allowing states to make more intentional decisions about what they test and report.
- **Provide examples of innovative designs that meet current requirements.** A key need is specific guidance and examples that would meet statutory requirements in ways that highlight the flexibility available to states. States have shared that it would be valuable if the Department of Education could highlight instances using alternative alignment or assessment design approaches that have passed peer review. For example, the Department of Education could highlight examples that focus on the claims the assessment is designed to make or how well the assessment meets commonly agreed-upon criteria for high-quality assessments, rather than on a decontextualized readout of the percentage of items that cover specific grade-level standards.

### **Recommendation 3: Update peer review guidance to emphasize requirements for test security that are appropriate to the design of the assessment.**

#### *Current federal technical guidance*

Peer Review Critical Element 2.5 requires that “the State has implemented and documented an appropriate set of policies and procedures to prevent test irregularities and ensure the integrity of test results through: Prevention of any assessment irregularities, including maintaining the security of test materials (both during test development and at time of test administration), proper test preparation guidelines and administration procedures, incident-reporting procedures, consequences for confirmed violations of test security, and requirements for annual training at the district and school levels for all individuals involved in test administration.”

#### *Why it matters*

While the overall intent of the critical element is important, the conception that the only way to achieve integrity of assessment is through item secrecy is unnecessarily narrow and—given the availability of technology and social media that can be used to share memorable tasks and items—increasingly difficult for states to implement. There are other ways to ensure that tests are “secure.” For example,

some certification programs achieve test security by simply releasing the full pool of possible items, making it nearly impossible for test-takers to memorize the answers without understanding the targeted content. Similarly, the GRE Analytical Writing assessment prompt pool is publicly available—this does not invalidate the test, since all candidates must nonetheless develop the skills needed to write those types of issue and argumentative essays if they are to successfully demonstrate competency on the prompts they are assigned. As the goals of schooling move beyond memorization of transmitted information to the development of performance abilities, older notions of test security are increasingly out of date.

Furthermore, states exploring the use of authentic, high-quality classroom performance tasks as part of their state systems may want such tasks to be openly used, discussed, and interpreted by students and teachers in keeping with making assessments more instructionally relevant. In these cases, states may make a wide variety of tasks available, using only a subset on the assessment, or may focus on realistic tasks, such as carrying out investigations or conducting research, that are not compromised when students and teachers know what students are asked to accomplish. For example, conducting a science investigation in which students must collect and analyze original data is not undermined by knowing the nature of the inquiry ahead of time, as long as the investigation requires sense-making using science ideas and practices.

Importantly, test security becomes increasingly irrelevant as the tests themselves involve increasingly authentic performances. For example, the specifics of each state's driving performance test are known and practiced, but the test is nonetheless valid because the components of it must still be accomplished. When an assessment task represents a category of performance that is itself a complex, valued activity in the tested content area (e.g., writing, modeling, conducting an investigation, assembling and analyzing data and making a persuasive argument about its meaning), students' ability to demonstrate the targeted knowledge and skill is not invalidated by their knowledge of the task in advance. In fact, it could be argued that insofar as the prior knowledge of a task encourages teachers and students to practice and develop the associated concepts and skills, this knowledge is actually how high-quality assessments should guide learning. Such knowledge becomes a problem only when the items on a test can be easily "gamed" through memorization (i.e., the tasks are designed such that advance knowledge renders generalizations from performance on that task unsupported) or when the task is susceptible to test-taking strategies that mask true understanding.

This fact is particularly important when considering alternative approaches to test security. For example, the performance task in the Advanced Placement (AP) Computer Science Principles class (see Figure 1) asks students to invent, develop, test, submit, and explain a computer program as part of their final score on the AP exam, which is used to confer college credit to students. This task is widely known and is completed (in part collaboratively) as part of classroom activities, but because knowing what the task is does not compromise interpretation of students' performance, the lack of test security does not compromise validity.

**Figure 1**  
**Advanced Placement Computer Science Curriculum-Embedded Performance Task**

## Create Performance Task




Programming is a collaborative and creative process that brings ideas to life through the development of software. In the Create performance task, you will design and implement a program that might solve a problem, enable innovation, explore personal interests, or express creativity. Your submission must include the elements listed in the Submission Requirements section below.

You are allowed to collaborate with your partner(s) on the development of the program only. **The written response and the video that you submit for this performance task must be completed individually, without any collaboration with your partner(s) or anyone else.** You can develop the code segments used in the written responses (parts 3b and 3c) with your partner(s) or on your own during the administration of the performance task.

*Please note that once this performance task has been assigned as an assessment for submission to College Board, you are expected to complete the task without assistance from anyone except for your partner(s) and then only when developing the program code. You must follow the Guidelines for Completing the Create Performance Task section below.*

### General Requirements

You will be provided with a minimum of 12 hours of class time to complete and submit the following:

-  **Final program code** (created independently or collaboratively)
-  **A video that displays the running of your program and demonstrates functionality you developed** (created independently)
-  **Written responses to all the prompts in the performance task** (created independently)

Scoring guidelines and instructions for submitting your performance task are available on the [AP Computer Science Principles Exam page](#) on AP Central.

*Note: Students in nontraditional classroom environments should consult a school-based AP Coordinator for instructions.*

Source: CollegeBoard. (2020). *AP computer science principles: Course and exam description*. <https://apcentral.collegeboard.org/media/pdf/ap-computer-science-principles-course-and-exam-description.pdf>

These kinds of solutions are becoming more important as an increasing number of states consider how to use realistic performance tasks as part of their assessment programs. Test security measures should match the intended purpose and design of a given assessment. By expanding the critical element to allow other approaches to test security in service of trustworthy assessment results, states seeking to use authentic and memorable tasks, to provide teachers and students with appropriate choices, and to provide teachers with access to tasks and student work in ways that can more directly support their teaching and learning can more readily enact these innovations.

### *Possible actions*

- **Revise the critical element language in peer review guidance to focus on ensuring high-quality and trustworthy scores.** The Department of Education could modify the language of the critical element to focus on policies and procedures to protect validity without imposing a particular approach to item security. This would allow states to more intentionally match their approach to test security to their test and task design and intended use.
- **Provide examples of alternative approaches to maintaining the integrity of state assessment programs.** It would be helpful if the Department of Education shared (e.g., through examples of evidence provided as part of the peer review guidance or through a separate guidance document) examples of alternative ways to maintain the integrity of state assessment programs. These examples could include (1) approaches to item formats that limit the likelihood of memorization (e.g., surface variations in task context or answer choices) and (2) alternative ways to maintain assessment integrity that are more performance-based, so that memorization is irrelevant (e.g., performance assessments used by certification programs, international K–12 assessments, and other large-scale programs, such as Advanced Placement, International Baccalaureate, and postsecondary and graduate admissions tests).

### **Recommendation 4: Revitalize foundational elements of the peer review process—peer reviewer selection and moderation of the process—to ensure that states can take full advantage of the opportunities provided by federal law and the Department of Education’s technical guidance.**

#### *Current process*

As described above, states are required to submit evidence for peer review showing that their statewide summative assessments meet all requirements under ESSA. The Department of Education recruits and convenes rotating panels of experts in the field of educational standards and assessments. Peers are selected based on “the individual’s experience and expertise, with an emphasis on knowledge of technical aspects of large-scale assessments, experience with the operation of state assessment systems, and relevant specialized expertise.” Peer reviewers are often those who have served in peer review roles (in this capacity or other related capacities) for the Department of Education before, or been recommended by staff or the educational field.<sup>23</sup> The peers are charged with determining whether the evidence submitted by a state meets the peer review criteria, which often relies on compliance with conventional psychometric standards.

## ***Why it matters***

Many of the recommendations outlined represent high-leverage changes to the technical guidance provided by the Department of Education that could serve to encourage more innovative assessments. However, two important elements of the federal peer review process must be addressed globally for the prior recommendations to have impact:

- updating the psychometric standards that underlie many notions of quality and sufficient evidence embedded within the peer review criteria, and
- expanding the people and processes involved in operationalizing updated guidance.

Without attention to these foundational components of the peer review process, even the most profound changes run the risk of being words on a page. For these recommendations to enable innovation, it will be essential that the peer review process be updated in terms of what psychometric conventions are held up as the appropriate standard, the people selected as peers, moderation and calibration of peer review, and state and partner engagement during the process.

## ***Possible actions***

- **Convene a task force to define the range of appropriate psychometric approaches for statewide summative assessments.** Many concerns with the limitations of current assessment programs, including those described above, stem from the ways in which conventional psychometric standards for testing in the United States have been used to prevent states from making commonsense decisions about better assessments. All assessment decisions reflect a set of trade-offs. Current Department of Education interpretations of psychometric standards have required trading off depth in service of coverage, validity of test scores in service of reliability, and instructionally useful data in service of scores that can be used to sort and label students and schools. While this may be appropriate for some assessments designed for some purposes, by basing compliance with federal peer review on a particular set of psychometric conventions, states are given very little choice in pursuing a different purpose for assessment that could better meet expectations within ESSA while meeting psychometric standards appropriate to those purposes.

The Department of Education could convene a task force to explore the appropriate psychometric guardrails needed for innovative assessments designed to support high-quality teaching and learning. Activities the task force may consider could include exploring the technical approaches taken in international assessment systems that intentionally make different psychometric trade-offs, considering more contemporary models for and evidence of testing quality, outlining special studies to be conducted by measurement leaders, and outlining opportunities to explore innovative psychometric models under current and future innovation demonstration authorities.



- **Ensure that selected peers consistently include experts in innovative assessment practices, such as performance assessments and their associated psychometrics, multipronged assessment systems, and alternative alignment approaches.** Operationalizing the flexibilities suggested above will require peers who can recognize high-quality, innovative systems that maintain the intent and integrity of state assessment systems under law. The Department of Education can intentionally diversify the peers recruited to ensure that experts in innovative, instructionally meaningful assessment are consistently represented and that they constitute enough of each peer review panel to advocate for innovative assessment designs that may be unfamiliar to more traditional assessment and measurement experts.
- **Carefully moderate peer review, including paying attention to flexibility intentionally embedded within the requirements.** In addition to selecting the right peers, the Department of Education can guide the peer review process—including training for peers, moderation within review discussions, and calibration across different panels—in ways that intentionally highlight the features of the law that invite more instructionally supportive assessments and the new and expanded expectations recommended above. For example, the Department of Education could focus peer training and calibration on the kinds of documentation that might be submitted to show alignment to depth of standards, and emphasize how this alignment evidence could be different than prior experiences. As part of this effort, the Department of Education could ensure that examples of systems that meet the peer review guidelines in innovative and more effective ways are provided as part of reviewer training and calibration activities, as well as within the guidance throughout the process.
- **Establish clear and open communication between the Department of Education and states/partners to communicate expectations, goals, and potential flexibility.** Perhaps the most important constituencies for the Department of Education to reach are states and their partners who are working on innovative designs. In addition to open and routine communication with states, the Department of Education could consider highlighting examples and communicating changes in expectations to those partners—assessment developers, technical assistance agencies, and alignment study providers—who are intimately involved in assessment design decisions.

## Enable IADA to Better Support Innovation in Assessment

While the inclusion of IADA within ESSA was first met with excitement by states, this optimism has waned as states recognized the limitations to innovation that are embedded within the authority's requirements. Put plainly, IADA does not currently offer states enough opportunity and flexibility to make the tremendous effort needed to create new assessment systems worthwhile. In fact, many of IADA's requirements are viewed as onerous and may actually limit efforts to develop innovative systems.

The recommendations below highlight ways executive action could shift the cost-benefit trade-offs to both open opportunities for innovation and remove barriers to state participation.

### **Recommendation 5: Update the interpretation of current IADA regulations to better enable high-quality innovative assessment approaches.**

#### ***Current requirements***

Multiple, inconsistent uses and interpretations of “comparability” within ESSA pose a barrier to states pursuing innovative assessment approaches. ESSA does not define the term “comparability,” and the term is not used consistently within the legislation. “Comparability” is explicitly used in two different contexts within ESSA: (1) with respect to the option to choose a national high school assessment in lieu of one associated with the state standards, and (2) within requirements for IADA. As described above, comparability is also implied within certain elements of the federal peer review guidance.

**Comparability in ESSA (generally).** With respect to statewide academic assessments pursuant to Section 1111, the statute requires that any locally selected, nationally recognized high school assessments approved by the state as an alternative to the state assessment “provide comparable, valid, and reliable data on academic achievement, as compared to the State-designed assessments, for all students and for each subgroup of students.”<sup>24</sup>

**Comparability within IADA.** Within IADA requirements, “comparability” has several meanings.

- *Comparability of coverage and difficulty.* Authority to use innovative assessments can be withdrawn if, at any time during its demonstration period, the state education agency cannot “demonstrate comparability to the statewide assessments ... in content coverage, difficulty, and quality.” This requirement suggests that the new assessments must measure the same things as the old assessments and do so in ways that do not result in more or fewer students doing well.
- *Comparability of task scoring.* ESSA also requires that states applying for IADA describe how they will “engage and support teachers in developing and scoring assessments that are part of the innovative assessment system, including through the use of high-quality professional development, standardized and calibrated scoring rubrics, and other strategies, consistent with relevant nationally recognized professional and technical standards, to ensure inter-rater reliability and comparability.”<sup>25</sup>

- *Comparability of results.* ESSA requires that states applying for IADA demonstrate that their innovative assessment systems “generate results that are valid and reliable, and comparable, for all students and for each subgroup of students described in Section 1111(b)(2)(B)(xi), as compared to the results for such students on the State assessments under Section 1111(b)(2).”<sup>26</sup>

IADA regulations<sup>27</sup> go beyond the statute in how they define comparability of results between the innovative assessment and the currently administered state assessment in their requirement that states annually determine comparability of results. IADA regulations specify that state education agencies (SEAs) must plan to determine comparability of results during each year of the demonstration authority period in one of five ways:

1. “Administering full assessments from both the innovative and statewide assessment systems to all students in participating schools [at least once per grade span]”
2. “Administering full assessments from both the innovative and statewide assessment systems to a demographically representative sample of all students and subgroups of students [at least once per grade span]”
3. “Including, as a significant portion of the innovative assessment ... items or performance tasks from the statewide assessment system”
4. “Including, as a significant portion of the statewide assessment system ... items or performance tasks from the innovative assessment system”
5. “An alternative method for demonstrating comparability that an SEA can demonstrate will provide for an equally rigorous and statistically valid comparison between student performance on the innovative assessment and the statewide assessment, including for each subgroup of students”

### ***Why it matters***

It is essential to ensure that a given assessment provides comparable standards-aligned tasks that generate comparable student scores across schools, students, and groups of students. Comparable assessments help schools, districts, and states ensure that all students, regardless of race, income, or zip code, are supported in finding academic success. However, some approaches to comparability—for example, requiring comparable scores from old and new tests even though they measure different things—prevent states from developing higher-quality assessment systems intended to better support and serve the very students comparability requirements are intended to protect. Of all the different ways that comparability can be considered or defined, the one that is most problematic for innovation is this expectation that a new system of assessment will generate comparable results with the pre-existing system it is seeking to replace. This criterion is essentially asking to what degree the

results from one assessment (e.g., a state’s innovative assessment) mirror the results on a different assessment (e.g., the current statewide assessment). Put another way, for student subgroups, how confident could you be that a student taking one assessment would get the same score or proficiency level on the other assessment?

If comparability is defined as generating the same results across the innovative and current state assessment, the innovative assessment will be automatically constrained by the same design and reporting decisions of the current summative test, including any limitations in surfacing useful student data. This approach positions the current state assessment as producing the “right” results, even as many efforts for innovation seek to produce better, more meaningful results that can be used to support teaching and learning more effectively. For example, many states pursuing assessment flexibilities or innovative assessment systems are seeking to measure deeper learning and other standards not well represented in their current tests (e.g., problem formulation, investigation, data set development and analysis, and writing and speaking). These innovations should be expected, by design, to produce *different* student scores. Thus, the requirement for comparability of results may inadvertently serve to prevent higher-quality assessments from being developed.

State leaders consistently note that IADA’s requirements around comparability of results are burdensome, unnecessarily stifle innovation in assessments, and place an undue emphasis on result comparability that does not allow states to accomplish ESSA’s assessment aspirations. For example, in the suggested approaches to establishing comparability explicitly described within IADA regulations (1–5 above), methods 1 and 2 burden pilot schools and districts, and their students, with double testing and place additional costs on the state in implementing parallel systems. Further, it is difficult to build capacity and buy-in among educators and local leaders for a new assessment approach while they are still required to participate in and are held accountable to the traditional tests. Methods 3 and 4, with their requirements of a “significant portion” of shared items across the statewide and innovative assessment systems, substantially constrain assessment design, or, if used in addition to innovative items and tasks, can extend testing time in ways that limit the viability of truly innovative designs. Method 5—which allows states some flexibility to demonstrate comparability through an alternative method—has not been defined and does not seem to be generally utilized by the Department of Education to support states. States report that while the regulations allow for a “state-developed” option, in practice, only methods 1–4 (described above) are treated as acceptable options.

Other commonly accepted ways to conceptualize comparability would be appropriate for maintaining assessment quality without this chilling effect. These include:

- *Comparability of assessments with respect to the standards measure.* That is, to what degree do different assessments measure the same learning goals?

- *Comparability of tasks and scores across students and schools.* That is, to what degree can students' assessment scores be interpreted in similar ways across students and groups of students (e.g., across different administration contexts and disaggregated subgroups)?
- *Comparability of task scoring.* That is, for a single assessment (e.g., a performance-based assessment), how likely is it that a student's essay would receive the same score from two different raters?

Under these conceptualizations, a state seeking to design an innovative, instructionally relevant system of assessments can explore the use of high-quality curriculum-embedded performance tasks in an assessment design that assesses the same standards as the current summative assessment, but does so in a way that allows for students to demonstrate aspects of the standards (e.g., more evaluative and critical thinking, writing ability, and problem-solving capacity) that were not assessable on the current test. Because students would be demonstrating different skills, one might expect that the innovative assessment could provide markedly different—and more instructionally productive and useful—information about student performance. The same may be said of states seeking to develop culturally responsive assessments. If those assessments are successful in allowing students to more effectively show what they know and can do, there should be some changes in subgroup performance.<sup>28</sup>

### **Possible actions**

- **Develop guidance on the fifth option under IADA regulations to allow states an alternative method for demonstrating comparability that focuses on the quality and standards alignment of the assessments and the comparability of their tasks and scoring across students and schools.** Rather than comparing the results of new assessments to the existing state tests that may be of lower quality and utility, this recommendation would further the goal of high-quality assessments designed to support teaching and learning, while ensuring transparency and accountability for results for all students and subgroups. For example, alignment methodologies can use common criteria for high-quality assessments, such as those developed by the Council of Chief State School Officers, to both evaluate individual programs and compare across programs. Specifically, a comparability requirement that is focused on ensuring that the assessments are high quality and aligned to standards would allow new assessments to be more challenging and sophisticated while also illustrating how they are aligned to the same core content.

For example, curriculum-embedded performance assessments—such as those used by new Advanced Placement courses and the International Baccalaureate programs—can provide comparable, high-quality data at scale. These assessment systems use centrally designed tasks and rubrics, training for administration and

scoring, and back-reads or audits where needed to allow instructionally relevant performance assessments to be used as part of a valid, comparable, and reliable data system.

Language for guidance on the fifth option for demonstrating comparability under IADA might be updated to state: “(5) an alternative method for demonstrating comparability that an SEA can demonstrate will provide for a rigorous and valid comparison between the innovative assessment and the statewide assessment, such that:

- a. the innovative assessment is demonstrated to be of equal or higher quality than the state assessment in terms of measurement of academic standards, and
- b. all tasks and scoring processes on the innovative assessment generate valid and reliable measures of student performance that are comparable across students, schools, and districts engaged in the innovative assessment.”

- **Ensure guardrails and protections for student subgroups by asking states to describe how the innovative assessment will accomplish its goals while maintaining comparability and reliability in assessment scores across students.** Scaling innovative assessment systems requires appropriate guardrails to ensure that innovation does not undermine equity or quality. It will be important to ensure that concerns for high-quality, comparable data, objective scoring of assessments, and equitable supports for student success are addressed. For example, states should explain how the differences in the new system will (1) accomplish important goals for information about learning and growth, (2) do so in ways that preserve comparability and reliability in assessment scores across schools, *and* (3) support fairness in opportunities for students to demonstrate their knowledge. This assurance should be done in a systematic way so that the intended protections and integrity remain intact while allowing for new systems that may generate more useful information.
- **Revise the technical guidance for state assessments to focus on a clear and appropriate definition of comparability.** Assessments developed under IADA will need to meet federal peer review guidelines, offering another opportunity to clarify the intended emphasis on comparability within state assessment systems. The Department of Education could update the language of peer review critical elements and evidence for state compliance to focus on (1) comparability of instruments based on alignment between the test design and its intended claims and uses, (2) robust rubrics and task scoring processes that yield reliable and comparable scoring of open-ended tasks, and (3) evidence that comparable determinations about student proficiency can be made across students and groups. It may also be helpful to position standardization of test administration

as just one example or way of demonstrating comparability (rather than as the critical element in its entirety) and include examples of other ways states may demonstrate comparability aligned to the specified definition.

- **Highlight examples that focus on establishing comparability through alignment to standards and intended interpretation of student performance.** Useful examples may highlight how (1) student and educator choice, (2) assessments aligned to different high-quality curricula, and (3) a range of accommodations and options for test implementation can be part of robust, fair, and trustworthy state assessment systems. Furthermore, it could be useful to highlight examples of processes found to produce high levels of scoring comparability, such as through training raters and auditing of cross-site scoring.

### **Recommendation 6: Utilize existing flexibilities and promulgate new regulations to allow for additional time to scale innovative assessment systems statewide.**

#### ***Current requirements***

ESSA requires states participating in IADA to scale innovative assessment systems statewide within 5 years,<sup>29</sup> while allowing for the possibility of a 2-year extension<sup>30</sup> as well as an additional undetermined amount of time with a waiver.<sup>31</sup> However, IADA regulations state that the Secretary of Education may grant only a 1-year extension after the 5- to 7-year period.<sup>32</sup> IADA regulations are silent on whether a planning year is allowed.

#### ***Why it matters***

IADA regulations governing the time to scale—allowing only a 1-year extension—have served as a barrier to innovation. Innovative systems—even those that begin with a preliminary design—take substantial time and resources to develop, pilot, refine, and scale. This time includes (1) engaging stakeholders sufficiently to build awareness and buy-in for new systems; (2) providing professional learning for educators and system leaders to enable them to implement and use new systems; (3) establishing sustainable and scalable scoring systems, particularly for those innovative systems that may rely more heavily on teacher/expert hand-scoring; (4) considering pilot data and making appropriate adjustments to assessment instruments, delivery mechanisms, and scoring approaches; and (5) implementing assessment maintenance and continuous improvement mechanisms that can sustain innovative systems over time.

By definition, innovation systems are new. They need space and time to ensure quality of the instruments and appropriate supports for users, as well as flexibility to course-correct during the scaling process. The limited timeline in IADA effectively requires that states have not only a predetermined plan for their innovative assessment, but also evidence and confidence in the functioning of the new system before entering the 5-year pilot rather than allowing for true innovation as part of the pilot. Acquiring this

evidence and confidence requires launching a system that states have no guarantee will be approved under federal law, which means that doing so before securing IADA approval (itself no certain guarantee) is not generally practicable.

### ***Possible actions***

- **Clarify and highlight the existing timeline and waiver opportunities.** Given this barrier, new regulations could clarify the timeline to scale statewide, providing states the 5- to 7-year time frame explicitly allowed under IADA, as well as the opportunity for additional waivers to extend that time frame pursuant to Section 1204 (j)(3) so that states could have additional time to implement new systems of assessments.
- **Consider a planning period.** Another approach that would not require amending current regulations (but could be included in a refresh) is to provide states a year or more of planning before the start of the 5- to 7-year demonstration period. Such an approach could be paired with planning grants, as described below, in Recommendation 8.

## **Recommendation 7: Lift the cap on the number of states able to participate in IADA and allow for states to collaborate on assessment designs.**

### ***Current requirements***

ESSA specifies that the Department of Education may grant demonstration authority to seven states, including those participating in consortia, during the first 3 years of IADA.<sup>33</sup>

The current federal administration is permitted to remove the seven-state limit on IADA after the third year of the program and the completion of the Institute of Education Sciences (IES) report on initial progress of innovative assessment systems required pursuant to Section 1204(c).<sup>34</sup>

### ***Why it matters***

Given the growing interest in new approaches to assessment, combined with the urgency of pandemic recovery, IADA could become an important vehicle for innovation if it is made less onerous to leverage and more responsive to state desires to innovate, as described in the recommendations above. Unfortunately, the program is currently capped at the seven participating states initially permitted in the statute, prior to the IES report on system progress.

### ***Possible action***

- **Complete the IES report and eliminate the seven-state limit.** The Department of Education can prioritize the completion of the IES report and take steps to eliminate the seven-state limit. This elimination would give more states the



ability to plan and engage in improved assessment practices. Lessons from these pioneering states could be invaluable as the nation crafts a new vision for assessment. The Secretary of Education could lift the cap on the number of states to allow other states that might be interested in exploring innovative systems to begin to do so. Pursuant to Section 1204(c)–(d), the Department of Education has the authority to take these actions now that the initial 3-year demonstration period has passed and once the IES report is complete.

## Create Additional Pathways to Innovation

While IADA represents one major effort to create opportunities for assessment innovation, there are other ways the Department of Education can signal, incentivize, and support change. The recommendation below focuses on alternative opportunities the government can engage to catalyze meaningful assessment reform.

### **Recommendation 8: Use the Competitive Grants for State Assessments (CGSA) program to stimulate individual or multistate efforts to develop and pilot new approaches that are instructionally useful and responsive to the broader view of assessment in ESSA.**

#### *Current requirements*

Title I Part B of ESSA, Grants for State Assessments (GSA), provides funding for states to support the costs of the development of state assessments and standards, and continual improvement of assessments.<sup>35</sup> If GSA is funded above \$369.1 million, any excess funds can be used to support CGSA, which is designed to enhance the quality of assessments or assessment instruments by states or a consortium of states.<sup>36</sup>

CGSA has been used to support improvements in assessments, including through multistate collaboratives, such as those awarded in 2019 to a 10-state collaborative to improve assessments for English language learners with significant cognitive disabilities, and in 2020 to an 8-state collaborative for instructionally relevant science assessments. It can also be leveraged to support states or multistate collaboratives in engaging in assessment innovation that improves state summative tests. Innovation through CGSA for this purpose would be more productive than IADA in its current form, as innovation would be tied to the provisions that govern state summative assessment in ESSA Section 1111(b)(2) rather than IADA's pilot authority, whose chief flaw is its conception of comparability that is not consistent with Section 1111(b)(2).

The Department of Education's 2022 CGSA grant announcement<sup>37</sup> and subsequent awards focus the program in a productive manner to support innovative assessments by requiring grantees to use funds for activities under Section 1201(a)(2)(K) and/or (L):

- (K) Measuring student academic achievement using multiple measures of student academic achievement from multiple sources.

- (L) Evaluating student academic achievement through the development of comprehensive academic assessment instruments (such as performance and technology-based academic assessments, computer adaptive assessments, projects, or extended-performance task assessments) that emphasize the mastery of standards and aligned competencies in a competency-based education model.

### ***Why it matters***

Assessment innovation is a resource- and time-intensive endeavor. The Obama administration recognized this fact when it provided over \$350 million to two state consortia, the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced, to develop new summative assessments aligned to state standards.<sup>38</sup> At the time, the two consortia involved 44 states. Funding and authority were tied to the American Recovery and Reinvestment Act's (ARRA) State Incentive Grant Program (Sections 14005 and 14006),<sup>39</sup> where among other uses, funding had to be used to enhance the quality of the Elementary and Secondary Education Act's required academic assessments.

Unlike with ARRA's State Incentive Grant Program, funding for states to innovate under CGSA has been anemic. For the current fiscal year (FY 2023), Congress appropriated \$390 million for GSA, leaving \$20.9 million for CGSA. Further, from FY 2018 to FY 2021, Congress appropriated only \$378 million to GSA, leaving only \$8.9 million annually for CGSA.<sup>40</sup>

In February 2022, when the Department of Education announced a CGSA competition, it combined \$8.9 million from FY 2021 and an expected \$8.9 million to make \$17.71 million available.<sup>41</sup> The Department of Education estimated an average grant award of \$2.5 million for a period of no more than 4 years.<sup>42</sup> At this average grant size and funding level, seven states could stretch \$2.5 million over 4 years, or about \$625,000 a year.

This funding level will not satisfy innovation needs. Based on our conversations with state leaders, it is likely that to fund deeply innovative approaches to assessment, an investment of \$2 million to \$7 million per state *annually* may be necessary. States need resources for different purposes that have different costs along the innovation continuum from planning through implementation. Likewise, states also need resources that align with timelines for innovation. Increased funding could fund more states, allow states to receive funding that supports their context, and sustain states participating in innovation over the timeline necessary to scale.

### ***Possible actions***

- **Increase funding for GSA to facilitate assessment innovation.** To remove cost as a barrier to assessment innovation, the Department of Education could consider a larger request. For example, increasing funding for CGSA to roughly

\$230 million could allow for 15 states to receive substantially increased grants of \$3 million annually for 5 years. Higher funding would allow for CGSA to be used for multiple purposes that support assessment and innovation and could provide additional resources to more states over a longer time period. In addition to supporting innovative assessments generally, additional funding under GSA could be used to support innovation through IADA. To remove cost as a barrier to engaging in IADA assessment innovation, the Department of Education could request \$600 million for GSA in future budget requests, which would leave roughly \$230 million to support CGSA, including funds that could be allocated for IADA purposes.

- **Utilize CGSA to support assessment innovation.** Prioritize support for states to engage in innovation. In addition to using the absolute priorities in the CGSA competition, the Department of Education could use CGSA funds for different discrete purposes that support assessment innovation. These purposes include (1) planning and implementation, (2) capacity building and assessment literacy, and (3) new technological methods.

Small planning grants could be provided to a larger number of states to support a rigorous design phase of the work, with larger implementation grants provided to a smaller number of promising models. Initial planning phases could be focused on activities that would position states to have a clear plan for how their innovative designs will support high-quality teaching and learning. Implementation grants could include support for capacity-building efforts such as the allocation of both sufficient time and funding for educator training for design, administration, scoring, and use of assessment results, as well as ongoing improvements in curriculum and instruction informed by assessment results. Planning and implementation grants could be used in many useful ways, such as:

- a. *Support for capacity building and assessment literacy.* For more innovative models, addressing the capacity building and assessment literacy needs will be imperative to successful system implementation. Additional funding could be specifically allocated to high-quality and high-leverage professional learning, capacity building, and assessment literacy connected to instructionally relevant assessments to incentivize these efforts.
- b. *Investment in new technical methodologies.* The prevailing traditional technical tools and models for assessment are sizable barriers to the creation of meaningful innovative designs. Many of the challenges described in this report could be addressed through technical innovation (e.g., psychometric models that allow for more expansive definitions of comparability, more innovative technology-enhanced items and simulation-based performance tasks, more effective automated scoring options)—but this kind of technical innovation requires considerable investments and is out of scope for states to pursue on their own.<sup>43</sup> By dedicating some specific and substantial

funding to innovative technical approaches, the Department of Education could catalyze the development of much-needed technologies that would accelerate assessment innovation, while providing a clear signal to states about its commitment to innovative assessment practices.

- **Allocate a portion of CGSA funding for planning and implementation grants to engage in IADA innovation.** Small planning grants could allow a state or a number of states to support a rigorous design phase before applying to IADA. Large implementation grants can support entities with approved applications to engage in implementing and scaling their innovative assessments.

## Conclusion

The Every Student Succeeds Act (ESSA) provided significant new opportunities for states to develop more robust assessments. These assessments have the potential to better measure higher-order thinking skills necessary for success in college and career, allow for better integration with curriculum, and provide timely information to inform instruction. However, additional action is needed to realize ESSA's promise of stronger state assessment systems that support, rather than hinder, teaching and learning.

A variety of federal executive action strategies could be implemented in the short term to encourage more innovative state assessment systems that better support teaching and learning, particularly as states work to support learning recovery related to the COVID-19 pandemic. Some strategies can help to strengthen state assessment systems in all 50 states under Section 1111(b). Other strategies can help to foster innovative assessments in the subset of states participating in the Innovative Assessment Demonstration Authority. In particular, those strategies include revising the options available for assessing comparability of new assessments in relation to standards for high-quality assessments generally and adjusting time frames to allow for design and scaling of new assessments. Additional headway can be made through an expanded Competitive Grants for State Assessments program. A strategic approach could enable significant advances at this time, as the field is focused on dramatic improvements needed to support learning recovery, which require assessments more tightly tied to curriculum and instruction.

## Endnotes

1. KnowledgeWorks. (2021). *Measuring forward: Emerging trends in k-12 assessment innovation*. <https://knowledgeworks.org/resources/emerging-trends-k12-assessment-innovation/>
2. Stecher, B. (2014). Looking back: Performance assessment in an era of standards-based accountability. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 17–52). Jossey-Bass.
3. Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests*. RAND Corporation.
4. Every Student Succeeds Act, 20 U.S.C. §6311(b)(2)(B)(vi) (2015).
5. Every Student Succeeds Act, 20 U.S.C. §6311(b)(2)(B)(vi) (2015).
6. Darling-Hammond, L., & Adamson, F. (2014). *Beyond the bubble test: How performance assessments support 21st century learning*. Jossey-Bass; Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). *The promise of performance assessments: Innovations in high school learning and college admission*. Learning Policy Institute; Newmann, F. M., Marks, H. M., & Gamoran, A. (1995). Authentic pedagogy: Standards that boost performance. *American Journal of Education*, 104(4), 280–312.
7. Elementary and Secondary Education Act of 1965 (P.L. 89-10, as amended through P.L. 116-260) Section 1204 (a)(1)–(2). <https://www.govinfo.gov/content/pkg/COMPS-748/pdf/COMPS-748.pdf>
8. For examples, see: Darling-Hammond, L. (2017). *Developing and measuring higher order skills: Models for state performance assessment systems*. Council for Chief State School Officers & Learning Policy Institute.
9. For a synthesis of research on how performance assessments can be designed to provide comparable data that can be aggregated for policy decisions and disaggregated to reflect the performance of student groups, see Lane, S. (2014). Performance assessment: The state of the art. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 133–184). Jossey-Bass.
10. For more information, see [www.summitlearning.org](http://www.summitlearning.org) and <https://www.assessmentforlearningproject.org/summit-public-schools/>
11. Fine, M., & Pryiomka, K. (2020). *Assessing college readiness through authentic student work: How the City University of New York and the New York Performance Standards Consortium are collaborating toward equity*. Learning Policy Institute.
12. Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). *The promise of performance assessments: Innovations in high school learning and college admission*. Learning Policy Institute.
13. Topol, B., Olson, J., Roeber, E., Darling-Hammond, L., & Adamson, F. (2014). Investing in assessments of deeper learning: The costs and benefits of tests that help students learn. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 259–276). Jossey-Bass.
14. For example, see the Building Educator Assessment Literacy initiative. Arnold, J. (2016). *Making the most of performance tasks in summative assessment: Building educator assessment literacy in Oregon*. WestEd.
15. ESSA requires the U.S. Department of Education to “conduct a peer review of the technical quality of all statewide assessment systems” to demonstrate that required statewide math, English language arts, and science assessments meet all of ESSA’s requirements (ESSA §1111(a)(4); 1111(b)(2)(B)(iii)-(iv); 34 CFR § 200.2(b)(4) and (5) and (d)).
16. U.S. Department of Education Office of Elementary and Secondary Education. (2018). *A state’s guide to the U.S. Department of Education’s assessment peer review process*. <https://www2.ed.gov/admins/lead/account/saa/assessmentpeerreview.pdf>
17. Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment* [White paper]. Center for K–12 Assessment & Performance Management, Educational Testing Service.

18. Darling-Hammond, L. (with Wentworth, L.). (2014). Reaching out: International benchmarks for performance assessment. In L. Darling-Hammond & F. Adamson (Eds.), *Beyond the bubble test: How performance assessments support 21st century learning* (pp. 93–132). Jossey-Bass.
19. ESSA §1111(b)(2)(J) (2015).
20. ESSA §1111(b)(2)(B)(J) (2015); 34 CFR § 200.2(c) (2016).
21. New Classrooms. (2019). *The iceberg problem: How assessment and accountability policies cause learning gaps in math to persist below the surface . . . and what to do about it.* <https://newclassrooms.org/icebergproblem/>
22. It's important to note that focusing on subscores undermines state assessment programs in several ways. For example, states are encouraged to build lengthier, more superficial assessments with enough points across the breadth of standards to support subscores while simultaneously trading off the very kinds of performances (e.g., those that focus on depth) that would actually support better classroom practice. Subscores are also rarely fine-grained enough or based on sufficiently compelling evidence from student performance for making meaningful decisions about student-level interventions. While there may be some limited potential uses of subscore information for making programmatic decisions at the school or district level, the requirement that state assessment programs report subscores at the individual student level is not necessary and interferes with the use of a number of innovative designs and measurement models that could better serve the intended uses of state assessments. Few, if any, states use subscores in their school accountability systems for identifying schools in need of support.
23. For a complete description of the process, please see the Department of Education's guide to the assessment peer review process. <https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/standards-and-assessments/>
24. ESSA §1111(b)(2)(H)(v)(II) (2015).
25. ESSA §1204(e)(2)(B)(v) (2015).
26. ESSA §1204(e)(2)(A)(iv) (2015).
27. 34 CFR §200.105(b)(4) (2016).
28. It's important to note that IADA requires that an innovative assessment “express student results or student competencies in terms consistent with the State’s aligned academic achievement standards” where “aligned academic achievement standards” is often interpreted as the achievement levels (e.g., emerging, proficient, advanced) and the associated descriptors that states develop as part of their assessment development process. States have shared that this requirement on its own acts as a guardrail to ensure that new assessments are still meaningfully connected to common expectations for meeting state standards but does not pose a significant barrier to innovation until it is coupled with the expectation for results that are comparable to the traditional test (e.g., that students who would be identified at a given proficiency level or score point on the old test would also be identified at the same level on the new test).
29. ESSA §1204(b)(2) (2015).
30. ESSA §1204(g) (2015).
31. ESSA §1204 (j)(3) (2015).
32. 34 CFR §200.108(c)(2) (2016).
33. ESSA §1204(b)(3) (2015); 34 CFR § 200.104(d) (2016).
34. ESSA §1204(d) (2015).
35. ESSA §1201 (2015).
36. ESSA §1203(b)(1)(A) (2015); ESSA §1201(a)(2) (2015).
37. Federal Register. (2022, February 16). *Applications for new awards; Competitive Grants for State Assessments Program.* <https://www.federalregister.gov/documents/2022/02/16/2022-03290/applications-for-new-awards-competitive-grants-for-state-assessments-program>
38. Heintz, L. (2013). *Next generation assessments.* National Conference of State Legislatures.

39. Federal Register. (2010, April 9). *Race to the Top Fund Assessment Program; notice inviting applications for new awards for fiscal year (FY) 2010*; The American Recovery and Reinvestment Act (Public Law 111-5) <https://www.congress.gov/111/plaws/publ5/PLAW-111publ5.pdf>
40. U.S. Department of Education. (n.d.). *2022 School Improvement Programs: Fiscal year 2022 budget request*. <https://www2.ed.gov/about/overview/budget/budget22/justifications/c-sip.pdf> (accessed 03/17/22).
41. CGSA's notice inviting applications was issued before the FY 22 bill was passed. The Department of Education projected a lower funding level for CGSA (\$8.9 million) than the \$20.9 million appropriated. It is assumed that the \$12 million worth of funding over the Department of Education's project will be committed to the notice as issued.
42. Federal Register. (2022, February 16). *Applications for new awards; Competitive Grants for State Assessments Program*. [https://www.govinfo.gov/content/pkg/FR-2022-02-16/pdf/2022-03290.pdf?utm\\_source=federalregister.gov&utm\\_medium=email&utm\\_campaign=subscription+mailing+list](https://www.govinfo.gov/content/pkg/FR-2022-02-16/pdf/2022-03290.pdf?utm_source=federalregister.gov&utm_medium=email&utm_campaign=subscription+mailing+list)
43. Pellegrino, J. W. (2023). Conclusions and implications. In M. Piacentini & N. Foster (Eds.), *Innovative assessment*. OECD.



## About the Authors

**Aneesha Badrinarayan** is the Director of State Performance Assessment Initiatives at the Learning Policy Institute (LPI), where she supports states in developing meaningful assessment systems that are a force for better teaching and learning. Badrinarayan is a national expert in innovative assessment systems, with a focus on meaningful science assessments. Her portfolio includes authoring more than 100 resources designed to support educators and leaders in the development of better assessment systems; leading collaborative national efforts to redefine “alignment” in the era of new state standards; partnering with states to develop instructionally relevant assessment systems; and leading multiple multistate collaboratives focused on better assessments in service of student learning. Prior to LPI, Badrinarayan served as the Director of Special Initiatives at Achieve, a museum educator, and a bench scientist. She earned an MS in Neuroscience at the University of Michigan, where she served as a research fellow for the National Institute of Mental Health, and a BA in Biology from Cornell University.

**Linda Darling-Hammond** is President of LPI and is the Charles E. Ducommun Professor of Education Emeritus at Stanford University. Darling-Hammond is past president of the American Educational Research Association and recipient of its awards for Distinguished Contributions to Research, Lifetime Achievement, and Research-to-Policy. She is also a member of the American Association of Arts and Sciences and of the National Academy of Education. She has been involved in the design and implementation of innovative assessments as Chair of the New York State Curriculum and Assessment Council, a founder of and technical advisor to the Smarter Balanced Assessment Consortium, and one of the initial designers of the Performance Assessment for California Teachers. Among her more than 600 publications are a number of studies of performance-based assessment, including *Authentic Assessment in Action* and *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning*. She received an EdD from Temple University (with highest distinction) and a BA from Yale University (magna cum laude).



1530 Page Mill Road, Suite 250  
Palo Alto, CA 94304  
p: 650.332.9797

1100 17th Street, NW, Suite 200  
Washington, DC 20036  
p: 202.830.0079

[@LPI\\_Learning](#) | [learningpolicyinstitute.org](http://learningpolicyinstitute.org)

The Learning Policy Institute conducts and communicates independent, high-quality research to improve education policy and practice. Working with policymakers, researchers, educators, community groups, and others, the Institute seeks to advance evidence-based policies that support empowering and equitable learning for each and every child. Nonprofit and nonpartisan, the Institute connects policymakers and stakeholders at the local, state, and federal levels with the evidence, ideas, and actions needed to strengthen the education system from preschool through college and career readiness.