COMMONWEALTH of VIRGINIA

Jenna Conway
Superintendent of Public Instruction

DEPARTMENT OF EDUCATION
P.O. BOX 2120
RICHMOND, VA 23218-2120

Office: (804) 225-2057
Fax: (804) 371-2099

June 3, 2026

The Honorable L. Louise Lucas
Chair, Senate Finance and Appropriations Committee 201 North 9th Street, Room 1404
Richmond, VA 23219

The Honorable Barbara A. Favola
Chair, Senate Education and Health Committee 201 North 9th Street, Room 509
Richmond, VA 23219

The Honorable Luke E. Torian
Chair, House Appropriations Committee
201 North 9th Street, Room 123 Richmond, VA 23219

The Honorable Sam Rasoul
Chair, House Education Committee 201 North 9th Street, Room 910 Richmond, VA 23219

Dear Chairs Lucas, Torian, Favola, and Rasoul:

Pursuant to Item 119.A.3.d. (Chapter 725, 2025 Acts of Assembly), please find the Virginia Department of Education's Report on the Virginia Assessment System.

Item 119.A.3.d., requires submission of a report that details options for building and implementing a high-quality assessment system to support high-quality teaching and learning in public K-12 schools in Virginia and associated costs.

If you have questions or require additional information relating to this report, please contact me.

Sincerely,

Jenna Conway
Superintendent of Public Instruction

CC: The Honorable Dr. Jeffery O. Smith
Secretary of Education

Virginia SOL Assessment System Modernization Study: Final Report Pursuant to House Bill 1600

Quality Information Partners, Inc.
www.qi-partners.com

TABLE OF CONTENTS

I. Executive Summary **1**

II. Overview **6**

A. Virginia's Current Assessment System *6*
B. Legislative Requirements: Chronological Development *10*
C. Project Scope *17*
D. Methodology *18*
E. Stakeholder Engagement *19*

III. Key Findings from Comprehensive State Scan, Organized by Integrated Priority Cluster **21**

Alignment, Cognitive Rigor, & Transparency (a, f, g) *23*
Non-traditional & Growth Measures (b, e) *24*
Accessibility & Multilingual Supports (c) *26*
Technology Integration and Scoring & Reporting (d, h, i, j) *26*
Implications for Virginia *28*

IV. Work Group and Stakeholder Engagement **30**

A. Work Group *30*
B. Additional Stakeholder Engagement *32*

V. System Design Options and Recommendations **34**

A. Options Framework *34*
B. Option Descriptions *36*
C. Comparison Table *44*
D. Recommended Approach *46*

VI. Implementation Considerations **49**

A. Transition Timeline and KPIs *49*
B. Funding Requirements *54*
C. Professional Development and Field Training Plan *57*
D. Communication and Change Management Strategy *58*
E. Legislative and Policy Considerations *62*
F. Risk Mitigation *64*

VII. Emerging Technologies and Future Trends **66**

A. Current Innovations *66*
B. Near-Term Opportunities for Virginia *70*

VIII. Procurement Guidance Overview **72**

A. Procurement Strategy Highlights *72*
B. Key RFP Components *72*
C. Next Steps for VDOE *77*

Appendix A. Virginia Innovative Assessment Work Group Members **78**

Appendix B. Exemplar State Profile by Key Focus Area 80

Appendix C. Federal Assessment Requirements Under ESSA 99

Appendix D. What We Heard: Stakeholder Feedback on Virginia Innovative Assessment 103

Appendix E. References 120

I. Executive Summary

Pursuant to Item 119 of the 2025 Appropriations Act (HB 1600), the Virginia Department of Education (VDOE) contracted with WestEd and Quality Information Partners (QIP) to evaluate exemplar assessment systems across 10 focus areas and develop actionable options for modernizing Virginia's Standards of Learning (SOL) assessment system. This report presents findings from that analysis, along with system design options, implementation guidance, and procurement recommendations for consideration.

Context

Virginia's assessment modernization effort reflects a multiyear legislative trajectory that began with HB 585 in 2022 and has been shaped by subsequent statutes, most significantly HB 1957 (enacted May 2025) and Senate Bill (SB) 200, along with its House companion HB 299, signed into law in April 2026. Together, these statutes establish additional requirements beyond federal compliance under the Every Student Succeeds Act (ESSA): assessments must evaluate application of knowledge and higher-order skills, individual student reports (ISRs) must be delivered within 45 days of the close of each assessment window, results must be reported on a 100-point scale, and local alternative assessments (LAAs) must meet Board of Education (Board) established guidelines for quality and consistency. The state is entering a new procurement cycle that will determine the design and delivery of the next generation of SOL assessments. These statutes include specific operational provisions that VDOE should implement alongside the broader modernization goals.

Evidence Base

WestEd conducted a systematic scan across more than 15 state assessment systems and multistate consortia, examining 318 pieces of evidence across the 10 focus areas specified in HB 1600. These focus areas span alignment with standards and instructional goals; non-traditional, project-, performance-, and competency-based approaches; accessibility and multilingual supports; advanced technology integration; authentic growth measures and formative feedback; application of knowledge and higher-order reasoning; transparency; innovative scoring practices; reporting to students, families, and educators; and timely reporting to the state educational agency.

The Virginia Innovative Assessment Work Group, established by state statute, met six times between October 2025 and March 2026. WestEd facilitated the final three sessions, which focused on reviewing comparative findings, developing draft recommendations, and validating the implementation framework. In addition, QIP conducted 13 stakeholder engagement sessions with 65 participants across all Virginia regions, including state and division leaders, school administrators, classroom educators, parents, and community-based organizations.

Findings

State scan. The 10 focus areas function as 4 integrated priority clusters rather than as isolated dimensions: (1) alignment, cognitive rigor, and transparency; (2) non-traditional and growth measures; (3) accessibility and multilingual supports; and (4) technology integration, scoring, and reporting. States with the strongest systems, notably Massachusetts, the Smarter Balanced consortium, and Florida, treat assessment modernization as a coherent system design problem in

which item types, scoring models, reporting timelines, and transparency practices are interdependent.

Several cross-cutting findings emerged from the scan:

- Assessment modernization should be treated as a coherent and integrated system redesign problem.
- Assessment expectations should be simultaneously technical and public-facing, so that educators can connect what is tested to what they teach.
- Innovation in assessment format is viable only when matched by educator capacity, infrastructure, and psychometric comparability across formats.
- Technology decisions about adaptive engines, automated scoring, and reporting are consequential beyond test delivery and should be evaluated together.
- Transparency and accessibility are foundational to equity and public trust.
- Successful modernization requires phased implementation, stakeholder engagement, and validity research at each stage.

Work Group recommendations. The work group reached substantial agreement on the following six priorities: (i) prioritize the summative assessment system; (ii) focus on alignment and rigor across the full depth and breadth of Virginia's standards; (iii) incorporate accessibility from the outset using universal design for learning (UDL) principles; (iv) modernize data systems to reduce reporting burden and support timely, accurate scoring and reporting to stakeholders; (v) provide consistent and simple reporting on student outcomes, including standardizing to four performance levels across all assessments; and (vi) increase public trust through transparency and consistency, including release of sample items, rubrics, and scored student work. Smaller subgroups within the work group highlighted additional priorities for VDOE to consider.

Stakeholder feedback. Educators and families reported communication gaps, including misconceptions about test formats and confusion about scoring scales. Teachers identified a lack of predictable test design documentation as a barrier to aligning instruction with state expectations. Parents described the current reporting system as inaccessible and poorly timed, with results arriving too late to inform summer interventions or the following year's instructional planning. Division leaders called for structured, multiyear assessment cycles and expressed concern about the pace and frequency of changes to the assessment system. While the transition to digital assessment formats was perceived as having significant advantages, all stakeholders reported concerns about the "digital divide" among students, as device navigation skills can affect performance. Finally, stakeholders emphasized that there should be a balance between innovation and academic rigor, and any shift toward performance-based assessments should address their concerns about administrative burden, local scoring validity, and the potential impact of artificial intelligence (AI).

System Design

The report presents three tiered options that satisfy all federal requirements under ESSA and all enacted state requirements. They differ in scope of innovation, implementation timeline, and level of investment and risk.

Core: Foundational Modernization. This option retains the current summative model with computer-adaptive delivery in Grades 3 through 8 mathematics and reading. Improves item quality (particularly at the Proficient and Advanced performance levels), implements the 0–100 reporting scale required by statute, standardizes four performance levels across all assessments, deploys UDL-based accessibility decision tools, expands released items and scored student work samples, and addresses LAA requirements under HB 1957 and SB 200. This option does not require structural changes to the current assessment model. Overall, this option is highly feasible. The estimated cost is $30 to $38 million per year for a scope similar to the current contract, based on comparable state contracts and nonbinding vendor estimates from the 2024 RFI. Virginia's current contract costs approximately $46.5 million per year. Several cost drivers, including end-of-course (EOC) retesting volume, item development intensity, and whether the Virginia Alternate Assessment Program (VAAP) is bundled or maintained separately, could push actual costs above the estimated range. These estimates should not be read as projected savings without careful analysis of the final contract scope.

Enhanced: Innovation Core

This option adds multistage adaptive testing (MST) for EOC subjects, AI-assisted scoring for constructed-response items with required fairness monitoring across all demographic subgroups, and an individual student growth model. Each feature requires pilot testing and validation before operational deployment. While this option offers solutions that go beyond the Core tier, it is still moderately feasible through phased implementation over a multiyear period and does not require federal approval. Estimated incremental costs for this option are $5 to $8 million per year above the Core tier for item bank expansion, AI scoring validation, and growth model development.

Transformative: Structural Redesign. This option incorporates competency-based assessment pathways and expanded local assessment flexibility with statewide moderation. However, the transformative nature of this option requires federal approval (through the Innovative Assessment Demonstration Authority or Section 8401 waiver), multiyear pilot programs, and infrastructure that Virginia does not currently have in place, including statewide scorer training, cross-division scoring coordination, and moderation protocols. The timeline is 3–5 years or more. The work group and broader stakeholders consistently prioritized summative system quality over structural redesign, and states that have pursued similar approaches have encountered significant challenges in scaling beyond pilot districts. Costs would be substantially above the Enhanced tier and would be incurred on top of ongoing Core tier operating costs.

Recommended Approach

The recommended approach is for Virginia to adopt the Core tier as the floor of the next procurement and structure the contract to build toward the Enhanced tier over the contract period. The Core tier addresses statutory requirements and sets the technical foundation that Enhanced features depend on, including the item bank depth. The Enhanced tier addresses documented needs, including individual student growth reporting, the 45-day ISR timeline for assessments with constructed-response items, and legislative direction on higher-order reasoning, through features that have been successful in peer states.

The Transformative tier is not recommended for the current procurement cycle. It requires not only substantial infrastructure and federal approval but also a policy determination that a fundamentally different approach to how student learning is defined and measured is the right direction for Virginia. The work group has not expressed that determination; its consistent priority has been a high-quality summative system.

The contract for the current procurement cycle should require the vendor to demonstrate capacity for Enhanced tier features and should include go/no-go decision points before those features move to operational use.

Implementation Timeline

The transition follows four phases and will overlap with the already-in-progress implementation of increased cut scores and a higher proficiency bar, scheduled to take effect through 2029–30. In the pre-procurement phase (2026), VDOE should finalize Board guidelines for mandatory LAAs, complete an ESSA compliance review, and issue the request for proposal (RFP). Year 1 (2027–28) will focus on delivering assessments without disruption under the new vendor. Additional Year 1 priorities, including the 0–100 reporting scale, the 45-day student score report requirement, accessibility tools, and the start of item bank development, may be achievable but will depend on the pace of vendor transition. Years 2 and 3 (2028–30) shift to Enhanced tier pilot testing, including MST for at least one EOC subject, and AI-assisted scoring validation. By Year 4 (2030–31), Enhanced tier features that meet pre-specified criteria for technical quality, fairness, cost, and educator readiness may be considered for operational use.

Near-Term Decisions

Several decisions should be resolved before procurement can proceed, including: the EOC assessment configuration (number of assessments, retake policy, and testing calendar); the 0–100 reporting scale implementation approach and associated communication strategy (including support for appropriately interpreting this scale to all intended stakeholders); Board guidelines and best practices for LAAs (due September 1, 2026, under SB 200); and timely RFP issuance to maintain the transition timeline.

Risks

The stakeholder engagement efforts revealed several misunderstandings about the assessment system among stakeholders. Lack of assessment literacy can undermine a system designed to provide useful data to teachers, families, and students. For example, the new 0–100 reporting scale will predictably be confused with percent correct or familiar grading scales. The report recommends a proactive communication strategy, including audience-tested messaging, an educator ambassador cohort, and materials available in the most commonly spoken community languages. Other risks include vendor transition disruption, item bank insufficiency at upper performance levels, AI scoring implementation delays, and VDOE internal capacity to manage the vendor relationship as well as a changing, complex assessment program.

VDOE Capacity

VDOE's ability to oversee an assessment program modernization depends on adequate internal staffing. Dedicated communications capacity would position VDOE to manage the public communication demands of the transition proactively rather than reactively. Additional content

staff would allow VDOE to keep pace with the volume of item review that an expanded development cycle requires, rather than creating a bottleneck at the state level. A data analyst would give VDOE independent capacity to evaluate vendor files, review score distributions and trends, and respond to technical questions from the field. These investments are modest relative to the contract and would strengthen both oversight and cost control.

Conclusion

Virginia has a clear legislative mandate, a defined set of priorities, and a competitive procurement opportunity to build an assessment system that is rigorous, transparent, and useful to educators and families. The exemplar evidence, work group input, and stakeholder feedback converge on a consistent message: get the summative system right first, build toward innovation through phased implementation, and invest in the communication and capacity infrastructure that the transition requires. The recommended approach—Core as the floor with a structured path to Enhanced—reflects that message and positions Virginia to deliver a modernized assessment system that meets statutory requirements, maintains public trust, and supports student learning.

II. Overview

The Virginia Standards of Learning (SOL) Assessment System Modernization Study provided information to the Virginia Department of Education (VDOE) to support the development of a modernized statewide assessment system aligned to the Commonwealth's SOLs and recent legislative direction. The work is designed to translate Virginia's policy goals into feasible design options by examining high-quality exemplar systems across key focus areas and identifying actionable pathways for implementation. This section establishes the foundation for that analysis: summarizing Virginia's current assessment system, then providing an overview of the legislative requirements and policy milestones that have shaped the modernization effort, and finally outlining the scope and methodology guiding this project. Together, this background provides the context necessary to evaluate exemplar approaches and frame recommendations for Virginia's next phase of assessment reform.

A. Virginia's Current Assessment System

Overview of Virginia's Current Summative Assessments

Virginia's current statewide assessment program (as of the 2024–25 school year) is designed to meet state and federal (Every Student Succeeds Act [ESSA]) requirements by measuring student performance on the SOLs in reading, writing, mathematics, science, and history/social science across Grades 3–8 and high school. The assessment is administered across multiple test administration windows throughout the year. Based on information from the Virginia SOL assessments 2024–25 technical report (Virginia Department of Education, n.d.), Table 1 presents an overview of the current SOL summative assessments for Grades 3–8, and Table 2 presents an overview of the current high school End-of-Course (EOC) assessments.

As shown in Table 1, in Grades 3–8, Virginia SOL summative assessments are administered in reading and mathematics across all grades, and summative science assessments and Integrated Reading and Writing (IRW) assessments are administered in Grades 5 and 8; in addition, content-specific history (Virginia studies) and civics and economics summative assessments are administered in upper elementary and upper middle grades, respectively. As shown in Table 2, EOC assessments are administered in high school in each of the following subjects:

- English: Reading, Writing, and an optional IRW test;
- Mathematics: Algebra I, Geometry, and Algebra II;
- Science: Biology, Chemistry, and Earth Science; and
- History/Social Science: Virginia & U.S. History, World History I, World History II, and World Geography.

Table 1. Virginia's SOL Summative Assessments in Grades 3–8

Subject	Grade Administered	Administration Mode*	Reporting Scale and Performance Level			
			Fail		Pass	
			Below Basic	Basic	Proficient	Advanced
Reading	3, 4, 5, 6, 7, 8	CAT	Varies by grade**	399 and below	400–499	500–600
Mathematics	3, 4, 5, 6, 7, 8	CAT	Varies by grade**	399 and below	400–499	500–600
			Fail		Proficient	Advanced
Science	5, 8	Fixed-form	Scale score of 399 and below		400–499	500–600
Integrated Reading & Writing (IRW)	5, 8	Fixed-form	Scale score of 399 and below		400–499	500–600
Virginia Studies (History)	Upper elementary	Fixed-form	Scale score of 399 and below		400–499	500–600
Civics & Economics	Upper middle	Fixed-form	Scale score of 399 and below		400–499	500–600

***Note:** Accommodated paper forms are available for students who need them.
****Note:** For Grades 3–8 Reading and Math, there are two failing categories; cut scores that classify students into the Below Basic category (i.e., students who are not yet Basic) for Reading and Math vary by grade; for Reading, these cut scores are 334, 338, 325, 317, 330, and 323 for Grades 3–8 respectively; for Mathematics, these cut scores are 342, 345, 334, 338, 340, and 339 for Grades 3–8 respectively.

Table 2. Virginia's SOL EOC Summative Assessments (High School)

Subject	Test	Administration Mode*	Reporting Scale and Performance Level		
			Fail	Pass	
				Proficient	Advanced
English	Reading; Writing; IRW (optional)	Fixed-form	399 and below	400–499	500–600
Mathematics	Algebra I, Geometry, Algebra II	Fixed-form	399 and below	400–499	500–600
Science	Biology, Chemistry, Earth Science	Fixed-form	399 and below	400–499	500–600
History/Social Science	Virginia & U.S. History, World History I, World History II, World Geography	Fixed-form	399 and below	400–499	500–600

***Note:** Accommodated paper forms are available for students who need them.

Administration Formats. The Grades 3–8 reading and mathematics SOL assessments are delivered as computer adaptive tests (CAT), while other SOL assessments (for both Grades 3–8 and EOC) are delivered as online linear fixed forms. For all of these assessments, paper-based tests are available for students who need them. The SOL summative assessments have defined assessment windows with retest SOL administrations available for eligible students, particularly for EOC tests. Virginia also administers fall and winter growth assessments in Grades 3–8 reading and mathematics, as described in the following section.

Growth Assessments. Virginia administers fall (baseline) and winter (mid-year) growth assessments in Grades 3–8 reading and mathematics in addition to the spring SOL assessments. These growth assessments are delivered as shorter CAT administrations, using existing SOL items to maintain alignment to SOL content standards; they are intended to measure student progress toward proficiency and inform instruction during the school year. The growth assessments primarily administer on-grade items, but they may also administer items below or above grade level, depending on student responses. Currently, these assessments are optional; school divisions may either use the state-provided growth assessments or request approval to use alternative assessments aligned to the SOL (e.g., NWEA MAP). These growth assessments also produce a vertical scale score to measure progress; however, these scales vary and are not on the SOL summative test scale. Performance levels are not reported for the fall and winter growth assessments. However, House Bill 199, signed into law on April 13, 2026, and effective July 1, 2026, eliminates the state mandate for fall and winter through-year growth assessments in Grades 3–8 reading and mathematics, returning Virginia exclusively to the end-of-year SOL testing framework.

Reporting Scale and Performance Levels for the Summative SOL Assessments. Virginia reports overall test scale scores for all SOL assessments on a 0–600 scale; 400 corresponds to the cut score that places students in the Pass/Proficient performance level, and 500 corresponds to the cut score that places students in the Pass/Advanced performance level. Scores below 400 are considered failing or not proficient. As shown in Tables 1 and 2, performance levels are structured differently for different tests: Grades 3–8 reading and mathematics are reported on four performance levels, while all other SOL assessments are reported on three performance levels, with the two lowest performance levels combined into a single level. In addition to overall scale scores and performance levels, SOL reports include reporting category scale scores ranging from 0 to 50, with a score of 30 indicating approximate grade-level mastery of content in that reporting category. These scores are intended for within-category interpretation, rather than for comparisons across categories.

Item Types. Virginia's SOL assessments consist primarily of multiple-choice (MC) items and technology-enhanced items (TEIs), with writing prompts included in the EOC Writing assessment and in the IRW component for Grade 5, Grade 8, and EOC. The TEIs are mostly in interactive formats that require students to interact with test content to select, construct, and/or support their responses; these formats are better suited to assessing deeper understanding than MC items alone. Several TEI types are already operational, including bar graphs, drag-and-drop items, and equation writing items, which allow students to demonstrate proficiency beyond MC responses. For paper-based assessments, the TEIs are replaced with another item type that assesses the same standards. All reading, mathematics, science, and history items are tagged with a cognitive complexity level. Virginia's SOL assessment system also includes writing prompts in the IRW and in the EOC Writing assessments that ensure that assessments adhere to SOL

specifications and elicit extended constructed responses. In addition, according to VDOE's November 2024 progress update on assessment modernization under House Bill 585 (HB 585; Chapter 760, 2022 Acts of Assembly) (Virginia Department of Education, 2024), several short constructed-response item types for both English language arts (ELA) and mathematics and other innovative (and TE) item types have been either field-tested or considered for future field testing.

Testing Accommodations. Virginia groups accommodations for students with disabilities into four standard categories:

- ***Timing/scheduling accommodations*** include multiple test sessions, time of day, order of tests, and flexible planned breaks.
- ***Setting accommodations*** include changes to test location, adaptive/special furniture, and special lighting.
- ***Presentation accommodations*** include oral directions accompanying written directions; specific verbal prompts; visual aids; amplification/noise dampening; large print, braille; read-aloud/audio test; and American Sign Language (ASL) interpreting/transliterating directions of the test itself.
- ***Response accommodations*** include examiner recording responses, word processor (including speech-to-text), augmentative communication devices, word prediction, spelling aids, dictionaries, dictation to a scribe, read-back of responses, and calculator/mathematics aids.

In addition, the technical report describes two categories of accommodations that are provided for students who are English learners:

- ***Direct linguistic accommodations*** include read-aloud/audio, bilingual dictionary, and translations or modifications to test directions.
- ***Indirect linguistic accommodations*** include adjustments to testing conditions such as multiple test sessions, visual aids, and flexible scheduling.

English Language Proficiency (ELP) Assessment. To meet federal requirements related to measuring ELP for English learners, Virginia uses the multistate WIDA ACCESS assessment as its ELP assessment. WIDA ACCESS measures proficiency in listening, reading, speaking, and writing, and it provides English learners with any additional necessary accommodations that are grounded in evidence. Virginia also administers Alternate ACCESS for English learners with the most significant cognitive disabilities.

Virginia Alternate Assessment Program (VAAP). Virginia offers an alternate assessment program for students with the most significant cognitive disabilities—approximately 1% of Virginia students—in reading, mathematics, and science for Grades 3–8 and high school. The VAAP is aligned to the Virginia Essential SOL, which are standards derived from the SOL but adjusted in depth, breadth, and complexity.

Virginia's Current Score Reporting System. Virginia provides a suite of score reports and data files to school divisions and schools after "authorization to report" (ATR) is granted following the administration of the summative SOL assessments. Virginia is currently able to provide test results to school divisions within 24 hours of test administration; this quick turnaround is feasible largely because the operational system relies heavily on selected-response formats. However, results may not be shared promptly downstream with school leaders, teachers, and families

because school divisions generate additional local reports or conduct internal analyses before distribution. Virginia's current suite of post-ATR score reports serves various audiences, including a two-page, family-facing individual student report (ISR) supplemented by a Student Detail by Question (SDBQ) report, a one-page, item-level diagnostic report that shows item descriptors, item difficulty, and student responses (correct or incorrect) for all items presented to the student. For spring 2026, VDOE worked with a national organization on parent communication to revise the ISR. The SDBQ is now included in the ISR, along with a QR code linking to additional resources for parents. Additional reports include a suite of school/division reports and dashboards that show overall and reporting-category results for various disaggregated groups of students, and various other summary dashboard reports that give administrators graphical overviews of group, school, or division performance on each test in a variety of formats.

Virginia's current academic assessment system is a large-scale, standards-based program with three primary components: SOL summative testing (computer-adaptive delivery in Grades 3–8 reading and mathematics, fixed-form online tests in other subjects); a statewide alternate assessment for students with significant cognitive disabilities; and (until July 1, 2026) fall (baseline) and winter (mid-year) growth assessments in Grades 3-8. Results are reported through a suite of student-, school-, and division-facing reports that include overall scale scores, performance-level classifications, and reporting-category performance.

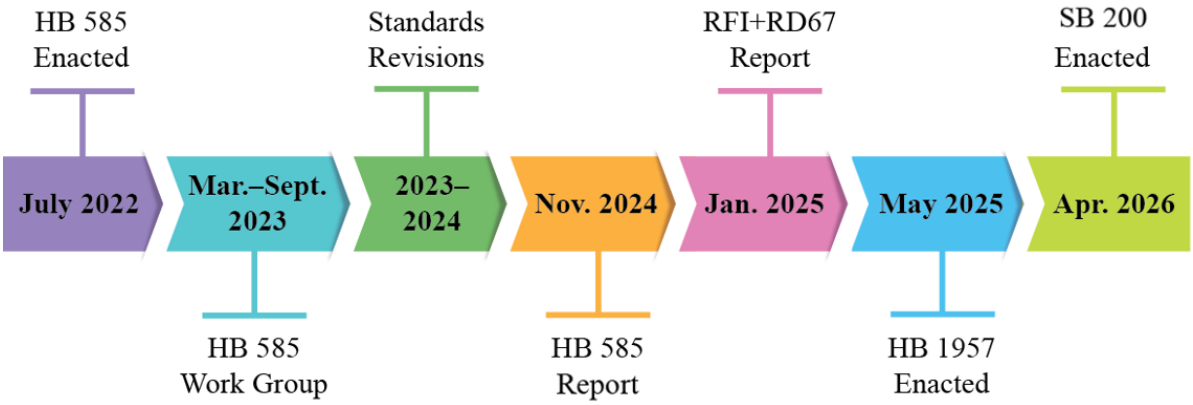
This brief overview of Virginia's current assessment system provides context for understanding both the strengths of Virginia's existing assessment infrastructure and the design and implementation considerations associated with any modernization effort. The following section reviews Virginia's recent legislative and policy history in detail to show how statutory requirements and statewide priorities have evolved and how they shape the goals and expectations reflected in Virginia's assessment modernization efforts.

B. Legislative Requirements: Chronological Development

This section describes how Virginia's laws and related state planning have progressively articulated new expectations, such as stronger alignment to updated standards, expanded item types to capture higher-order skills, improved accessibility, faster and more actionable reporting, and clearer rules for performance-based and alternative assessment pathways.

Beginning with the enactment of HB 585 (Chapter 760, 2022 Acts of Assembly), Figure 1 summarizes the sequence of major milestones in this policy evolution. This legislative timeline provides the policy context for the project's scope and clarifies why Virginia is seeking to evaluate exemplar state systems across the 10 focus areas described in this report.

Figure 1. Chronological Development of Legislative Requirements in Virginia's Assessment Modernization Timeline



HB 585 Enacted, July 2022

HB 585 took effect in July 2022, launching Virginia's multiyear effort to modernize its SOL assessment system. The law established an expectation for VDOE to plan for revisions to SOL summative assessments, including consideration of richer item types, pilots prior to full implementation, development of a vetted item bank, and identification of policy and funding needs to support the transition. The law created the statutory foundation for the state's subsequent activities: convening a work group, documenting early implementation actions, and testing feasibility and costs through a vendor information-gathering process.

HB 585 Work Group Convened and Modernization Blueprint Produced, March–September 2023

Beginning in March 2023, the Virginia Secretary of Education and VDOE convened the HB 585 Work Group, which included teachers, parents, Board of Education members, division and school leaders, and assessment experts, to review Virginia's assessment system, examine innovative assessment approaches, and compare Virginia to higher-performing states. The work group framed the need for assessment reform around improving the usefulness and transparency of assessment results for educators and families, and it situated assessment redesign within Virginia's Framework for Excellence in Education, aligning standards, assessments, and accountability.

The work group organized recommendations into five opportunity areas that became the state's modernization roadmap:

1. clearer and more rigorous standards
2. more rigorous assessment items (beyond selected-response)
3. more timely, clear, actionable reporting
4. improved system coherence (reducing duplicative assessments)
5. a runway for innovative assessment design over time

The work group's report also articulated a sequenced transition process for the assessment system that involved revising standards and then benchmarking expectations for what students know and can do, including via National Assessment of Educational Progress (NAEP) crosswalks, to strengthen credibility. The work group recommended that the state should then

pursue procurement to support new item type development before finally scaling up to statewide implementation.

Standards Revisions, April 2023–June 2024

From April 2023 to June 2024, VDOE updated its SOLs across major content areas: History/Social Science (April 2023), Mathematics (August 2023), and ELA (March 2024). These revisions served as a foundational step for modernization, establishing stable instructional expectations on which assessment specifications, blueprints, and item development could be rebuilt.

VDOE Assessment Modernization Progress Update, November 2024

In its November 2024 progress update, VDOE described progress on HB 585 Work Group recommendations that could move forward without new procurement or additional funding, and highlighted several concrete actions taken during 2023–24:

- ***Assessment alignment and item-type expansion.*** VDOE began reviewing test items and test structure and realigning them to the newly adopted mathematics and ELA SOL, and it provided school divisions with updated resources describing the revised structure and alignment. The update reported the development of a small set of short constructed-response items in reading and mathematics, with teacher review completed in summer 2024 and field testing planned for spring 2025. The report also referenced the planned development of additional TEI types (e.g., drag-and-drop, equation writing), also slated for spring 2025 field testing.
- ***Accessibility supports.*** VDOE documented continued and expanded accommodations and embedded accessibility features to support English learners and students with disabilities, including read-aloud, text-to-speech, screen reader/braille supports, ASL, speech-to-text, word prediction, translated glossaries, and additional interface features such as magnification and color contrast settings.
- ***Reporting improvements for families.*** VDOE partnered with Learning Heroes to examine the usability of parent reports and SDBQ reports, through parent focus groups and principal interviews, producing recommendations to improve clarity and navigation (e.g., simpler language, consistent labeling/color coding, added resources). The Office of Assessment then worked with the current assessment contractor to provide parent-facing report videos and parent-portal resources, beginning with fall 2024–25 reporting. The update noted that the expanded use of constructed-response and performance-task items in future assessments would require additional scoring capacity if VDOE plans to maintain timely reporting. The update referenced automated scoring as one strategy being considered to support timely score reports as item types evolve.
- ***System coherence through early literacy screening.*** As part of strengthening system coherence and reducing early-grade data gaps, Virginia implemented the Virginia Language and Literacy Screening System (VALLSS), developed by the University of Virginia, for prekindergarten–Grade 3 in 2024–25. The VALLSS measures alphabet knowledge, phonological awareness, decoding, encoding, passage comprehension, vocabulary, and rapid automatized naming. The VALLSS for eligible students in Grades

4–8 was also piloted in 2024–25, with implementation of the Grades 4–8 assessments beginning in 2025–26.

RFI to Test Vendors Issued, August–September 2024; RD67 Report Submitted to General Assembly, January 2025

Following the HB 585 Work Group recommendations and early implementation actions documented in the progress update, VDOE used a request for information (RFI) to test feasibility, vendor capacity, and indicative cost for a modernized statewide assessment program. The RFI, posted in late August 2024 and closed at the end of September 2024, was explicitly structured as an information-gathering step and not a vendor selection process. In the subsequent RD67 report (Virginia Department of Education, 2025), VDOE described how the RFI was aligned to the state's modernization priorities, with vendors asked to address major functional areas such as rigorous item development and item banking, assessment development and administration, scoring and reporting, performance tasks, and longer-term possibilities such as competency-based assessment approaches. Findings from the RD67 report inform the next phase of work: shaping a formal procurement pathway (an RFP) and laying out a transition strategy that would maintain statewide testing continuity while a new system is designed, field-tested, and implemented.

Item 119 of the 2025 Appropriations Act (HB 1600), 2025

Item 119 of the 2025 Appropriations Act (HB 1600) established this study and defined its scope, directing VDOE to evaluate exemplar assessment systems across 10 focus areas spanning assessment design, accessibility, technology, scoring, and reporting. With its emphasis on innovation, Item 119 signals that the General Assembly expects VDOE to consider how the next assessment system can be better than what compliance alone would produce.

HB 1957 Enacted, May 2, 2025 (Major Provisions Effective July 1, 2026; Some Provisions Contingent on Reenactment)

HB 1957 represents a significant statutory step that reinforces and operationalizes key themes from HB 585, particularly around what assessments should measure and how quickly results should be returned. This section broadly describes the requirements of the HB 1957 statute; ESSA requirements are presented in Appendix C.

In broad terms, HB 1957 strengthens the assessment statute to emphasize evaluation of knowledge as well as the application of knowledge and related higher-order skills. It also advances requirements that affect assessment operations and reporting, including timeliness expectations for returning ISRs to families after an assessment window closes. HB 1957 also provides additional structure and expectations around local alternative assessments (LAAs), including performance assessment and portfolio pathways, signaling that Virginia intends "authentic" assessment options to be defined and governed with clear parameters. HB 1957 includes an effective-date structure in which requirements are implemented on a staged basis. This structure has important implications for timeline planning. Many provisions are scheduled to take effect on July 1, 2026, and will be operational requirements after that date. However, two provisions—subsections K and L, which include making prior-year SOL tests available for educator use and auditing LAAs for consistency and rigor—were adopted as a "first enactment," contingent on reenactment by the 2027 General Assembly.

SB 200 Signed into Law, April 6, 2026

Senate Bill (SB) 200, along with its House companion HB 299, was signed into law on April 6, 2026. SB 200 makes several clarifying revisions to HB 1957, as described in the following section. Where SB 200 amends a specific HB 1957 provision, the amendment is noted in line below.

Virginia's planning and procurement timeline should account for three distinct categories of statutory requirements: provisions that take effect July 1, 2026; provisions whose effective date is tied to pilot implementation of the new assessment system; and provisions that require reenactment by the 2027 General Assembly.

Requirements Effective July 1, 2026 (HB 1957 as Amended by SB 200)

Beginning July 1, 2026, Virginia's assessment system must:

- ***Measure higher-order skills and validate assessment quality.*** Statewide assessment methods must evaluate knowledge, application of knowledge, critical thinking, and related skills, and the Board must conduct a regular analysis and validation process with independent experts (§ 22.1-253.13:3, subdivision C 1).
- ***Apply reporting requirements statewide.*** All assessments, including SOL assessments, mandatory LAAs, and permissive LAAs, must be reported on a 100-point scale in accordance with statewide guidance (subdivision C 10, as amended by SB 200; HB 1957 originally required assessments to be "scored" on a 100-point scale, and SB 200 changed this to "reported," preserving technical scoring flexibility while standardizing score communication). Virginia must provide ISRs within 45 days of the close of any state assessment window (subdivision C 11) and publicly release statewide SOL results no later than the date on which individual scores are released (subdivision C 12).
- ***Update and clarify LAA rules.*** The Board must amend guidelines to distinguish between mandatory and permissive LAAs and must establish criteria and parameters for their development and administration. The definition of "local alternative assessment" includes accepted college-readiness tests such as Advanced Placement (AP), International Baccalaureate (IB), or Cambridge assessments. School boards must administer mandatory LAAs in every SOL subject area where no statewide assessment is available during the school year. SB 200 added requirements that each mandatory and permissive LAA (1) permit the use of formative learning experiences designed to build student understanding toward the summative performance assessment, with limits on what materials may carry into the summative assessment (student's own work only, no drafts or outlines, no teacher-provided scaffolds), and (2) meet authenticity and integrity requirements, including completion in a supervised, secure setting with digital security measures such as a lockdown browser. SB 200 also requires that any teacher grading LAAs for verified credit purposes shall not grade assessments completed by students they taught during that school year. The Board must develop and make available best practices for grading and scoring both mandatory and permissive LAAs by September 1, 2026 (SB 200, enactment clause 4).
- ***Constrain assessment frequency.*** For each SOL subject area in a single school year, the total number of assessments (including SOL assessments, mandatory LAAs, and

permissive LAAs) shall not exceed one per academic quarter, or four per year (subdivision E 5). SB 200 added three exemptions not present in HB 1957: (i) AP, College-Level Examination Program (CLEP), IB, and other national norm-referenced achievement tests; (ii) assessment retakes; and (iii) performance assessments. The performance assessment exemption is significant for LAA implementation, as it means performance assessments administered as part of a mandatory LAA do not count toward the one-per-quarter limit.

- ***Constrain assessment timing for Grades 7 through 12.*** Any EOC SOL assessment or permissive LAA must be administered no earlier than 2 weeks prior to the last day of the applicable course of instruction, grade level, or academic quarter or semester (subdivision E 4 a). HB 1957 originally referenced "the last day of the school year"; SB 200 changed this to "the last day of the applicable course of instruction, grade level, or academic quarter or semester," providing greater scheduling flexibility for courses that do not end at the close of the school year. SB 200 also added exceptions for (i) graduating seniors during their final academic quarter or semester, (ii) assessment retakes, (iii) multiday assessments in Grades 7 and 8 (up to 3 weeks if needed after reasonable efforts to complete in 2 weeks), and (iv) transfer students or testing irregularities. These timing and frequency provisions do not apply to students with disabilities who participate in the VAAP or alternative SOL assessment administration (added by SB 200), to assessments administered as part of a competency-based assessment system, or to AP, CLEP, IB, and other national norm-referenced tests.
- ***Establish retake eligibility.*** Any student who performs below grade level on an SOL assessment or permissive LAA is eligible to retake the assessment, subject to the following conditions: no retake after June 30 of that calendar year unless the school board grants permission; students eligible for expedited retakes must retake within 2 weeks of scoring completion; and the retake score does not replace the initial score for purposes of calculating the student's final course grade (subdivision E 3 c). For Grades 3 through 8, recovery credits must be awarded to any student who performs below grade level on a reading or mathematics SOL assessment, receives remediation, and subsequently passes a retake (subdivision E 1 a (1)).
- ***Exempt VAAP students.*** SB 200 clarifies that students with disabilities who participate in the VAAP or in alternative SOL assessment administration are exempt from several requirements of subdivisions E 4 and E 5, including the timing, frequency, and grading provisions.

Requirements with Delayed Effective Dates

SB 200 delays two provisions of subdivision E 4. These provisions do not take effect until the beginning of the second full school year after the school year during which pilot implementation of assessment items for the new statewide assessment system takes place, pursuant to the HB 585 framework (SB 200, enactment clause 2). The Department of Education must certify in writing to the Virginia Code Commission the date on which this condition is met. The delayed provisions are as follows:

- ***No additional EOC assessment (subdivision E 4 b):*** No additional EOC assessment shall be administered for any SOL subject area beyond the required EOC SOL assessment or permissive LAA.
- ***Ten percent course grade requirement (subdivision E 4 c)***: Each student's score on a grade 7-12 EOC SOL assessment or permissive LAA shall account for at least 10% of the student's final grade in that course, as calculated based on the grading scale used in the applicable school division. This course grade requirement does not apply to students with disabilities who participate in the VAAP or alternative SOL assessment administration (added by SB 200), to assessments administered as part of a competency-based assessment system, or to AP, CLEP, IB, and other national norm-referenced tests.

Based on the implementation timeline described in this report, the pilot year would likely be no earlier than 2028–29, placing the earliest effective date for these provisions at the beginning of the 2030–31 school year.

Provisions Contingent on Reenactment by the 2027 General Assembly

Two subsections added by HB 1957, subsections K and L, do not take effect unless reenacted by the 2027 session of the General Assembly (SB 200, enactment clause 3). SB 200 also directs the Department, to the extent permitted by federal law, to include these provisions as "add alternate components" in the RFP for the new assessment vendor and to report their estimated costs to the General Assembly by October 1, 2026 (enactment clause 5).

- ***Subsection K*** requires the Board to (1) make available to each school division the previous year's SOL assessments and answer keys for use as practice tools and instructional resources, with alignment and refresh requirements, and a fallback to representative sample questions if full release would deplete the item bank below 70% coverage; and (2) perform annual audits of mandatory and permissive LAAs for quality and consistency, with consequences for audit failure including prohibition from administering permissive LAAs for 4 years.
- ***Subsection L*** requires all assessments in Grades 3 through 12 to include items requiring application of knowledge, critical thinking, and logic, including open-ended questions and long-form writing, and to use different items across assessment administrations within the same school year, grade, and subject. It also sets parameters for multiple-choice questions, including a 40-question thematic cap for mandatory LAA performance assessments.

Since 2022, Virginia has advanced a multiyear effort to modernize its statewide assessment system through a sequence of legislative actions and state planning steps that have progressively clarified both the rationale for change and the operational expectations for a redesigned program. Beginning with HB 585, Virginia launched a structured process to examine the limitations of the existing SOL assessment system and outline a pathway toward more rigorous, instructionally useful measures, including stronger alignment to updated standards, expanded item types, clearer reporting, and a longer term runway for innovation. Subsequent reporting and feasibility work further translated these priorities into practical considerations around test design, scoring, reporting timelines, accessibility, and statewide capacity. More recent statutory updates, including HB 1957 and SB 200, have reinforced these expectations that Virginia's modernized

assessments should measure application of knowledge and higher-order skills and that results from these assessments should be communicated in ways that support educators, families, and state-level decision-making. At the same time, the level of operational detail embedded in these statutes is unusual for assessment legislation and introduces implementation considerations that VDOE needs to manage alongside the broader modernization goals. Some provisions, such as specific constraints on assessment timing and frequency and requirements for particular scoring and reporting formats, may constrain VDOE's flexibility to sequence modernization work while meeting statutory requirements.

Collectively, this legislative and policy trajectory sets the context for the current project requiring a systematic review of exemplar assessment systems across 10 focus areas to identify models, evidence, and best practices that can inform feasible options for a modernized statewide assessment system.

C. Project Scope

This section describes the scope of the current SOL modernization project. WestEd partnered with VDOE and Quality Information Partners (QIP) to analyze options for implementing an innovative new statewide assessment system that supports high-quality teaching and learning, is aligned to Virginia's SOLs, and is designed to be implemented beginning in the 2027–28 school year. Consistent with the modernization rationale described in previous sections, the project's core work was a comparative analysis of how other states operationalize their assessment design choices and what those choices imply for Virginia's assessment system options in terms of feasibility, infrastructure, accessibility, scoring, and reporting needs. Pursuant to Item 119 of the Virginia Appropriations Act (HB 1600), Virginia explicitly asked that the core analysis of the exemplar evaluation approach center on 10 focus areas, across assessment design, accessibility, technology, scoring, and reporting, so that Virginia can evaluate these exemplar approaches and translate them into feasible options for a modernized statewide assessment system. The following section provides a brief overview of the 10 focus areas.

Summary of the 10 Focus Areas

a. ***Alignment with standards and instructional goals.*** Review how other states document alignment of assessments to standards (including claims/targets, blueprints, and item-to-standard mapping), how they verify alignment over time (including alignment studies and educator review processes), and how they communicate that alignment so teachers can use it to guide instruction.

b. ***Nontraditional, project-, performance-, and competency-based approaches.*** Review how other states and consortia have used project-, performance-, and competency-based assessments at scale, including the technical and policy conditions that make them feasible and the implementation pathways (e.g., pilots, phased rollouts, guardrails) that help states avoid equity and consistency problems.

c. ***Accessibility and multilingual supports.*** Review how states and consortia have used inclusive design and resource development practices that ensure appropriate supports and provide equitable access for all students, including multilingual learners.

d. ***Advanced technology integration and infrastructure.*** Review choices made by exemplar states that illustrate scalable, adaptive architecture and technology integration, such as

interactive items and accessibility features, comparing platforms and delivery models, and infrastructure requirements for rapid reporting.

e. ***Authentic growth measures that provide actionable, formative feedback.*** Review how other states use through-year assessments or other interim assessments with vertical scales, how they benchmark or interpret growth, and how they ensure that growth results are timely, understandable, and connected to instructional next steps.

f. ***Application of knowledge and higher-order reasoning skills.*** Review exemplar states to identify item/task types and rubrics that elicit higher-order reasoning, looking specifically at how states ensure that these tasks are aligned to standards and are reliable.

g. ***Transparency through release of rubrics, performance level descriptors, item maps, and sample items and tasks.*** Review what states release, how often, and how they protect item banks while staying transparent.

h. ***Innovative, reliable, and valid scoring practices.*** Review how other states use innovative scoring approaches (human, distributed, teacher-moderated, automated, or hybrid) and describe quality control (QC) processes (calibration, drift checks, audits, anomaly detection) that they use for clean, validated, and timely results.

i. ***Reporting to students, families, and educators.*** Review how states tailor reports by audience (e.g., family-friendly ISRs vs. educator dashboards), what level of detail is most useful for each audience, whether item-level reports have been used and where they are deemed appropriate, how quickly results are returned, and how reporting is linked to supports or resources that help users respond instructionally.

j. ***Timely reporting to state education agency (SEA).*** Review best practices across exemplar states to understand data pipelines (e.g., application programming interfaces [APIs], validation rules, refresh cycles), security and governance, and analytics-ready reporting structures and dashboards that help the SEA identify where school divisions need support while there is still time to intervene.

D. Methodology

WestEd conducted an analysis of exemplary practices from various states across the nation to identify high-quality models that address the 10 focus areas (a–j) required by HB 1600. Our approach leveraged WestEd's existing research base, professional networks, and national technical expertise to highlight the top exemplars for each focus area. This approach ensured efficiency, breadth, and immediate relevance to Virginia's design needs.

Exemplar Selection Approach

WestEd organized its analysis around the 10 focus areas and identified states or systems that are nationally recognized for excellence in each. Selection criteria included

- demonstrated implementation of innovative or effective practices within the past 5 years;
- availability of public artifacts (e.g., blueprints, rubrics, reporting templates, technology documentation);
- relevance to Virginia's goals for innovation, equity, and instructional usefulness; and

- evidence of positive reception, scalability, or successful evaluation, as available.

The exemplars for the focus areas may come from different states or consortia, depending on which system shows the strongest practice in a focus area. Based on WestEd's existing knowledge of state systems, potential exemplars were pre-identified for each focus area. This list was then vetted and refined through consultation with VDOE, additional research, and input from internal WestEd experts. States and consortia that were used as exemplars for each focus area, and broad reasons for their selection, are listed in Table 4 in Section III.

Evidence Collection

WestEd researchers gathered evidence across the identified exemplars. A lead researcher first oriented the team to the project's purpose and goals and established a shared understanding of the evidence needed across the 10 focus areas. After reviewing the project background materials, the researchers conducted an internal calibration to ensure consistent interpretation of the evidence-collection criteria and to align on the types of sources and documentation to be gathered for each focus area.

For each exemplar, WestEd researchers collected and analyzed the following types of documentation:

- technical reports, test blueprints, and content specifications
- administration guidance and timelines
- released items, rubrics, performance level descriptors, and scoring exemplars
- reporting templates and data dashboards
- accessibility manuals and technology infrastructure details
- cost, RFP, and contract documentation
- peer review documentation

WestEd researchers first collected publicly available documentation from the identified exemplars and conducted an initial review to confirm relevance and to extract high-level findings most pertinent to Virginia. In a few cases, WestEd collected additional documents directly from states as part of interviews focused on procurement practices. Next, researchers conducted a cross-exemplar analysis within each focus area, using a consistent framework to evaluate technical quality and validity evidence; compare approaches; identify strengths, limitations, scalability, cost, implementation complexity, and timeline; and consider feasibility for the Virginia context. Finally, they synthesized the evidence for each focus area, highlighting the most relevant features for Virginia's modernization efforts and their implications for future design choices.

E. Stakeholder Engagement

WestEd and QIP each played a role in engaging stakeholders to inform the final report. Prior to the work beginning, the Virginia Legislature directed VDOE to establish a work group related to implementation of an innovative assessment system. WestEd supported these work group meetings by developing agendas and delivering content to support the work group's efforts. In addition to the work group meetings, QIP conducted broader statewide engagement through a

series of focus groups with various groups of stakeholders from across the state. Engaging both a formal work group and broader stakeholders from across the state served two purposes: the work group provided representative and expert input on technical design, while the broader engagement with stakeholders across the state attempted to build public understanding and surface implementation concerns early, when they can still inform design decisions rather than derail implementation. For more details, see Section IV of this report.

Additional Engagement Activities

QIP conducted additional stakeholder engagement through a series of focus groups across the state; the methods and findings of these activities are described in Section IV.B and in Appendix D.

Recommendations Development

Recommendations based on these findings are organized into a tiered options framework that reflects different levels of change to Virginia's current assessment system. Each tier integrates features from multiple exemplar states and balances innovation with feasibility, cost, and implementation timeline. This framework allows Virginia to choose the right balance of innovation, cost, and implementation risk. Table 3 provides a high-level overview, incorporating feedback from VDOE. Section V details each option, including design components, alignment with Virginia legislation, exemplar states informing each feature, estimated costs and timeline, and readiness for implementation.

The work group met six times between October 2025 and March 2026. WestEd facilitated the final three meetings (December 2025, January 2026, and March 2026), which focused on reviewing comparative findings, developing draft recommendations, and validating the implementation framework. For details on the work group's review and recommendations, see Section V.

Table 3. System Design Options Framework

Option	Description	Feasibility	Cost Implication
Core	Makes improvements within the current SOL system, such as enhanced item quality, improved accessibility options, better reporting, and stronger alignment with standards. This option minimizes disruption and vendor transition costs.	High	Moderate
Enhanced	Builds on the Core tier with adaptive testing and development of performance tasks.This option requires vendor capability assessment and phased implementation but maintains system stability.	Moderate	Moderate–High
Transformative	Incorporates competency-based pathways and expanded local assessment flexibility as envisioned in HB 1957. This option requires pilot testing, teacher training systems, and scoring moderation processes before statewide scale.	Lower (requires extensive preparation and piloting)	Higher

III. Key Findings from Comprehensive State Scan, Organized by Integrated Priority Cluster

This section presents the key findings from a comprehensive scan of state assessment systems across the United States, organized around Virginia's 10 assessment priorities. For each focus area, the review synthesized evidence from a curated set of states and consortia identified as exemplars, drawing on publicly available technical reports, test design documentation (e.g., blueprints and specifications), scoring and reporting resources, and program guidance. The scan examined 318 pieces of evidence across the 10 focus areas, analyzing documentation from more than 15 state assessment systems and multistate consortia. The goal was not to catalog every feature of each system, but to surface the most actionable design patterns, implementation conditions, and documented evidence to inform Virginia's assessment modernization decisions.

The evidence from the systematic scan suggests that high-quality assessment systems have several common characteristics: strong alignment documentation; transparent communication of performance expectations; multiple pathways for students to demonstrate learning; and timely, actionable reporting. A consistent pattern also emerged across exemplars: the same states and consortia frequently served as strong models across multiple focus areas, indicating that the 10 focus areas are best interpreted as closely linked priority clusters rather than as isolated areas. Four broad integrated priority clusters emerged as a result of the state scan (see Figure 2 and Table 4). The analysis revealed several states demonstrating strong practices that Virginia may consider for its assessment system design. Figure 2 identifies states that emerged as significant exemplars within each priority cluster. In addition, as shown in Table 4, multistate consortia such as the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Dynamic Learning Maps (DLM) Consortium emerged as prominent exemplar systems across the priority clusters.

Appendix B includes detailed exemplar profiles organized by focus area from states and consortia. This section presents brief results organized by priority cluster.

Figure 2. Exemplar States by Integrated Priority Cluster

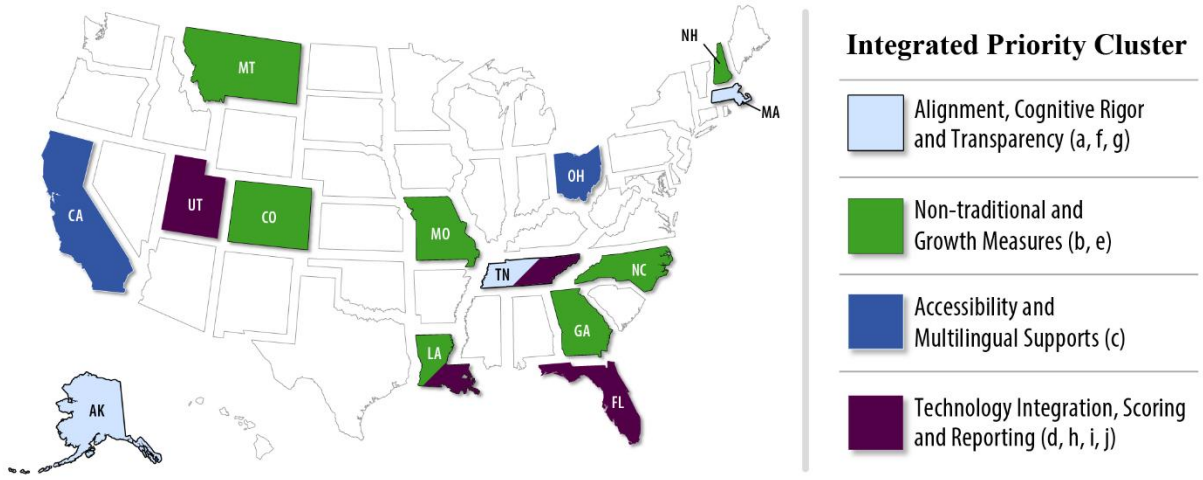


Table 4. Exemplar States and Consortia Across Integrated Priority Clusters

Integrated Priority Cluster	Illustrative Exemplar	Relevance
Alignment, Cognitive Rigor, & Transparency (a) Alignment with standards and instructional goals (f) Application of knowledge and higher-order reasoning (g) Transparency through release of rubrics, performance level descriptors, item maps, and sample items and tasks	(a, f, g) Massachusetts (MCAS) (a, f, g) Smarter Balanced (a, g) Tennessee (TCAP redesign) (g) Alaska (AK STAR)	(a) Demonstrate strong content coherence and instructional usability models for SOL alignment (f) Provide item-level exemplars and scoring frameworks for higher-order cognitive assessment (g) Offer strong release and communications practices to build public trust and instructional alignment
Non-traditional & Growth Measures (b) Non-traditional, project-, performance-, and competency-based approaches (e) Authentic growth measures and formative feedback	(b) Colorado (Alternate Assessment) (b) New Hampshire (PACE) (b, e) Louisiana & Missouri (IADA) (b, e) Georgia (IADA & Georgia Student Growth Model) (e) Montana (MAST) (e) North Carolina (NCPAT/IADA)	(b) Offer models for integrating authentic, classroom-based tasks within statewide systems (e) Demonstrate actionable, formative growth measures compatible with statewide systems
Accessibility & Multilingual Supports (c) Accessibility and multilingual supports	(c) California (CAASPP) (c) Ohio (c) Dynamic Learning Maps (DLM)	(c) Highlight inclusive design practices ensuring equitable participation and comparability
Technology Integration, Scoring & Reporting (d) Advanced technology integration and infrastructure (h) Innovative, reliable, and valid scoring practices (i) Reporting to students, families, and educators (j) Timely reporting to the SEA	(d, h, i, j) Florida (FAST automated scoring and dashboards) (d, h, j) Smarter Balanced (adaptive engine, distributed scoring & reporting APIs) (d, i) Utah (RISE technology integration and reports) (h) Louisiana (teacher-moderated scoring) (j) Tennessee (EdTools and InformTN)	(d) Illustrate scalable, adaptive architectures and operational readiness to inform Virginia's infrastructure planning (h) Demonstrate scalable, valid scoring models with robust QC and timely turnaround (i) Model rapid, understandable reporting that directly informs instruction and family engagement (j) Provide blueprints for state-level data pipelines, enabling timely division support and accountability

As summarized in Figure 2 and Table 4, the exemplar evidence is logically grouped into four integrated priority clusters that help organize the key findings from the scan. Following are brief descriptions of each of these priority clusters and exemplar systems that illustrate various states' approaches. Additional details on each priority cluster are described in Appendix B.

Alignment, Cognitive Rigor, & Transparency (a, f, g)

Alignment with standards and instructional goals (a), application of knowledge and higher-order reasoning (f), and transparency (g) function as a single integrated priority cluster. States and consortia with systems that exemplified alignment to standards and instructional goals often also demonstrated strong application of cognitive rigor, as well as strong transparency in how they release items and tasks. The strongest exemplars for this priority cluster were the Massachusetts Comprehensive Assessment System (MCAS) and the Smarter Balanced consortium. Both of these systems provide exemplary alignment documentation; show that rigorous alignment is not just a technical mapping exercise but is reinforced through item and task designs that elicit application and higher-order reasoning; and provide consistent public-facing resources (e.g., blueprints, released items, rubrics/exemplars, and performance descriptors) that make expectations visible and instructionally usable. The Tennessee Comprehensive Assessment Program (TCAP) redesign and Alaska's AK STAR system further underscore the implementation value of clear, educator-facing blueprinting and assessment transparency.

- ***Massachusetts (MCAS) (a, f, g).*** The MCAS is a strong exemplar of all three focus areas within this priority cluster. MCAS combines rigorous alignment documentation with clear expectations of cognitive demand in its item and task design, as well as strong transparency practices. MCAS alignment is supported through strong technical documentation and blueprinting that translates the Massachusetts Curriculum Frameworks into explicit design targets, while the system's cognitive rigor is reinforced through depth-of-knowledge levels assigned to items and tasks that require analytic writing and higher-order reasoning skills. Massachusetts also demonstrates extensive transparency practices through consistently releasing item sets and tasks across grades and subjects, published rubrics, and curated sample student work with scoring guidance. These released artifacts are not only instructionally useful to educators but also help other stakeholders, such as families, see what proficient performance looks like and how it is evaluated in practice.
- ***Smarter Balanced Assessment Consortium (a, f, g).*** Smarter Balanced exemplifies alignment, cognitive rigor, and transparency through the use of a coherent, system-level design framework. Smarter Balanced uses the evidence-centered design (ECD) framework that makes the logic of the assessment visible from standards to tasks. Its ECD approach and claims/targets architecture provide a strong alignment structure that explicitly organizes what the assessment intends to measure and the kinds of evidence that students must produce, including through performance tasks and constructed responses designed to elicit higher-order reasoning. Smarter Balanced reinforces transparency through publicly accessible specifications, blueprints, sample items, practice tests, and performance level descriptors. Professional learning resources, including Tools for Teachers, support educators in understanding the relationship between instruction, standards, and assessment expectations. The result is a well-documented model of how to maintain technical coherence while also supporting instructional usability and public understanding.
- ***Tennessee (TCAP Redesign) (a, g).*** Tennessee exemplifies this priority area through its emphasis on practical, educator-facing alignment tools that make standards-to-test expectations concrete. TCAP provides blueprints that clearly specify how Tennessee

Academic Standards are represented in operational forms, laying out the distribution of items/points across reporting categories and standards, as well as providing an accessible mechanism for educators to understand the intended emphasis of the assessment across grades and content areas. Tennessee further supports transparency through assessment overviews and released items that allow educators to examine operational content and see how standards are instantiated in item design. Although Tennessee's publicly available technical documentation is less expansive than those of Massachusetts or Smarter Balanced, its test blueprinting and item-release ecosystem provides a transparent, implementation-oriented model for communicating alignment and expectations to various stakeholders.

- ***Alaska (AK STAR) (g).*** Alaska provides an exemplary model for assessment transparency, showing what making assessment expectations and results accessible to different audiences looks like. The AK STAR portal provides a strong set of educator- and family-facing guides and an interactive Achievement Level Explorer that helps users understand what performance levels mean by grade and subject, making expectations easier to navigate than through static documents alone. Alaska also supports transparency through practice tests and readiness materials that help students and educators understand assessment formats and expectations.

Non-traditional & Growth Measures (b, e)

Non-traditional and performance-based approaches (b) and authentic growth measures and formative feedback (e) form a combined innovation-and-growth cluster because many exemplar models rely on distributed measurement across the year and localized classroom-connected evidence of learning to successfully implement innovative assessments. Exemplars such as New Hampshire (PACE), Colorado (Alternate Assessment), and Louisiana and Georgia (Innovative Assessment Demonstration Authority [IADA]–related work) illustrate the promise of performance tasks and competency-oriented systems, while Montana (MAST), North Carolina (NCPAT/IADA), and the Georgia Student Growth Model highlight how growth and in-year feedback can be structured to support instruction. Across these examples, the consistent finding was that innovation becomes viable when states pair standards-aligned innovative task formats and through-year designs with consistent psychometric comparability mechanisms (e.g., equating, scoring calibration), provide realistic and distributed local implementation opportunities, and provide clear interpretation support to reflect authentic classroom expectations and be instructionally useful.

- ***Colorado (Alternate Assessment) (b).*** Colorado's CoAlt Science assessment, designed for students with significant cognitive disabilities, provides a strong example of how performance-oriented tasks can be integrated into a technically robust statewide system. The assessment includes supported performance tasks alongside selected-response items, and its technical reporting documents structured item development, field testing, and reliability analyses. Although it is not a general-education IADA system, CoAlt demonstrates how performance-based components can coexist with standardized scoring and comparability controls.
- ***New Hampshire (PACE) (b).*** New Hampshire's Performance Assessment of Competency Education (PACE) is the first federally approved innovative assessment pilot, ensuring that students demonstrate mastery of critical knowledge and skills. Rather than relying

solely on annual statewide tests, PACE is centered on locally developed and locally administered performance assessments tied to clearly defined grade and course competencies, with cross-district common tasks used to calibrate scoring and ensure comparability. The system includes a task bank of previously validated assessments as well as educator networks for building local capacity that illustrate how performance-based approaches can be embedded within a coherent comparability framework. However, challenges have been encountered and reported in scaling the program statewide.

- ***Louisiana (IADA) (b, e).*** Louisiana's IADA model integrates curriculum-embedded performance tasks into a structured through-course design that supports both non-traditional measures and growth measurement. By administering assessments in multiple windows (fall, winter, and spring) and tying them directly to high-quality, state-endorsed curricula, Louisiana enables students to demonstrate recently acquired knowledge in authentic curriculum-embedded contexts, rather than relying on a single spring summative assessment. The technical reporting also documents comparability evidence across test forms, years, and languages, ensuring that distributed, curriculum-connected assessments can still yield defensible summative interpretations.
- ***Georgia (IADA and Georgia Student Growth Model) (b, e).*** Georgia illustrates the interplay between innovative assessment approaches and robust growth modeling. While its IADA application signaled interest in alternative structures, Georgia's most comprehensive innovation is the Georgia Student Growth Model, which uses vertically scaled assessment data to generate student growth percentiles and growth targets. Georgia demonstrates how states can integrate authentic assessment evidence with systematic growth reporting by pairing standards-aligned summative assessments with a well-developed longitudinal growth model and providing educators and families with guidance for interpreting those results.
- ***Montana (MAST) (e).*** The Montana Aligned to Standards Through-Year (MAST) program exemplifies a statewide shift toward distributed growth measurement through multiple assessment windows. MAST replaces the traditional, single end-of-year summative assessment model in Grades 3–8 ELA and mathematics with multiple, shorter standards-aligned assessment occasions administered throughout the year, with results aggregated to produce summative determinations for accountability purposes. The state's theory of action positions these shorter, instructionally aligned administrations as providing more authentic and continuous evidence of learning. Montana received federal flexibility approvals and Competitive Grants for State Assessments (CGSA) funding in 2022 and 2024, which support scaling this model to MAST Science.
- ***North Carolina (NCPAT/IADA) (e).*** The North Carolina Personalized Assessment Tool (NCPAT) integrates interim and adaptive summative components into a single growth-oriented system. The NC Check-Ins 2.0 provide interim feedback throughout the year, while the multistage adaptive end-of-grade assessments produce summative results on the same vertical scale. Data from interim administrations inform summative test form assignment and instructional adjustments, linking formative and summative measurement more tightly than in traditional systems.

Accessibility & Multilingual Supports (c)

Accessibility and multilingual supports that are grounded in universal design principles are central to building an equitable and technically defensible assessment system, especially as item types and online delivery become more complex. The exemplars for this priority area show how states operationalize accessibility through clear frameworks, decision supports, and documentation that protects both equity and validity.

- ***California (CAASPP).*** The California Assessment of Student Performance and Progress (CAASPP) system is grounded in UDL principles, ensuring that accessibility is built into assessment design, rather than added as an afterthought. The CAASPP system demonstrates comprehensive accessibility implementation, aligned with Smarter Balanced guidelines, through its state-maintained Accessibility Resources Matrix and Accessibility Resources Planning Tool. These tools provide detailed guidance on allowable supports and specify the universal tools, designated supports, and accommodations (embedded and non-embedded) that are available across assessments. In addition, companion quick-reference resources and implementation guidance allow educators to identify appropriate supports for individual students. For multilingual learners, California provides translated test directions in multiple languages and maintains separate assessments through its English Language Proficiency Assessments for California (ELPAC), with an associated information guide.
- ***Ohio.*** Ohio serves as an exemplar for strong statewide policy and decision support, using an accessibility manual that provides comprehensive guidance on accommodations. The accessibility manual frames accessibility as a core validity requirement, providing a structured system of supports (including administrative considerations, universal tools, designated supports, accommodations, and modifications) intended to ensure consistent implementation for students with disabilities and English learners. Ohio also provides policy-level guidance and assessment training resources for educators, which promote consistent, appropriate accessibility decisions and implementation statewide.
- ***Dynamic Learning Maps (DLM).*** The DLM Alternate Assessment System provides a model for assessing students with significant cognitive disabilities. The DLM consortium serves students in Grades 3–8 and high school across multiple participating states, with assessments in ELA, mathematics, and science. The DLM accessibility manual outlines a six-step process for selecting and implementing accessibility supports for students. It provides guidance for Individualized Education Program (IEP) teams, educators, and test administrators documenting required student supports and selecting appropriate accommodations. DLM's technical manual provides extensive validity evidence, including item development processes incorporating bias, sensitivity, and accessibility reviews.

Technology Integration and Scoring & Reporting (d, h, i, j)

Technology integration and infrastructure (d), scoring and validation (h), and reporting to educators and families (i) and to the SEA (j) operate as an end-to-end operational implementation cluster. The exemplars in this strand, notably Florida (FAST), Smarter Balanced's adaptive engine and reporting structures, and Utah (RISE) demonstrate that delivery platforms, scoring workflows, QC, and reporting timelines are interdependent. These systems

show that faster, more actionable reporting depends on design decisions made upstream (e.g., item types, scoring model, platform readiness), and that all reporting—regardless of the intended stakeholder—requires the same underlying data infrastructure (e.g., validation rules, refresh cycles, secure access, and role-based dashboards) and are therefore interconnected.

- ***Florida (FAST) (d, h, i, j).*** The Florida Assessment of Student Thinking (FAST) system exemplifies a fully integrated technology, scoring, and reporting pipeline. FAST operates as a statewide computer-adaptive progress monitoring system, with assessments administered three times per year, with each administration covering the full test blueprint aligned to the Florida Benchmarks for Excellent Student Thinking (B.E.S.T.) standards. The system relies on item response theory (IRT)–based adaptive delivery and centralized online administration infrastructure, supported by detailed technology and device guidance for districts. What distinguishes Florida most clearly is the integration of automated scoring and near-immediate reporting; predominant use of selected-response items and machine-scored components allow results to populate in the Family Portal within approximately 24 hours of test completion during the testing window. Florida also provides educator-facing reporting guides and structured dashboards that differentiate reporting for families, teachers, and administrators, with appropriate interpretative guides for each score user. At the SEA level, statewide data infrastructure supports centralized access and reporting, demonstrating how test delivery, scoring, and reporting operate as a single coordinated system.
- ***Smarter Balanced (Adaptive Engine, Distributed Scoring and Reporting APIs) (d, h, j).*** Smarter Balanced exemplifies a mature adaptive system (consistent since 2014–15) that is supported by robust scoring and reporting documentation. Smarter Balanced's CAT engine incorporates blueprint constraints to ensure content coverage, item exposure controls to protect item banks, and precision optimization procedures to maximize measurement accuracy. The consortium's technical reports document reliability, measurement precision, and error analyses, reinforcing that adaptive delivery decisions are grounded in psychometric evidence. Smarter Balanced also employs a hybrid human/AI scoring model for constructed responses and performance tasks, supported by documented reliability and calibration processes. On the reporting side, standardized score types enable member states to integrate results into their own dashboards and data systems through detailed API documentation. This combination of adaptive delivery logic, documented scoring validation, and interoperable reporting standards positions Smarter Balanced as a system-level coherence exemplar.
- ***Utah (RISE) (d, i).*** Utah's Readiness, Improvement, Success, Empowerment (RISE) assessment system illustrates how a statewide adaptive platform can be operationalized within a state-specific accountability and reporting framework. RISE provides standardized proficiency and reporting category data, aligned to Utah's standards, delivered through a CAT platform. Utah supplements its technical infrastructure with item type tutorials and online testing technology guides that prepare districts for implementation, signaling the importance of readiness and capacity-building alongside platform adoption. Reporting materials explain scale scores, proficiency levels, and reporting categories in plain language for families, while public dashboards provide aggregated results. Although Utah's publicly available technical documentation is less expansive than Florida's or Smarter Balanced's, the RISE model demonstrates how

adaptive delivery, centralized vendor partnership, and differentiated reporting tools can function together within a statewide accountability system.

- ***Louisiana (Teacher-Moderated Scoring) (h).*** Louisiana provides an exemplar for integrating scoring rigor into its innovative assessment design, which includes curriculum-embedded and performance-based components. Through its Louisiana Educational Assessment Program (LEAP) 2025 system, Louisiana documents scoring procedures, including inter-rater reliability evidence and calibration processes, within its technical reports, particularly for constructed-response components. The model incorporates QC procedures that generate response distribution reports to inform recalibration that maintain comparability across administrations. Louisiana's scoring timelines are coordinated with its statewide assessment calendar, reflecting the operational relationship between human scoring capacity and reporting expectations.
- ***Tennessee (EdTools and InformTN) (j).*** Tennessee exemplifies how assessment results can be integrated into a broader state data ecosystem. While TCAP delivery is not as heavily adaptive as Florida's FAST, Tennessee's EdTools and InformTN platforms illustrate how statewide data systems can centralize assessment information for educator and SEA use. InformTN functions as an interactive dashboard that provides data visualizations and filters organized around key accountability indicators, including achievement and growth, enabling state and district leaders to analyze results across domains. The Education Information System (EIS) is a statewide student information and reporting infrastructure that supports the collection and reporting of assessment and accountability data, ensuring statewide reporting consistency. Finally, while InformTN is a formal statewide data analytics platform, EdTools is a separate educator portal that provides assessment resources, practice tests, and other support materials for educator use.

Implications for Virginia

The findings from the state scan, consolidated across Virginia's 10 priority areas and across the four integrated priority clusters, suggest that Virginia will benefit most from treating assessment modernization as an integrated system redesign problem. Overall, several cross-cutting implications are clear for Virginia as it considers next steps in its assessment modernization timeline.

- ***Assessment modernization is best treated as a coherent and integrated system redesign problem:*** building instructional coherence through alignment and transparency; ensuring sufficient infrastructure for capacity and comparability for any shifts toward nontraditional models; integrating accessibility through universal design principles; and selecting technology, scoring, and reporting approaches as one coordinated pipeline that can deliver timely reporting to all educational stakeholder and score users at scale. The exemplars that appear across clusters, such as Massachusetts, Smarter Balanced, and Florida, demonstrate that alignment, cognitive rigor, technology integration, scoring models, and reporting structures function interdependently. Decisions about item types affect scoring timelines, scoring models influence reporting speed, reporting formats shape instructional use, and blueprint structures determine whether cognitive rigor is preserved at scale. For Virginia, this suggests that procurement, piloting, and implementation planning should be organized around integrated system architecture.

- ***The exemplar states show that assessment expectations should be simultaneously technical and public-facing, so that educators can connect what is tested to what they teach.*** Massachusetts and Smarter Balanced provide examples of claim/target frameworks, blueprint specifications, and validity documentation, and pair those technical artifacts with released items, rubrics, exemplars, and achievement level descriptors that educators can inspect and use. Tennessee reinforces the importance of blueprint transparency for instructional planning. For Virginia, this implies that modernization should include technical documentation for defensibility and peer review plus educator-facing transparency artifacts that make expectations stable and interpretable during transition.
- ***Innovation in format should be matched by psychometric comparability, educator capacity, and infrastructure.*** The non-traditional and growth-measures-focused exemplars (e.g., New Hampshire, Louisiana, Missouri, Montana, North Carolina) demonstrate that performance-based and through-year systems are viable only when supported by psychometric research, scorer training, data governance, and reporting clarity. Georgia's experience further illustrates that growth analytics should be integrated into system design, not layered on afterward. For Virginia, this means that any movement toward distributed testing, competency-based approaches, or performance tasks should include up-front investment in scoring calibration, audit mechanisms, and educator professional learning.
- ***Technology decisions are consequential far beyond delivery.*** Florida, Utah, and Smarter Balanced illustrate that adaptive engines, automated scoring, and rapid reporting are tightly connected to infrastructure readiness, vendor partnerships, and statewide data systems. These examples show that technology, scoring, and reporting operate as a single data pipeline, serving educators, families, and the SEA. For Virginia, this implies that platform selection, scoring strategy (human, hybrid, automated), and reporting dashboards should be evaluated together, with explicit attention to cost, sustainability, and data integration with existing state systems.
- ***Transparency and accessibility are foundational to legitimacy***. California, Ohio, and DLM demonstrate that accessibility frameworks need to be explicit and systemwide to protect equity and validity. Alaska and Massachusetts show that interactive tools, released items, and clear reporting guides strengthen public trust and instructional alignment. For Virginia, this means modernization should include deliberate transparency design—what will be released, how often, and in what formats—as well as a comprehensive accessibility framework aligned to both traditional and performance-based item types.
- ***Successful modernization requires phased implementation, stakeholder engagement, and iterative validation.*** States that have piloted innovative approaches (e.g., Louisiana) did so under structured waiver conditions with documented evaluation and technical reporting. For Virginia, this underscores the importance of sequencing: piloting before scaling; documenting validity evidence at each stage; and ensuring that infrastructure, educator capacity, and public communication keep pace with technical change.

In sum, the exemplar scan indicates that Virginia's modernization effort will be most successful if it is designed as an integrated, standards-anchored system that balances rigor with usability, innovation with comparability, and technological advancement with transparency and equity.

IV. Work Group and Stakeholder Engagement

WestEd and QIP carried out two main strands of stakeholder engagement: facilitating meetings with the Innovative Assessment Work Group and focus groups that engaged stakeholders across the state. The work group, established by state statute (Item 119.A.3.d. [Chapter 725, 2025 Acts of Assembly]), consists of members from across the state with experience with Virginia's assessment system. The focus of the work group is twofold: to provide input on the needs of the assessment system and to analyze options for implementing an innovative new statewide assessment system. The work group was specifically asked to focus on project- and performance-based approaches, accessibility for English learners and students with disabilities, integration of growth measures, and resources that clarify benchmarks for success. Appendix A provides a full list of the work group members.

In addition to the work group, QIP conducted 13 stakeholder engagement sessions with 65 participants across all regions of Virginia in January 2026. Respondents included state- and division-level leadership, school administrators, classroom educators, parents, and community-based organizations. The QIP team contacted multiple participants per region to ensure geographic diversity, with participants drawn from each of the eight superintendents' regions across the Commonwealth. QIP analyzed the data using a thematic approach, prioritizing the preservation of participant sentiment and terminology. Findings represent the depth of experience of focus group participants and may not capture the breadth of every position or school division in the state. Appendix D provides a complete report on the focus groups.

A. Work Group

The work group held six meetings from October 2025 through March 2026 to gain deeper insights into the current assessment system and explore potential options to improve the system, including a full-day in-person session on January 9 and a virtual session on March 17. These two sessions focused on identifying and articulating a set of draft recommendations; this section describes the product of these two meetings. The recommendations prioritize a high-quality, transparent, and accessible assessment system that ensures operational efficiency and clear communication with stakeholders.

Under Executive Order 4 (January 17, 2026), the Governor directed VDOE to continue convening the Virginia Innovative Assessment Work Group on a quarterly basis as an advisory body to VDOE and the Office of the Governor. The recommendations documented in this section are not final and will continue to evolve as the work group's charge continues under Executive Order 4.

Draft Work Group Recommendations

As part of the aforementioned two sessions, work group members came to substantial agreement around the following recommendations:

- ***Prioritize summative assessments.*** Work Group members mostly agreed that VDOE should focus its support for the state's assessment system on summative assessments to provide clear milestones for student outcomes. The work group particularly focused on the importance of a clear multiyear plan and timeline for implementing a robust and efficient summative assessment program while recognizing the time needed to build such

a system. While the work group tended to focus on VDOE's role in the summative assessment system, many members mentioned that formative and interim assessments also play an important role in a balanced, coherent assessment system. Most work group members felt that VDOE should allow school division leaders discretion in selecting and implementing formative and interim assessments while implementing reasonable guardrails that ensure alignment within the broader assessment system.

- ***Focus on alignment and rigor.*** Many work group members highlighted the importance of the assessment system aligning with the full depth and breadth of Virginia's academic content standards. However, the work group recognized that there is tension between this recommendation and reducing testing time. The work group also rejected the dichotomy between content knowledge and critical thinking, as students must possess foundational knowledge to apply higher-order reasoning, and differentiated state and local roles, with the state focused on assessing the most essential concepts in the summative system and school divisions focused on more granular assessment through formative and interim systems.
- ***Incorporate accessibility from the outset.*** Work Group members strongly agreed that new assessments must be developed using UDL principles that result in meaningful accommodations. The work group agreed that guidance around accommodations for specific groups (such as English learners) needs to be further refined and specified.
- ***Modernize data systems.*** The work group strongly recommended that VDOE invest in the implementation of automated data systems that reduce the reporting burden on local school divisions and encourage fast, accurate scoring of student assessment results. Work Group members emphasized that data systems should be implemented consistently, with clear communication and timelines that allow for educators and parents to learn the system. Attention should be paid to the overall ecosystem of data systems to ensure that the reported information balances comprehensiveness, usability, and efficiency.
- ***Provide consistent and simple reporting on student outcomes.*** The work group zeroed in on simplifying the reporting of student growth measurement and standardizing performance levels. These topics have connections to the assessment system, and there is also significant overlap with the School Performance and Support Framework, the state's accountability system. Any adjustments made on this topic should focus on alignment between the state's assessment and accountability systems and should be undertaken by the entities most closely related to the specific topic. In 2025, the Joint Legislative Audit and Review Commission published a study (2025) that provides direct ties and recommendations on this topic.
- ***Increase public trust through transparency and consistency.*** The work group identified two main areas where VDOE could improve trust related to the assessment system. First, the work group recommended releasing sample test bank items, rubrics, and examples to the public to provide a critical feedback loop and to give educators actionable tools for instruction. Second, many members emphasized the need for stable revision cycles for SOL assessments, with enough time to allow for adequate transitions. The work group suggested a minimum 4-year cycle for revisions, and suggested providing school divisions a minimum of 18 months with the full resources (e.g., standards, instructional guides, assessment blueprints) before implementing changes to the assessment.

In addition to these six recommendations, smaller subgroups within the work group highlighted three issues. While the following issues were less prioritized across the work group, compared to the six recommendations, , many members felt that they were still worthwhile for the state to consider.

- ***Reevaluate graduation standards.*** Given the recommendation for alignment between assessments and content standards, many work group members recommended that VDOE reevaluate graduation standards. Work Group members highlighted the need to reevaluate verified credits to ensure alignment with the assessment system. Although this topic relates to the assessment system and needs to be aligned with assessment system changes, it is outside the scope of this work group, and other entities would likely need to lead work in this area.
- ***Streamline high school assessments.*** Given the number of assessments administered in high school grades, some members of the work group recommended that VDOE conduct a cost and value analysis to identify areas for improved efficiency. Currently, there are several low-utilization assessments, including a number of EOC assessments, that could be combined with other assessments or eliminated to improve the efficiency of the assessment system.
- ***Optimize testing methods.*** Most work group members supported the continued use of CAT for summative assessment purposes, with some disagreement within the work group. Discussion within the work group highlighted a significant need for better education regarding CAT among stakeholders, to address misconceptions and to highlight the advantages that CAT assessments provide.

B. Additional Stakeholder Engagement

QIP staff conducted a thorough stakeholder engagement effort to understand user experiences with the current assessment system. Staff hosted 13 focused engagement sessions with 65 participants across all regions of Virginia in January 2026. Respondents included state- and division-level leadership, school administrators and classroom educators, and parents and community-based organizations. Following are key findings from these engagement sessions, which are explained in more detail in Appendix D.

Key Findings

- ***Strengthen communication and instructional alignment.*** Stakeholders reported that they desired enhanced transparency about test design and mechanics, and that they wanted structured design documentation to better align daily classroom instruction with state expectations and to reduce burden.
- ***Enhance data utility and reporting accessibility.*** Stakeholders emphasized the importance of quick data turnaround to inform summer interventions and start-of-year placements. Parents said they desired reports that contextualize a student's growth trajectories and their performance relative to their peers.
- ***Optimize developmental and curricular appropriateness.*** Participants noted that the length of testing sessions can exceed the physical stamina of younger learners and students with specific learning needs, potentially compromising these students'

demonstration of academic knowledge, and that language complexity can be a barrier for English learners or students with low reading skills.

- ***Improve efficiency and technological integration.*** While the transition to digital-only formats offers significant advantages, stakeholders identified a "digital divide" where device navigation skills impact performance. They called for better alignment between state assessments and external diagnostic tools to reduce test-taker confusion.
- ***Balance innovation with academic rigor.*** Participants discussed the future of assessment formats, including technology-enhanced items and simulations. While there was interest in these approaches, participants emphasized that any shift toward performance-based assessments should address concerns about administrative burden, local scoring validity, and the potential impact of AI. Participants also stressed that critical thinking and transferable skills should remain anchored in a solid foundation of factual knowledge and procedural fluency.
- ***Develop systemic stability and accountability frameworks.*** Stakeholders were concerned about the high-stakes nature of verified credits required for graduation. Division leaders sought a return to structured, multiyear assessment cycles to improve instructional alignment. Participants expressed a desire for accountability systems to incentivize resource allocation rather than to punish poor performance.

V. System Design Options and Recommendations

Virginia's assessment modernization is not a blank slate exercise. It takes place within the context of federal requirements, enacted state statutes, and legislative directives that, together, establish what is federally mandated, what Virginia statute requires, what the General Assembly has directed, and what room remains for design choices beyond compliance.

A. Options Framework

Any modernized assessment system must satisfy two layers of non-negotiable requirements: federal requirements under ESSA and state statutory requirements. This section provides an overview of both. Appendix C provides additional information about federal requirements under ESSA, and Section II.B provides additional information about state statutory requirements.

Federal Requirements. Under ESSA, Virginia must administer annual assessments in ELA and mathematics in Grades 3–8 and at least once in high school, and in science at least once in each of three grade bands (Grades 3–5, Grades 6–9, and Grades 10–12). These assessments must align to state content standards, meet technical quality standards for validity and reliability, provide appropriate accommodations for students with disabilities and English learners, and produce disaggregated, accessible results that can be readily interpreted by all stakeholders. Virginia must also administer an annual ELP assessment and maintain a statewide accountability system that meaningfully differentiates schools. These requirements are fixed regardless of which design option Virginia selects, and any system that fails to satisfy all of them will fail federal peer review.

Virginia Statute: HB 1957. HB 1957 was enacted in May 2025, with many provisions effective July 1, 2026. It establishes a second layer of requirements. The statute requires that statewide assessments evaluate not only content knowledge but also application of knowledge, critical thinking, and related skills. It directs the Board of Education, with the assistance of independent testing experts, to conduct a regular analysis and validation process of these assessments. It sets a 45-day deadline for returning ISRs after each assessment window closes, a requirement that has direct implications for scoring and reporting system design. As amended by SB 200, it requires that all assessments be reported on a 100-point scale, in accordance with statewide guidance. It strengthens and further defines the existing framework for local alternative assessments, formalizing "mandatory" and "permissive" categories and establishing Board guidelines and quality criteria that govern how those assessments are developed and administered. SB 200 requires the Board to develop and make available best practices for grading and scoring LAAs by September 1, 2026. Two subsections, K and L, are contingent on reenactment by the 2027 General Assembly; these would require making prior-year assessments available to teachers, establishing a formal annual audit process for both mandatory and permissive LAAs, and setting item design requirements including open-ended questions and long-form writing.

Beyond what state law requires, the General Assembly has directed specific attention to particular dimensions of assessment modernization:

- ***HB 1600 Item 119.*** The 2025 Appropriations Act established this study and defined its scope. Section A.3.c identifies 10 focus areas that Virginia's assessment system should address: (a) alignment with academic standards and instructional goals; (b) non-traditional, project-based, competency-based, or personalized learning approaches; (c)

accessibility for students with disabilities and English learners; (d) advanced technology integration; (e) authentic growth measures and formative feedback; (f) higher-order reasoning, critical thinking, and real-world application; (g) transparency and public accessibility of assessment information; (h) reliable and valid scoring; (i) timely and actionable reporting to students, families, and educators; and (j) timely reporting to the state agency to support interventions and accountability. Each recommendation in this section can be traced directly to one or more of these focus areas. Section A.3.d directs VDOE to establish the work group whose recommendations have informed this report and requires that this final report be delivered to the Board of Education and designated legislative committees.

- ***SB 200.*** SB 200 was signed into law on April 6, 2026 and makes several significant revisions to HB 1957. SB 200 repeals the through-year growth assessment mandate, leaving through-year assessment as a design option rather than a requirement. It changes the 100-point scale requirement from "scored" to "reported," preserving technical scoring flexibility. It provides greater scheduling flexibility by modifying the EOC timing window and adding exceptions for graduating seniors, retakes, and multiday assessments. It exempts performance assessments from the one-per-quarter frequency limit, which has significant implications for LAA implementation. It exempts VAAP students from the timing, frequency, and grading provisions. It delays the prohibition on additional EOC assessments and the 10% course grade requirement until the second full school year after pilot implementation of the new assessment system. It strengthens LAA requirements around formative learning experiences, authenticity, and assessment security, and requires the Board to publish best practices for grading and scoring LAAs by September 1, 2026. A detailed analysis of all HB 1957 provisions as amended by SB 200 is provided in Section II.B.

What Innovation Means in This Report

In this report, innovation refers to specific, evidence-based design choices that go beyond what federal law and Virginia statute require, informed by the exemplar analysis described in Section III and Appendix B, and grounded in the 10 focus areas identified by the legislature.

Innovation in this framework is a spectrum of design decisions, each with its own evidence base, cost profile, timeline, and risk. Some are relatively straightforward extensions of current practice. For example, expanding item types to include more constructed-response and technology-enhanced formats builds on work that Virginia has already piloted (Virginia Department of Education, 2024) and reflects a clear direction in both HB 1957 and the exemplar evidence. Other decisions require more substantial changes to infrastructure, scoring operations, or educator capacity. Multistage adaptive testing (MST) for content-heavy subjects, hybrid scoring models that pair automated and human scoring, and growth reporting systems all involve vendor capabilities, psychometric development, and field testing before they can operate at scale.

The most dramatic design shifts involve structural changes to how Virginia defines and measures student learning. Competency-based assessment pathways, expanded local performance assessment systems with state-level moderation, and integrated formative-to-summative assessment architectures represent shifts that other states have pursued under federal waiver authority, typically with multiyear pilot periods. New Hampshire, Louisiana, Georgia, and

Missouri have all implemented versions of these approaches; their experiences are documented in Section III and Appendix B. However, these approaches require conditions that Virginia does not yet have in place: statewide scorer training and calibration systems, audit and moderation infrastructure, and sufficient item bank depth.

The tiered options presented in the following section are designed to satisfy all federal and state requirements. What distinguishes them from one another is how far beyond compliance they reach, how quickly they can be implemented, and what they demand in terms of cost, capacity, and risk tolerance. The Core tier addresses statutory obligations and makes targeted improvements within the current system architecture. The Enhanced tier adds design features that require pilot testing and phased implementation over 1–3 years. The Transformative tier incorporates structural changes that require federal approval, extensive field testing, and a 3- to 5-year development timeline.

This framework is designed to support VDOE in improving its state assessment system. Each tier represents a coherent package of design choices, but the boundaries between tiers are not rigid. Virginia may choose to implement the full set of features within a single tier, or it may adopt a phased approach that begins with one tier and builds toward another over successive procurement cycles. However, a constraint on implementation order is sequencing: features in a higher tier generally depend on capacities built in a lower tier. For example, adaptive testing for EOC subjects requires an item bank large enough to support adaptive routing, which, in turn, requires sustained item development investment that begins in the Core tier.

B. Option Descriptions

The tiers differ in scope of innovation, implementation timeline, and level of investment and risk.

Core: Foundational Modernization

The work group's strongest consensus, with approximately 75% of members voting it as a top priority, was that the state's first responsibility is a high-quality summative system. The Core tier reflects that priority. It addresses what Virginia needs to do under statute and sets the technical foundation for future improvement. It does not require structural changes to the current assessment model. The procurement and transition to a new vendor can proceed under this tier without federal approval, waiver applications, or multiyear pilot programs.

Design Overview. The Core tier retains the current summative, end-of-year model with computer-adaptive delivery in Grades 3–8 mathematics and reading. It does not add new assessment modes or alter the fundamental test architecture. It improves the quality of items, the clarity of reporting, the rigor of accessibility procedures, and the consistency of technical documentation.

Consistent with existing provisions in the Standards of Quality and as further specified by HB 1957 and SB 200, school boards must administer mandatory LAAs in every SOL subject area where no statewide assessment is administered during that school year. For Grades 3–8, this includes history and social studies (Grade 3 History, Virginia Studies, Civics and Economics, U.S. History to 1865, and U.S. History 1865 to Present) and science in non-tested grades. The statute requires Board guidelines, comprehensive grading rubrics, and annual certification. SB 200 adds that teachers may not grade their own students' LAAs when those assessments are used

for verified credits, and that all summative performance assessments must be completed in a supervised, secure setting. These are new operational requirements that school divisions will need guidance and support to implement. The cross-teacher grading requirement poses a particular challenge for small and rural school divisions that may lack the personnel to comply. If a division has only one or two teachers in a subject area, finding qualified scorers who did not teach the students being assessed requires cross-division coordination that may not currently exist. Work Group members recommended elevating social studies as a leading subject for curriculum-embedded performance assessment, noting that its alignment with the College, Career, and Civic Life (C3) Framework and its position outside federal testing requirements create an opening to develop a scalable LAA model without triggering ESSA compliance constraints.

Features. The most immediate requirement is the continued implementation of the 0–100 reporting scale required by the statute. Although SB 200 changed the requirement from "scored" to "reported" on a 100-point scale, which means the underlying psychometric scale can use any metric that is technically defensible, the reporting phase still requires revised score reports, interpretive guides, and a proactive communications strategy to prevent the scale from being misread as percent correct. Currently, Virginia's planned approach is to report the 0–100 score alongside the existing scale score as a conversion, rather than conducting a new standard setting on the 0–100 metric (as South Carolina did for its End-of-Course Examination Program). This avoids the cost and timeline of a full standard setting, but it means the 0–100 number will be a mapped derivative of the existing scale. The mapping should be evaluated to ensure sufficient score differentiation, particularly near cut points.

The work group also recommended standardizing four performance levels (Below Basic, Basic, Proficient, Advanced) across all assessments. Grades 3–8 reading and mathematics already use four levels, and Grade 5 and Grade 8 science moved to four performance levels beginning in 2025–26. EOC and remaining assessments currently collapse the two lower levels into a single nonproficient category. Standardizing to four levels provides more diagnostic information and aligns with NAEP reporting conventions. However, expanding to four performance levels for the remaining assessments would require developing new performance level descriptors for Below Basic and Basic, conducting standard setting or other cut score processes to establish the new cut points, and updating all score reports and interpretive materials. This change also introduces complexity connected to graduation requirements and verified credits because the cut score that determines whether a student passes an assessment for verified credit purposes must be clearly defined within the revised performance level structure.

The Core tier also requires strengthening the item bank. The current item bank has greater depth at the lower end of the performance distribution than at the upper end and expanding coverage in the Proficient and Advanced ranges would strengthen measurement precision where classification decisions matter most. Under the new, higher performance standards, that imbalance will produce larger measurement errors and weaker decision consistency. Additional item development in the Proficient and Advanced ranges is a foundational investment that every higher tier depends on. Current retesting policy, which provides multiple attempts on EOC assessments, increases the number of forms required and places additional demand on the item bank by requiring more unique forms than most states need. Virginia should evaluate whether the current retake policy is sustainable, given the item bank investments this tier requires.

Test security requirements increase with each tier but should be addressed beginning with the Core tier. The Core tier requires item exposure controls sufficient to support multiple EOC testing windows and retake administrations without compromising item pools; secure browser requirements for online delivery; and protocols for incident reporting and form invalidation. The Enhanced tier's MST routing and the Transformative tier's locally administered assessments each introduce additional security considerations. The RFP should specify security requirements as a distinct evaluation category (see Section VIII.B).

Accessibility improvements in this tier include adopting UDL decision tools (e.g., student accessibility profiles, student needs matching tools), such as those used in the California approach, to help IEP teams match student needs to specific accommodations and modifications more consistently and systematically. Educators in focus groups cited inconsistent accommodation practices as a recurring concern; a more structured decision framework addresses that gap.

Transparency improvements include an expanded released-item library, documented scoring rubrics with student work samples at each performance level, and audience-specific interpretation guides for teachers and families. Family members in focus groups described the current state reporting system as a "black hole" for parents trying to understand their children's results. The work group's transparency recommendations align directly with this feedback. Alaska's interactive Achievement Level Explorer offers a useful model: an online tool that lets families see exactly what performance at each level looks like in practice. These transparency practices should extend beyond the SOL summative assessments. For LAAs, divisions will need access to model rubrics, anchor papers at each performance level, and exemplar tasks that illustrate what proficient student work looks like under the new requirements. Without these resources, the quality and consistency of LAAs will vary widely across divisions. Similarly, for the VAAP, released tasks, scoring rubrics, and exemplar student responses help educators working with students with significant cognitive disabilities understand assessment expectations and prepare students appropriately. States with mature released-item programs, such as Massachusetts, have found that publishing scored student work alongside rubrics is one of the most effective tools for building shared expectations across schools and divisions. The work group also recommended a broader assessment literacy plan, using plain language and developed in coordination with educator associations and parent networks. That plan is addressed in Section VI.D as a communication and change management requirement, but it should be understood as a Core tier deliverable. The transparency features of this tier will not achieve their purpose without a sustained effort to build assessment literacy among educators, families, and school board members.

Performance level descriptors (PLDs) can serve as one component of that assessment literacy effort. Virginia's performance levels (Fail/Below Basic, Fail/Basic, Pass/Proficient, Pass/Advanced) provide a shared definition of what it means to meet or not meet the proficiency standard, and professional learning structured around PLDs can help educators connect assessment expectations to instructional practice. Specifically, PLDs can support assessment literacy in several ways: they help educators interpret score reports by grounding reporting categories in descriptions of what students at each level know and can do; they provide a framework for analyzing student work and calibrating expectations across classrooms and schools; they inform the selection of classroom and interim assessments by giving educators criteria for evaluating alignment to standards and reporting categories; and they connect

assessment outcomes to instructional planning by helping educators identify where students fall on a learning progression and what instructional moves would support growth. While PLDs alone are not sufficient to build broad assessment literacy, incorporating them into a structured professional learning program that includes released items, scoring rubrics, and student work samples would strengthen educators' capacity to use assessment data meaningfully.

Legislative Alignment. This tier directly addresses the statutory requirements on reporting scale, reporting timelines, accessibility, and transparency. This tier also satisfies the HB 1600 focus areas of (a) alignment with standards; (c) accessibility; (g) transparency; (i) reporting to students, families, and educators; and (j) reporting to the state agency.

Work Group Alignment. The work group's recommendations align closely with the Core tier. Members specifically endorsed the four-level performance structure, UDL-based accessibility procedures, and expanded item and rubric release. The work group also recommended a minimum 4-year revision cycle for any new assessment design, with at least 18 months of lead time for educator resource development before changes take effect. VDOE may wish to consider reflecting this recommendation in the procurement process; for example, by specifying that no major changes to the assessment design, scoring, or reporting framework would occur within the first 4 years of operational use except to correct identified problems. There was no significant divergence on Core tier features.

Timeline. This tier's implementation may begin in Year 1 of the new contract, contingent on a timely RFP release and contract award. The 2027–28 testing year will likely need to focus on the foundational requirements of transitioning to the awarded vendor and delivering assessments without disruption. Phased implementation of additional Core tier features would follow as vendor capacity and system stability are established.

Risks. Work Group members noted the confusion caused by the vertical and horizontal scale scores, and although SB 200 repealed the statutory mandate for the through-year growth system, the underlying concern remains relevant: adding a 0–100 reported score to the existing scale score system increases the number of metrics that families are asked to interpret. In addition, a 0–100 scale is easily confused with percent correct and carries connotations of familiar grading scales, and this misinterpretation is difficult to prevent regardless of communication investment. The reporting strategy for the new system should address how these scores relate to each other and, ideally, reduce that complexity. VDOE will need to plan for significant communications and engagement as part of the implementation of this requirement in 2026–27.

Cost Range. Based on other state contracts and nonbinding vendor estimates gathered through Virginia's 2024 RFI, a scope similar to the current contract is estimated at $30M to $38M per year. Virginia's current contract costs approximately $46.5M per year, under a contract that has been in place since 2005. A competitive procurement would allow VDOE to evaluate pricing across multiple vendors and identify the most cost-effective approach for the state's needs. For context, Tennessee's recent contract ($28M per year) covers a broadly similar scope but with fewer EOC assessments and no growth component; and Massachusetts's contract ($37.9M per year, including alternate assessment) is for a system with fewer tests but with additional cost drivers such as translated assessments and ongoing item development to replace released items. The median of vendors' low-end and high-end estimates from the 2024 RFI spanned $23.5M to $31M for core assessment tasks; it is unclear whether those estimates included Virginia's alternate assessment.

These cost estimates carry important limitations. State assessment contracts are difficult to directly compare because states include different grade levels, subjects, and deliverables; and bundles cost differently across development, administration, and reporting and vary in makeup, retest, and accommodation policies. Several cost drivers could push Virginia's actual costs above the $30M to $38M range. EOC retesting across multiple windows adds meaningful test volume that is not reflected in most comparator contracts; spring 2025 alone included more than 540,000 EOC administrations. Item development intensity is the single largest variable: RFI estimates ranged from $2.3M per year for minimal development (10 items per grade/subject) to $9.6M per year for intensive development (200 items per grade/subject, up to 5 forms), and Virginia's item bank rebalancing needs and release requirements will push the cost toward the higher end of that range. Whether the RFI estimates included VAAP is also unclear; the current alternate assessment item development contract costs an additional $1.6M (approximately $600K per year, for just over 2 years) over and above the main contract, and if excluded from vendor estimates, actual costs would be correspondingly higher. Virginia may wish to explore participation in the DLM consortium, which manages alternate assessment development and administration across 24 states, at an estimated $1.5M to $2M per year.

Approximately 60% to 70% of contract costs are typically attributable to test administration, with the remainder covering item development, program management, and reporting. These estimates do not include internal VDOE capacity costs. VDOE should consider investing in a state data analyst, a communications specialist, and one or two additional content staff, at an estimated total of $400K to $600K per year in salary and benefits, to manage the procurement and the ongoing vendor relationship effectively. These roles would provide VDOE with independent technical capacity to evaluate vendor deliverables, manage communications during transition, and reduce reliance on the vendor for functions that are better handled in-house.

Alternate Assessment (VAAP). Virginia's current contract covers administration, scoring, and reporting for 17 alternate VAAP tests, with item development handled through a separate $1.6M contract (approximately $600K per year) as previously described. VDOE needs to determine whether to bundle VAAP into the general assessment contract or maintain it as a separate procurement. Bundling simplifies contract administration and may allow shared infrastructure across general and alternate assessments. Maintaining a separate procurement allows VDOE to evaluate specialized alternate assessment options on their own merits, including consortium participation, without tying those decisions to the general assessment award timeline. If it is bundled, the aforementioned cost estimates would increase accordingly. If it is maintained separately, Virginia may choose to consider joining an alternate assessment consortium such as DLM, which can reduce per-state development and maintenance costs, provide access to a shared item bank and established psychometric infrastructure, and offer peer-state collaboration on accessibility and scoring. Accessibility and accommodation features that serve both general and alternate assessment populations, such as UDL decision tools and accommodation assignment protocols, are addressed within each tier, regardless of how the VAAP contract is structured.

Enhanced: Innovation Core

The Enhanced tier adds design features that go beyond compliance and that require pilot testing, phased implementation, and expanded vendor capacity. Its timeline spans 1–3 years beyond the Core tier. The recommendation is designed to offer the option to dedicate the early contract years to Core tier implementation and later years to adding Enhanced tier features.

Design Overview. The Enhanced tier adds substantive changes to the Core tier architecture: MST for EOC assessments and AI-assisted scoring for constructed-response items. Each requires design work, field testing, and validation before operational use. None requires federal approval, which distinguishes this tier from the Transformative tier.

Features. MST delivers items in stages (testlets) rather than item by item, which preserves the integrity of passages and complex stimuli while still adapting to student performance. Virginia already uses item-level CAT in Grades 3–8. Extending MST to EOC subjects such as history and science addresses a measurement limitation in the current system: the EOC tests are not adaptive, which limits their precision. Stage-adaptive delivery also places less strain on the item bank than item-level adaptation (see Section III and Appendix B).

AI-assisted scoring of constructed-response items addresses the 45-day ISR delivery requirement in HB 1957. Constructed-response items, including extended writing tasks, cannot meet that timeline under traditional human scoring at scale. To meet reporting timelines, many states use a hybrid approach in which AI scores responses initially and human raters review flagged cases in addition to a random sample (see Section III and Appendix B for exemplar profiles). The vendor should be required to validate the scoring engine against human scores for all student demographic groups and to report results disaggregated by group. Research on automated scoring of NAEP constructed-response items identified English learner status as a significant source of differential scoring accuracy (Gregg et al., 2021; also see Ramineni & Williamson [2018] on differential scoring accuracy by demographic group in automated essay scoring more broadly). Virginia's RFP should build in explicit fairness monitoring requirements for both human and automated scoring.

Some approaches to growth measurement can be implemented with Virginia's current infrastructure, while others depend on Enhanced tier capabilities. Value tables, which credit schools for moving students toward proficiency using existing assessment results, do not require new measurement infrastructure. Other growth models, such as vertical scale growth, require an extended item bank and a field test that includes off-grade items to anchor the scale. Whatever growth construct Virginia selects should work alongside the 0–100 reporting scale without producing contradictory signals for individual students. Formalized IRW assessment in Grades 5 and 8 is also an Enhanced tier feature. The Virginia TAC recommended keeping reading and writing scores separate for growth analysis and potentially combining them for overall ELA accountability purposes. The RFP should ask the vendor to support both uses.

Legislative Alignment. This tier builds on the Core tier, which addresses all ten HB 1600 focus areas and advances HB 1957 requirements on growth measurement and reporting timelines. It adds particular depth in HB 1600 focus areas (b) performance-based approaches, (d) technology integration, (e) growth measures, (f) higher-order reasoning, and (h) scoring.

Work Group Alignment. The work group endorsed growth measures and AI-assisted scoring in principle, with specific cautions. In terms of growth, members preferred simpler models with strong educator interpretation resources, consistent with the Georgia Student Growth Model approach (see Appendix B for more details.) In terms of AI scoring, the work group accepted the approach, contingent on rigorous validation and fairness monitoring. There was no work group consensus on which growth construct to adopt, which underscores the need for VDOE to make a policy decision.

Timeline. This tier will require 1–3 years of pilot testing before full operational use. Structuring this work within a single multiyear contract is feasible and is preferable to issuing a separate procurement for Enhanced features.

Risks. AI scoring carries fairness risk that is manageable with proper validation requirements, but the research to successfully score a new program can take time. MST for EOC assessments requires sufficient item bank depth in multiple content areas (although less than item-level CAT); that capacity should be built during the Core tier phase.

Cost Range. These costs are incremental above the Core tier. The largest cost drivers are (1) item bank expansion to support MST routing, (2) AI scoring validation studies, and (3) standard setting. Based on comparator state pricing for selected-response item development, item development for MST-capable EOC pools could add a minimum of $500K per year above the Core tier's development baseline, depending on the number of subjects and target pool depth needed. (This figure assumes approximately 45–50 new items needed per EOC per year and an item cost of approximately $750/item). Additional detail on Enhanced tier pilot costs is provided in the Recommended Approach section (Section V.D). The validation studies required for AI-assisted scoring, including disaggregated fairness analyses across all demographic subgroups, are typically priced as special research studies; Texas's contract prices comparable work at approximately $205 per hour. These costs should be specified as discrete line items in the RFP so that VDOE can evaluate and control them independently of the base contract. VAAP costs are addressed under the Core tier regardless of procurement structure; no additional VAAP costs are anticipated for Enhanced tier features.

Transformative: Structural Redesign

The Transformative tier incorporates structural changes that go beyond what a standard procurement contract can deliver. It requires federal approval, multiyear pilot programs, and infrastructure that Virginia does not yet have: statewide scorer training and calibration systems, moderation protocols for locally developed assessments, and technical and governance frameworks. Critically, pursuing the Transformative tier does not pause the need to deliver assessments statewide. Virginia would need to continue operating at least the Core tier for all students while piloting Transformative approaches in a subset of districts, which means running two systems in parallel and absorbing the associated capacity demands and costs.

Design Overview. The Transformative tier centers on two changes: competency-based assessment pathways and expanded local assessment flexibility as envisioned in HB 1957 Subsections K and L. Both represent fundamental shifts in how Virginia defines and measures student proficiency, and both take time to build.

Features. Competency-based pathways allow students to demonstrate proficiency through multiple modes: locally developed performance assessments, portfolio submissions, and project-

based tasks moderated for comparability across school divisions. New Hampshire's PACE program, approved under IADA in 2018, is the most documented example. PACE allows participating districts to administer the state summative assessment only once per grade span, with locally developed assessments providing the annual measurement required by ESSA. The program demonstrated that performance-based systems can meet federal technical quality standards, but it also encountered significant challenges in statewide scaling. In 2021, the New Hampshire Department of Education withdrew formal support for PACE and returned to traditional statewide testing. The work of PACE continues through the New Hampshire Learning Initiative's Performance Learning and Assessment Consortium for Educators (PLACE) program, an educator-led network that builds on PACE's design principles without state infrastructure or IADA authority behind it.

Louisiana's CGSA-funded through-year assessment model (approved since 2022) and Missouri's IADA (approved for pilot testing in the 2025–26 school year) represent newer iterations of federally funded programs to pilot alternatives to traditional standardized testing. Massachusetts successfully transitioned science simulation tasks from an IADA pilot to operational use in its MCAS Science, Technology, and Engineering assessments, which is the closest example of a performance-based innovation reaching full statewide scale within a summative system. For additional details on the state approaches described in this section, see Section III and Appendix B.

HB 1957 Subsections K and L, which are contingent on reenactment by the 2027 General Assembly, would require VDOE to make all previous years' assessments available to teachers as practice tools and establish an annual audit process for local alternative assessments. These provisions, if reenacted and operational, would represent a starting point for the governance infrastructure that a Transformative tier approach requires, but significant additional capacity in scorer training, moderation, and cross-division coordination would still need to be built. SB 200 also exempts performance assessments from the one-per-quarter assessment frequency cap established in HB 1957, which creates statutory room for performance-based approaches even if Subsections K and L are not reenacted.

Legislative Alignment. This tier is most directly responsive to HB 1957 Subsections K and L and to the local alternative assessment provisions of the statute more broadly. It also addresses HB 1600 focus area (b) on non-traditional, project-based, and personalized learning approaches.

Work Group Alignment. The work group did not recommend the Transformative tier for immediate implementation. Members consistently prioritized a high-quality summative system over expanded local flexibility, and several raised concerns about the burdens that portfolio and performance assessment processes place on teachers. An instructional supervisor in a focus group noted that portfolio-based assessments resulted in more time assessing than instructing. The work group and broader stakeholders engaged through this study consistently prioritized improvements to the summative system over structural redesign. Pursuing the Transformative tier would require building both the technical infrastructure and the stakeholder understanding necessary to support a substantially different model of measuring student learning.

Timeline. This tier would require 3–5 or more years to develop and implement. Federal approval, whether through IADA or through a Section 8401 waiver, requires a detailed pilot design, evaluation plan, and evidence of comparability with the state summative system. The current federal administration has signaled support for state assessment flexibility, including

through Section 8401 waivers. Virginia would need to build that evidence base before applying, then implement a multiyear pilot in a limited number of districts before any transition to broader operational use. This timeline is not incompatible with starting now; it simply means that starting now produces results in the early 2030s, not in the current contract cycle.

Risks. Federal approval is not guaranteed and can take multiple legislative cycles to secure. Scaling competency-based systems statewide has proven difficult even in states with strong infrastructure and long lead times. The teacher capacity required to develop, score, and moderate performance assessments at scale is not currently widely available in Virginia. These are known challenges that are documented in the research scan, not theoretical concerns. New Hampshire's experience with PACE, described above, illustrates the risk that a state may invest years of development only to find that statewide scaling is not viable. Pilot costs would also be in addition to, not instead of, the costs of maintaining the statewide summative system for all non-pilot students and schools.

Cost Range. The Transformative tier would cost substantially more than the Core and Enhanced tiers, and these costs would be incurred on top of Core tier operating costs since the statewide system must continue to function during any pilot period. Precise estimates are not available because the pilot scope, federal approval pathway, and infrastructure requirements have not yet been defined. For reference, states that have pursued IADA-level innovation have typically relied on federal grants to offset development costs: Louisiana received $5.9M in CGSA funding across two awards (2022 and 2024), Montana received $7M, and Massachusetts invested $26M in Statewide Longitudinal Data System (SLDS) grants to build reporting infrastructure that supports its assessment innovations (SLDS grants fund data systems, not assessment development directly, but the reporting capacity that they create is a prerequisite for the kind of performance-based system the Transformative tier envisions). Virginia should explore CGSA and SLDS grant opportunities as funding vehicles for Transformative tier development but should not plan on grant funding as a substitute for sustained state investment in calibration systems, moderation networks, and training programs.

C. Comparison Table

Table 5 compares all three tiers across dimensions. Features are additive: the Enhanced tier includes all Core features, and the Transformative tier includes all Enhanced features.

Table 5. Tier Comparison by Dimension

Dimension	Core	Enhanced	Transformative
Assessment model	End-of-year summative; CAT in Grades 3–8 mathematics and reading; Grades 5 & 8 science	Added MST for EOC subjects; formalized IRW in Grades 5 and 8	Added competency-based pathways; LAAs with state moderation
Item types and item bank	Strengthened item bank (especially at Proficient and Advanced ranges); expanded constructed-response items; TEIs	Expanded bank to support MST routing for EOC; simulation and scenario-based items in EOC subjects	Performance tasks and portfolio components scored through statewide moderation
Scoring	Improved item quality; 0–100 reporting scale (statutory); four performance levels for all grades	Added AI-assisted scoring of constructed responses with fairness monitoring	Performance task/portfolio statewide scorer training

Dimension	Core	Enhanced	Transformative
Writing/IRW	Current writing assessment structure maintained	Formalized IRW assessment in Grades 5 and 8; separate scores for growth, combined option for accountability	Research emerging AI technologies to personalize writing prompts
Growth	Existing model maintained	Formalized individual student growth model if vertical scale or other infrastructure-dependent approach is selected	Explore multiple methods for demonstrating growth
Accessibility	UDL decision tools; improved accommodation/ modification decision-making practices	No structural change beyond Core	No structural change beyond Core
Reporting	Expanded released items; rubrics and student work samples; audience-specific guides; 45-day ISR target	Added family portal development; rapid reporting infrastructure	Research large language model–assisted personalized score reporting
Local flexibility	Current local alternative assessments maintained under HB 1957	No structural change beyond Core	Federally approved competency-based pathways; expanded local assessment options with statewide moderation
Federal requirements	ESSA compliant; no waiver needed	ESSA compliant; no waiver needed	Requires federal approval (IADA or Section 8401 waiver)
Primary legislation addressed	HB 1957 (core); HB 1600 focus areas a, c, g, i, j	HB 1957 (growth, reporting timelines); HB 1600 focus areas b, d, e, f, h	HB 1957 Subsections K and L; HB 1600 focus area b
Work Group alignment	Strong	Moderate (with caveats on AI scoring validation)	Conditional (work group prioritized summative system quality over local flexibility and innovative alternative approaches)
Implementation timeline	Current procurement cycle; operational by 2027–28	1–3 years pilot; operational by 2029–30	3–5+ years pilot; operational in early 2030s
Feasibility	High	Moderate	Lower (requires several pilot years, substantial costs, and federal approval)
Cost range	$30M–$38M/year	$5M–$8M/year above Core for item bank expansion, AI scoring validation, and growth model development	Substantially above Enhanced; states have relied on federal grants (CGSA for assessment innovation, SLDS for reporting infrastructure) to offset development costs

D. Recommended Approach

The exemplar analysis, work group input, and stakeholder feedback all point in the same direction: adopt the Core tier as the floor of the next procurement and structure the contract to build toward the Enhanced tier over the contract period. This is a sequencing argument, grounded in what Virginia's assessment system needs right now and what the field has consistently identified as necessary before adding complexity.

The Case for Core as the Floor

Virginia is entering a new procurement cycle with a fundamentally different set of statutory requirements than the last cycle. HB 1957 is not a marginal adjustment, and the majority of the changes are effective July 1, 2026. Implementing those changes well, under a new vendor, within a tight transition timeline, is a substantial undertaking. The work group was explicit: the state needs a high-quality summative system before it adds new components to that system. Teachers and administrators in focus groups said the same thing in different language. The first responsibility of the next procurement is to deliver a system that works, that reports results families can understand, that measures what it claims to measure across all student groups, and that arrives on time.

The item bank investment required in the Core tier is also foundational. Virginia cannot likely implement elements of the Enhanced or Transformative models (e.g., MST for EOC subjects), AI-assisted scoring at scale, or reliable growth scores without first establishing a high-quality, balanced item bank.

The Case for Building Toward Enhanced

The Core tier alone is not a sufficient long-term vision for Virginia's assessment system. The statutory individual student growth reporting requirements, the 45-day reporting deadline, and the legislature's direction on higher-order reasoning all point toward Enhanced tier features as the appropriate horizon. The work group endorsed growth measures, AI-assisted scoring with proper safeguards, and improved family reporting. Educators in focus groups emphasized the data lag as one of their most serious operational complaints. These are not aspirational preferences; they are documented needs with documented solutions in peer states.

Structuring the contract to reach Enhanced features through a phased approach within a single procurement reduces transition risk and maintains continuity of the student data record. The RFP should require vendor capacity to support Enhanced features and should establish pilot requirements and go/no-go decision points that VDOE controls.

On the Transformative Tier

The Transformative tier is not recommended for the current procurement cycle. The infrastructure that Virginia would need to implement competency-based pathways at scale, including statewide scorer training and moderation networks, does not currently exist and cannot be built within a single contract period. Federal approval, whether through IADA or a Section 8401 waiver, would also need to be secured. The work group's caution about teacher burdens related to locally developed performance assessments is well-founded and is consistent with the scaling challenges documented in every state that has attempted this model. None of this means that the Transformative tier is off the table for the future. However, the Transformative tier is simply not the next step after Enhanced. It represents a fundamentally different vision of how

student learning is defined and measured, and pursuing it requires not only the infrastructure described above but also a policy determination that a different approach is necessary. The work group and broader stakeholders have not expressed that determination; their consistent priority has been a high-quality summative system. If Virginia builds successfully through the Core and Enhanced tiers, if HB 1957 Subsections K and L are reenacted and operational, and if the state invests in the infrastructure that competency-based assessment requires and obtains federal approval (possibly with federal funding), a future procurement could incorporate Transformative features.

These recommendations align with and extend the work group's recommendations. The work group recommended a cost and value analysis of high school EOC assessments before reducing the number of those assessments. That analysis is a parallel workstream that VDOE should pursue, though it is upstream from procurement design, rather than being an RFP requirement. Regarding the growth construct, the work group expressed a clear preference for value tables and simpler interpretive models over complex value-added approaches. The growth model should prioritize educator interpretability. However, the specific construct decision should be made by VDOE before the RFP is issued, since that decision may drive the assessment design requirements that the vendor will be asked to meet. Section IV discusses the work group recommendations in more detail.

Feasibility and Cost-Effectiveness of Pilot Testing

Item 119 (A.3.c.2) specifically directs an evaluation of the feasibility and cost-effectiveness of pilot testing components of the proposed system before full implementation.

For the Core tier, no pilot testing is required. These features can be implemented operationally in the first contract year, other than the usual field testing that is required of newly developed items.

For the Enhanced tier, pilot testing is both feasible and necessary. MST for EOC subjects, AI-assisted scoring, and growth model implementation all require field testing with student data before operational deployment. Structuring these pilots within a single multiyear contract is the most cost-effective approach because it avoids a separate procurement cycle, leverages the item bank and platform investments already made in the Core tier, and maintains a single vendor relationship. The primary cost drivers for piloting are item bank expansion (to support MST routing), AI scoring validation studies across demographic subgroups, and growth model calibration. Based on comparator state pricing for selected-response item development, item development for MST-capable EOC pools could add a minimum of $500K per year above the Core tier's development baseline, depending on the number of subjects and target pool depth needed. (This figure assumes approximately 45–50 new items needed per EOC per year and an item cost of approximately $750/item). AI scoring validation and growth model calibration are more modest in scale but require specialized expertise typically priced at $200 to $250 per hour for psychometric and research services. These costs are incremental to the Core tier and should be specified as line items in the RFP, rather than left to vendor discretion.

For the Transformative tier, the pilot requirements are more extensive and less predictable in cost because they depend on federal approval processes and infrastructure that do not yet exist. Section VI.A addresses the proposed implementation timeline and key performance indicators (KPIs) for the transition.

Procurement Implications

The RFP should be structured to require full Core implementation as the baseline deliverable, with Enhanced features specified as required capabilities that the vendor needs to demonstrate and pilot on a defined schedule. The contract should include go/no-go checkpoints before Enhanced features move to operational use. Vendor evaluation criteria should weigh demonstrated experience with MST, AI-assisted scoring with fairness monitoring, and growth model implementation because Virginia will need those capabilities within the contract period. Section VIII addresses the recommended procurement structure in detail.

VI. Implementation Considerations

A. Transition Timeline and KPIs

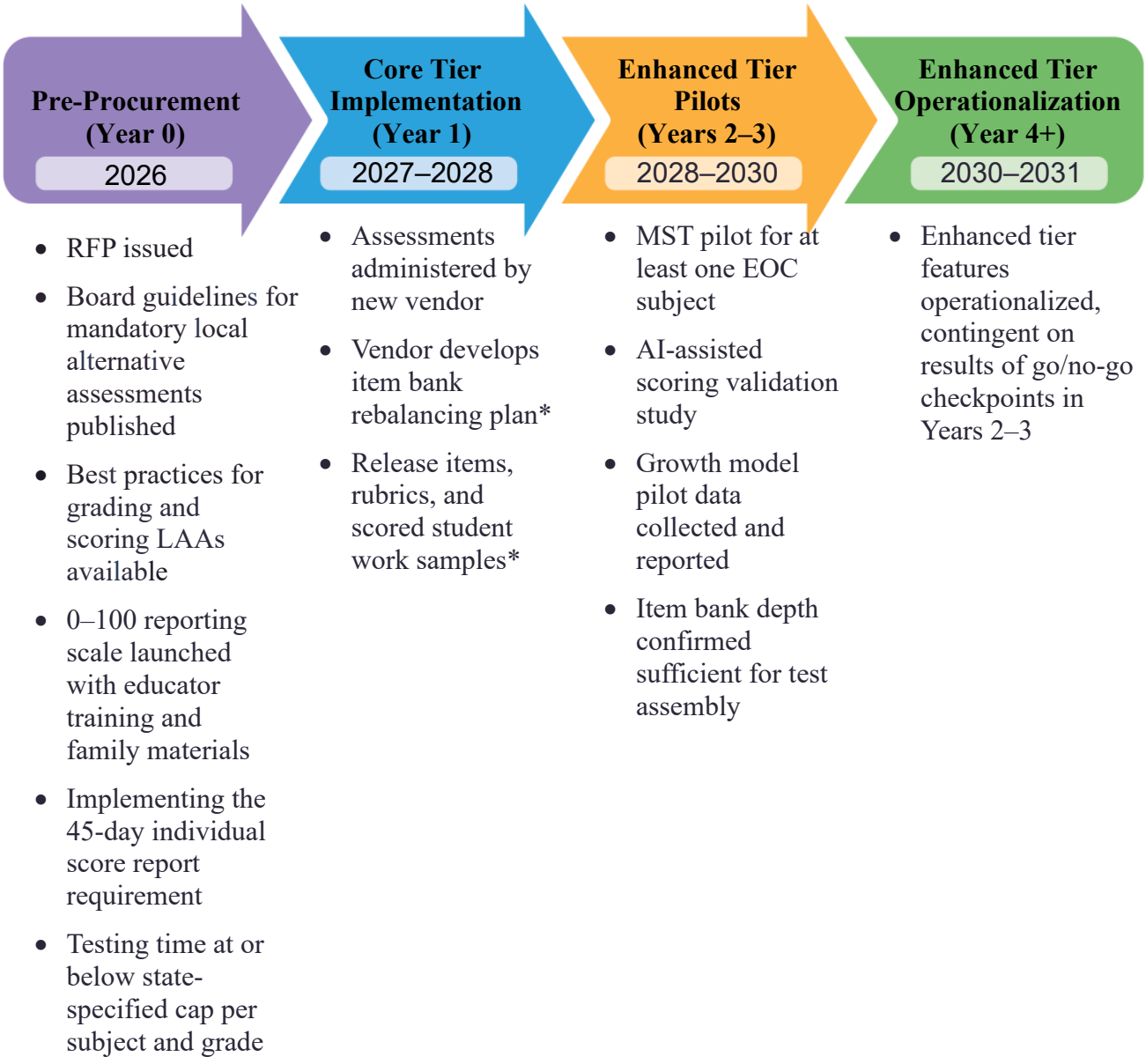
The transition from Virginia's current assessment system to the modernized design described in Section V is constrained by several fixed dates that together will determine the pace of implementation. The current vendor contract has been extended through December 2027, which provides a backstop but no margin for delay. SB 200 requires the Board to publish best practices for grading and scoring LAAs by September 1, 2026. HB 1957 requires that most provisions, including the 0–100 reporting scale and the 45-day ISR delivery timeline, take effect by July 1, 2026, meaning they will be implemented under the current contract. Item bank rebalancing, which is a prerequisite for the Enhanced tier's MST pilots, should begin in the first contract year to build sufficient depth by Years 2 and 3.

These constraints produce a four-phase timeline. The Pre-Procurement phase (2026) is when the critical policy decisions should be made: finalizing guidelines for mandatory LAAs, completing the ESSA compliance review, and issuing the RFP. Delaying any of these decisions pushes downstream milestones into the contract extension window and eliminates the time available for phased implementation. QIP's stakeholder engagement identified transition stability as a top priority. Division leaders and state staff reported that frequent changes without adequate lead time have created a perception of "building the airplane while flying it," undermining confidence in assessment data and preventing longitudinal comparisons. Stakeholders expressed strong preference for a structured, multiyear assessment cycle with a minimum development and testing period of 2–3 years before field implementation. These findings reinforce the phased approach recommended here and underscore the importance of communicating the transition timeline clearly and early.

Year 1 (2027–28) is the first operational year under the new vendor and should focus on the foundational requirements of delivering assessments without disruption, including item and blueprint portability from the prior vendor. Additional Year 1 priorities, such as deploying UDL decision tools and beginning the item bank development that the Enhanced tier depends on, may be achievable depending on the pace of vendor transition, but should not be assumed as guaranteed in the first operational year. Years 2 and 3 (2028–30) shift to Enhanced tier pilot testing, including MST for at least one EOC subject, AI-assisted scoring validation across all demographic subgroups, and collection of growth model pilot data. Each pilot feature is subject to go/no-go decision points before VDOE commits to operational deployment. In 2030–31, Enhanced tier features that meet pre-specified criteria for technical quality, fairness, cost, and educator readiness move to full operational use.

Figure 3 summarizes the major milestones across these four phases. Tables 6–9 provide measurable indicators, tied to specific statutory requirements or work group priorities, for tracking whether the transition is proceeding on schedule and producing the intended results.

Figure 3. Major Milestones Timeline



***Note:** These priorities may be achievable depending on the pace of vendor transition but should not be assumed as guaranteed in the first operational year. Any items not completed during the Core tier implementation should be carried over into Years 2–3.

KPIs

Tables 6–9 describing KPIs are organized by transition phase. They include both milestones (binary completion targets) and performance indicators (measures of quality or progress) tied to specific statutory requirements or work group priorities. Milestones confirm that a required

action occurred, and performance indicators assess whether the system is functioning as intended.

Table 6. Pre-Procurement Phase (2026)

KPI/Milestone	What It Measures	Statutory or Work Group Basis
Milestones		
RFP issued	Procurement launch occurs on schedule per the transition timeline in Section VI.A	Item 119 deliverable; Section VIII
ESSA compliance review completed	Federal requirements for state assessment systems verified and documented before procurement	Item 119 A.3.c.5; Appendix C
0–100 reporting scale implemented on all score reports	All student, school, and division score reports use the new scale; no legacy scale values appear	HB 1957 as amended by SB 200 (100-point reporting scale requirement)
45-day ISR delivery target met	Percentage of ISRs delivered within 45 calendar days of test completion	HB 1957 (45-day ISR requirement)
Testing time at or below state-specified cap per subject and grade	Actual median testing time by subject and grade compared to the maximum session length specified in the RFP	HB 1957 (assessment frequency and duration limits); work group priority
Performance indicators		
Growth construct defined by VDOE	VDOE reached a policy decision on how individual student growth will be defined, enabling the RFP to specify growth model requirements accordingly	Section V.D recommendation; work group priority
Board guidelines for mandatory LAAs published	Local school divisions have operational guidance with enough lead time to plan implementation for the 2027–28 school year	HB 1957 subdivision E.1.a(2) (Board guidelines); SB 200 enactment clause 4 (best practices due September 1, 2026)
Best practices for grading and scoring LAAs available	School divisions can train teachers on scoring expectations before local assessments go live	SB 200 enactment clause 4 (September 1, 2026, deadline)

Table 7. Year 1: Core Tier Implementation (2027–28)

KPI/Milestone	What It Measures	Statutory or Work Group Basis
System operations		
New vendor delivering assessments on schedule with no disruption to testing windows	Operational continuity confirmed; schools experience no gaps in test availability or score delivery during vendor transition	Work Group priority
0–100 reporting scale implemented on all score reports (continued under new vendor)	All student, school, and division score reports use the new scale; no legacy scale values appear	HB 1957 as amended by SB 200 (100-point reporting scale requirement)

KPI/Milestone	What It Measures	Statutory or Work Group Basis
45-day ISR delivery target met (continued under new vendor)	Percentage of ISRs delivered within 45 calendar days of test completion	HB 1957 (45-day ISR requirement)
Four performance levels reported on all assessments	All score reports include four labeled performance levels as recommended by the work group	Work Group recommendation
Testing time at or below state-specified cap per subject and grade (continued under new vendor)	Actual median testing time by subject and grade compared to the maximum session length specified in the RFP	HB 1957 (assessment frequency and duration limits); work group priority
Item bank and technical foundation		
Item bank rebalancing plan in place with year-over-year targets for Proficient and Advanced ranges*	Vendor has submitted a plan with measurable annual targets for increasing item coverage at upper performance levels; VDOE has approved the plan	Section V.B Core tier; prerequisite for MST
Accessibility and accommodations		
UDL decision tools deployed to all school divisions	All school divisions have access to the decision-making tool for selecting appropriate accommodations; usage data are being collected	Work Group priority; HB 1600 focus area (c)
Accommodation utilization rates tracked and disaggregated by disability category and English learner status	Baseline data collected for ongoing monitoring of whether students who need accommodations are receiving them	HB 1600 focus area (c); ESSA peer review requirements
Transparency and communication		
Released items, rubrics, and scored student work samples published*	Materials publicly available for each assessed subject and grade within 90 days of score release	Work Group priority; HB 1600 focus area (g)
Assessment literacy materials distributed to families*	Materials available in the languages and formats specified in the communication plan (Section VI.D); distribution confirmed at school level	HB 1957 (interpretation guidance requirement in ISRs)
Educator training completion rate for score report interpretation*	Percentage of teachers and school administrators who have completed training on interpreting and using new score reports	Work Group priority (assessment literacy); Section VI.D
Local alternative assessments		
Mandatory LAAs administered in all required subject areas	All school divisions have implemented local assessments in the subjects required under the Standards of Quality and as specified in HB 1957; participation data reported to VDOE	HB 1957 subdivision E.1.a(2)

KPI/Milestone	What It Measures	Statutory or Work Group Basis
Cost**		
Actual contract costs within projected range from Section VI.B cost estimates	Year 1 vendor costs compared to the cost projections in the feasibility study; variance calculated and reported to VDOE and legislative committees	Item 119 A.3.c.2 (feasibility and cost-effectiveness); Section VI.B
Per–test administration cost calculated as baseline	Total contract cost divided by total number of test administrations across all windows (including retests), establishing the baseline for the ongoing cost tracking indicator (see Table 9)	Item 119 A.3.c.2 (feasibility and cost-effectiveness); Section V.D

***Note on implementation timeline:** The priority in Year 1 will be on the transition to a new test vendor and test delivery. Additional items noted in this table may be achievable but are not guaranteed. Any items not completed during the Core tier implementation should be carried over into Years 2–3.

****Note on cost metrics:** Virginia's retest volume (540,000+ EOC administrations in spring 2025 alone, plus fall and summer windows) means that per-pupil cost figures understate the true unit cost of service delivery. Both per-pupil and per-test-administration metrics should be reported to VDOE and to legislative committees. The per-pupil metric supports cross-state comparisons; the per–test administration metric captures the impact of retest policy on unit economics.

Table 8. Years 2–3: Enhanced Tier Pilots (2028–30)

KPI/Milestone	What It Measures	Statutory or Work Group Basis
MST pilot launched for at least one EOC subject	Pilot is operational with a defined sample of schools; student-level routing data are being collected for analysis	Section V.B Enhanced tier
AI-assisted scoring validation study completed for all demographic subgroups	Differential scoring accuracy evaluated across race, gender, disability status, and English learner status; results reviewed by VDOE and TAC before operational use	Section V.B Enhanced tier; Gregg et al. (2021) findings
Growth model pilot data collected and reported to VDOE	At least 1 year of individual student growth scores calculated under the growth construct; results compared to existing growth proxies	SB 200 (individual student growth scores); work group priority
Go/no-go decision points met for each Enhanced feature	Each pilot feature evaluated against pre-specified criteria (technical quality, fairness, cost, educator readiness) before VDOE decides whether to proceed to operational deployment	Section V.D recommendation
Item bank depth sufficient to support MST routing in pilot subjects	Item pool meets minimum size and content coverage thresholds defined in the vendor contract for adaptive test assembly	Section V.B Enhanced tier

Table 9. Ongoing Indicators

KPI/Milestone	What It Measures	Statutory or Work Group Basis
Annual Board audit of mandatory LAAs completed (beginning Year 1)	Board reviews quality and consistency of local assessments across school divisions annually; governance framework is functioning	HB 1957 subdivision K 2 (if reenacted)
Minimum 4-year revision cycle maintained	No major changes to assessment design, scoring, or reporting framework within the first 4 years of operational use except to correct identified problems	Work Group priority
Division-level participation rates meet ESSA thresholds	95% participation rate met at division and subgroup levels; any school divisions falling below threshold identified and supported	ESSA (95% participation requirement); HB 1957 (opt-out provision; note potential tension between state opt-out allowance and federal threshold)
Disaggregated results published on the same date as individual scores	School-level, division-level, and state-level disaggregated results released simultaneously with student reports; no delayed reporting for subgroups	HB 1957 (simultaneous public release requirement for statewide results)
Educator and administrator satisfaction with the assessment system surveyed annually (beginning Year 1)	Teachers', school test coordinators', and division assessment directors' perceptions of the quality, usability, and instructional usefulness of the system	Work Group emphasis on assessment literacy and educator voice; Section VI.D
Cumulative per-pupil and per–test administration cost tracked and compared to baseline	Annual reporting of total assessment system cost divided by (a) number of tested students and (b) total test administrations across all windows, including retests, compared to the pre-transition baseline and projected costs in Section VI.B	Item 119 A.3.c.2 (feasibility and cost-effectiveness); Section VI.B

Note: Indicators apply across all phases. Start dates are noted where relevant.

B. Funding Requirements

We estimate that costs for a scope similar to the current Virginia contract may range between $30M and $38M per year, with about 60–70% related to test administration and the remainder for test development program management and reporting. This estimate is based on a comparison of Virginia's current assessment scope to recent contracts in other states and to nonbinding cost estimates from four vendors who responded to Virginia's 2024 RFI.

- Tennessee (approximately $28M/year) has a relatively similar scope, with a mix of summative and alternate assessments under a single contract. However, Virginia has more EOC tests (13 vs. 10), and (until July 1, 2026) includes growth assessments in reading and mathematics, which Tennessee does not. These additions would push Virginia's likely future costs above the Tennessee baseline.

- Massachusetts (approximately $37.9M/year including an alternate assessment) has a system with fewer tests than Virginia, but its contract includes some cost drivers that Virginia may also face, including translated assessments and ongoing item development to replace released items. This provides an approximate upper bound for what an enhanced Virginia system might cost, based on current discussion.
- Florida and Texas offer useful reference points but are harder to compare directly to Virginia. Florida's base contract ($26.5M/year) includes a similar Grades 3–8 and high school scope, but also includes K–2, and serves a far larger student population, which may act to reduce unit testing costs. Florida amended this contract to incorporate additional item development and other tests, with an amended value of about $49M/year, which suggests that Virginia's costs might be expected to fall somewhere in a wide range depending on scope. Texas ($124M/year across two contracts) covers a much larger scope in terms of assessments administered. A rough estimate derived from Texas suggests that Virginia's costs could be around $40M, though the calculation is highly approximate.
- Four vendors responded to the RFI. The median of vendors' low-end estimates was $23.5M; the median of their high-end estimates was $31M. One vendor's estimates were significantly higher than the other three. Excluding that vendor, the low- and high-end medians were $21M and $30M, respectively. Several factors limit the ability to accurately estimate costs based on other states' contracts.
- State-level contract information on assessments is difficult to compare directly. States include different grade levels, subjects, tasks, and deliverables in their contracts. Number of students to be tested varies, as can makeup, retest, and accommodation policies.
- Pricing information is often not available or not transparent.
- In particular, contracts may differ in how they handle costs for item and test development. Costs may be spread over years or calculated within a given year, which can greatly affect annual costs, depending on assessment activities at a given moment in time.
- Contract structure could also affect total cost: multiple assessment contracts may drive up program management and coordination expenses, while a single large contract can include pass-through costs or reduce transparency in bundled pricing.

Further, Virginia has not yet made final decisions on what will be included in an assessment contract. We therefore assumed a scope similar to the current scope for comparison purposes. However, changes in the following areas could significantly affect these estimates:

- ***EOC retesting.*** Virginia administers EOC tests in multiple administrations every year. The fall and summer windows serve students who are retesting for graduation, enrolled in block-scheduled courses, or attending summer school. The spring 2025 administration alone included more than 540,000 EOC test administrations. Fall and summer retest volumes are not reported in the technical report (Virginia Department of Education, n.d.-b) but represent a meaningful addition to total test volume. Each retest requires test delivery, scoring, and reporting. The comparator states may not have equivalent retest structures, which means their contract costs may understate the costs of what Virginia needs.

- ***Item development intensity.*** The RFI vendor estimates show a wide range for item development costs, from $2.3M per year (minimal development, 10 items per grade/subject) to $9.6M per year (200 items per grade/subject, up to five forms). Virginia's needs will depend on item release policies, the pace of standards revision, and the desired size of the CAT item pools. This single line item could shift annual costs by millions of dollars. Detailed pricing from Texas and Florida offers some useful information on test development and other project activities. For example, for academic assessments, per-item development costs range from $749 for new mathematics selected-response items, to $1,004 for ELA performance task items, to $1,239 for new TEIs, to $1,345 for alternate assessment items. Passage costs are $4,175 per passage. Florida's pricing estimate for performance tasks is about $11,500 per task.
- ***Alternate assessment.*** Whether the RFI vendor estimates include Virginia's alternate assessment (VAAP) is not clear. The current Virginia contract covers administration, scoring, and reporting for 17 alternate assessments, with item development handled through a separate $1.6M contract. If the RFI estimates excluded alternate assessment, the actual cost would be higher than reported. Virginia may wish to explore participation in an alternate assessment consortium such as Dynamic Learning Maps (DLM) (n.d.). This consortium manages all aspects of alternate assessment design and delivery across nearly half the country (24 states). Based on other state costs, we estimate that participation might cost $1.5–$2M per year. Benefits of participation would include shared costs for peer review and documentation and likely lower administration burden on VDOE staff.
- ***ELP assessment.*** These costs do not include ELP assessment, which is currently managed through Virginia's participation in the WIDA consortium.
- ***New or expanded scope.*** Additions such as performance tasks, translated assessments, new subjects, or expanded scoring and reporting would increase costs beyond the current scope estimate. For reference, the only RFI vendor to provide detailed assumptions on translation costs estimated test translation at $165K. Massachusetts's contract includes translation and a student survey and Tennessee's includes formative assessment platforms and a forensic analysis of test results. Virginia's final scope decisions will be the largest determinant of where costs fall within or above the estimated range.

Non-Contract Costs

The cost estimates in the preceding section reflect only contract costs. It is important to note that, with enhanced assessment needs, Virginia may also require additional internal capacity to support a robust assessment system. A well-resourced state office reduces dependence on vendors for functions that are better handled in-house, improves the state's ability to manage contracts effectively, and can lower overall program costs over time by shifting certain responsibilities out of vendors' scopes. Specific areas where additional capacity would have the greatest impact include psychometrics and content oversight. A data analyst, a communications specialist, and one or two additional content staff might cost a total of $400K to $600K per year in salary and benefits. That is roughly 1–2% of the estimated contract cost, and it would likely improve VDOE's ability to manage both the procurement process and the ongoing vendor relationship.

- A full-time state data analyst would give VDOE independent capacity to review vendor data files, evaluate score distributions and trends, and respond to technical questions from the field without relying on the contractor for every request. Several comparator states maintain in-house research staff. This position would also strengthen VDOE's hand during contract negotiations by ensuring that the state can evaluate technical proposals on their merits.
- Assessment results reach parents, educators, legislators, and the public. Dedicated communications staff, either within VDOE's assessment office or elsewhere in the agency, would ensure that public-facing content is developed with consistent messaging in accordance with agency timelines. Alternatively, communications capacity could be incorporated as a requirement in the assessment RFP or a separate procurement. This is especially important during high-visibility moments such as the release of statewide results or the introduction of new assessments aligned to revised standards. The new mathematics (2023) and English (2024) standards, along with upcoming new cut scores, make this an especially good time to invest in this area.
- If Virginia moves toward increased item development to support item release, maintain robust CAT pools, or develop new item types such as additional performance tasks, additional content specialists at VDOE would be needed to manage the volume of review required. Every new item goes through internal review at VDOE before external committee review. More items mean more review cycles, more committee meetings, and more QC. Without adequate staffing, state-level review becomes a bottleneck, which can delay development timelines and increase costs. For context, the RFI vendor estimates showed that moving from minimal development (10 items per grade/subject) to intensive development (200 items per grade/subject) increases item development costs from $2.3M to $9.6M. That level of development would require proportionally more state oversight.

C. Professional Development and Field Training Plan

Through our stakeholder engagement efforts, educators and administrators repeatedly asked for greater and more consistent communication and information from VDOE. Stakeholder engagement conducted by QIP reinforced this finding. Across 13 focus groups with 65 participants statewide, educators and parents reported persistent misconceptions about SOL assessments, including beliefs that tests are solely multiple choice and confusion about how computer-adaptive tests function. Teachers described a lack of predictable test design documentation as a barrier to aligning instruction with state expectations. Parents and community organizations reported a "data lag" in which score reports arrive too late to inform summer interventions or the following year's placement decisions. These communication gaps are addressable, but they require sustained, audience-specific outreach rather than one-time announcements. (See Appendix D for complete stakeholder engagement findings.)

As VDOE releases changes to the assessment system, the agency should develop and implement a comprehensive technical assistance plan, whether through the assessment RFP, a separate contract, or a combination of both, that gives educators and administrators the knowledge and tools needed to implement the new assessment system. The training team should bring subject-matter expertise to develop a cohesive, scalable approach to training that is specifically targeted to key audiences across the Commonwealth.

Outreach and dissemination strategies should be tailored to effectively reach a wide range of stakeholders, including division staff, school and district administrators, educators, families, and students. QIP's engagement sessions identified specific audience needs that should shape the communication strategy. Teachers want structured design documentation, released items, and training on new item formats. Parents want visually intuitive reports that show how their child performs relative to peers and growth over time, with translations available in multiple languages. Division leaders want firm release dates for instructional tools and crosswalks, which are frequently labeled "forthcoming" but delivered after testing windows close. Assessment coordinators want reduced back-end administrative burden in platform management. VDOE or its vendor should ensure that communications are strategically timed to align with the academic year, ensuring that each audience receives relevant, timely, and actionable information. For example, messaging about item types and recently released items should be targeted in the winter months, before educators begin preparing for end-of-year assessment.

Technical assistance should be delivered through a coordinated suite of communication products to ensure the broadest possible reach. The strategy should include self-paced online training modules, live and on-demand webinars, short-form instructional videos, visually engaging infographics, and targeted email communications. Multiple formats will reinforce key messages across multiple channels to increase retention of key messages. All materials should be designed using plain-language principles and best practices in adult learning, to ensure clarity and accessibility.

Training should be focused on critical topics, such as released assessment items, new item types, updates to scoring scales, and any changes in reporting. Each topic should be broken down into digestible, actionable components, with examples and practical guidance to support implementation in real-world educational settings. VDOE should ensure that materials remain current by providing resources to update existing materials as policies and practices evolve.

D. Communication and Change Management Strategy

Virginia is introducing several simultaneous changes to its assessment system: a new vendor, a new reporting scale, standardized four-level performance categories across all assessments, new LAA requirements, and, potentially, new item types and score report formats. These changes will overlap with already-in-progress implementation of increased cut scores and a higher proficiency bar, scheduled to take effect through 2029–30. Each of these changes individually would require a communication effort. VDOE should plan for significant communications and engagement as part of implementation. The work group identified communication as a recurring theme across its recommendation categories. Twenty-two of 23 members present at the March 2026 meeting voted to prioritize modernizing data systems and reporting. Twelve voted for transparency through released items and rubrics. The work group's assessment literacy recommendation called for a comprehensive educational effort, using plain language, directed at policymakers, school boards, educators, and families. Focus group participants reinforced these priorities: family members described the current reporting system as a "black hole," and educators flagged the existing two-score system (vertical and horizontal scale scores) as confusing for parents, students, and teachers alike.

This section outlines the communication and change management actions required to support a successful transition. It is not a comprehensive communications plan; VDOE's communications

team will need to develop detailed messaging, materials, and distribution strategies. The following sections identify the priorities, audiences, and sequencing that should shape that work.

New Reporting Scale

SB 200's change from "scored" to "reported" on a 100-point scale means the underlying psychometric scale does not need to change. Virginia's planned approach is to map the existing scale to a 0–100 reporting metric. This avoids the cost and timeline of a full rescaling (such as South Carolina's), but it means every communication should make clear that the 0–100 number is not a percentage, not a grade, and not directly comparable to any other 0–100 metric that families encounter.

Four actions are required before rollout:

1. Develop and test "This is NOT a grade" messaging for the 0–100 reporting scale in parent focus groups, with particular attention to families with limited English proficiency, families unfamiliar with the state testing system, and families of English learners. The messaging should be tested in the languages most commonly spoken across school divisions and should explicitly address the fact that the assessment score will eventually count as part of a course grade once the 10% requirement takes effect (see Section II.B). Proactive messaging should be developed well in advance of that effective date.
2. Create one-page "What to Expect" guides addressing the test format, what the 0–100 scale means, what the four performance levels mean, and why students may see test content they have not yet been taught. These guides should be distributed to teachers before the testing window; sent home with students; and available in print, online, and in the most commonly spoken community languages.
3. Establish multiple distribution channels for all communication materials: online, through school mailings, and as print copies sent home with students. Families without internet access must receive information through mail or in person at school events.
4. Train an ambassador cohort of teachers and administrators to lead peer-to-peer explanation before reporting cycles, as described in the following section.

Educator Training and Ambassador Networks

Teachers are the primary interpreters of assessment results for families. If teachers cannot explain what the scores mean, no amount of state-level communication will compensate.

VDOE should train an ambassador cohort of teachers and administrators before score release. Ambassadors should represent schools across a range of community contexts, including rural school divisions and high-poverty schools. The ambassador model draws on NWEA's approach to MAP implementation in which peer-to-peer explanation from trained teachers proved more effective than top-down communication from the state.

Ambassador functions should include leading peer explanation sessions within their schools, providing feedback to VDOE on how materials are being received in schools, and serving as a rapid-response resource during the score release. VDOE should deploy ambassadors for regional information sessions before scores are released, and equip them with FAQ documents, sample score reports with annotations, and talking points for common parent questions.

Educator training should cover, at a minimum, how to interpret the 0–100 reporting scale and how it differs from a grade; how to explain CAT to students and families; how to read and use the new four-level performance reports; and how growth data, if reported, relate to proficiency classification. Training completion rates should be tracked as a KPI (see Section VI.A).

Audience-Specific Communication

Different audiences need different information at different levels of detail.

Families need to understand what their child's score means, whether their child is on track, and what they can do to support learning. They do not need legislative history, psychometric terminology, or technical documentation. Score reports should lead with the performance level and a plain-language description of what it means (e.g., "your child understands 5th-grade content" rather than "your child scored at the Proficient level on a criterion-referenced assessment"). Alaska's interactive Achievement Level Explorer Tool (see the Focus Area G section of Appendix B) is a useful model for letting families see what performance at each level looks like in practice. Score reports must include the interpretation guidance required by HB 1957, and that guidance should be written and tested for a nontechnical audience.

Educators need instructional detail: where a student is strong, where the student needs support, and how the results compare to prior performance. Audience-specific report formats should separate the teacher view (e.g., item-level detail, strand-level performance, growth data) from the family view (e.g., performance level, growth trajectory, next steps). The work group recommended that VDOE follow Massachusetts's MCAS model of publishing annotated student work samples at each performance level, which serves both transparency and instructional planning (see the Focus Areas F and G sections of Appendix B).

Communications should also help educators distinguish between the purposes of different assessments, including the role of formative assessment in providing day-to-day instructional feedback, so that the summative system is understood as one component of a broader picture of student learning.

School board members and local administrators need to be equipped to answer constituent questions without relying on VDOE staff. They should receive proactive briefings before each score release, with talking points, anticipated questions, and context for their division's results. Board members who are blindsided by parent complaints about "failing" scores will generate political pressure that VDOE then has to manage reactively.

Legislators and legislative staff need a clear, one-page summary of what changed, why, and what the scores mean at the system level. They do not need the same materials as families or educators. The executive summary of this report and the Board of Education presentation serve this function for the initial transition; ongoing legislative communication should be built into VDOE's annual reporting cycle.

What VDOE Can Stop Emphasizing

Effective communication is as much about what to remove as what to add. VDOE's current parent-facing materials could be restructured to lead with what scores mean for students and families, rather than with technical terminology or legislative context. Score reports and landing pages are most effective when they prioritize plain-language explanations of student performance and next steps.

VDOE should move technical language to appendices and technical documentation. Parent-facing materials should lead with what scores mean instructionally. The shift to standardized four-level performance reporting across all assessments is an opportunity to change the framing entirely, from "Did my child pass?" to "What is my child ready to learn next?"

Initial Reporting Under the Revised Scale

The first score release under the revised reporting scale will set the tone for public perception. VDOE should plan for it as a managed communication event, not a routine data release.

VDOE should monitor social media and district communications during and after the first release, using a rapid-response capacity to correct misinterpretations before they spread. Board members should be briefed proactively. Ambassadors should be deployed for peer-to-peer explanation in every region before scores are released.

Ongoing Communication and Evaluation

After the first operational year, VDOE should evaluate the communication strategy against measurable outcomes such as decline in help-desk inquiries about "failing" scores, teacher-led explanations replacing VDOE-led defenses, growth language appearing in district communications, and fewer board-level complaints about grading confusion. These outcomes are reflected in the KPI framework in Section VI.A (educator satisfaction survey, educator training completion rates).

Priorities for the period following the first operational year include publishing annotated student work samples, developing audience-specific report formats with separate teacher and family views, and evaluating rapid reporting infrastructure. Florida's FAST Family Portal (see the Focus Area I section in Appendix B), which provides scores within 24 hours, represents an aspirational long-term target for Virginia.

Communication for Alternate Assessment (VAAP)

Families of students with significant cognitive disabilities need clear information about how the VAAP will be affected by the vendor transition, any changes to score reporting, and the standardization of performance levels. Whether VAAP is bundled into the new contract or maintained as a separate procurement (see Section V.B), VDOE should ensure that communication to VAAP families addresses their specific questions and does not assume familiarity with the general assessment changes. If Virginia pursues consortium participation through DLM, the shift in assessment format and reporting would require its own dedicated communication effort.

Communication for Local Alternative Assessments (LAAs)

HB 1957 and SB 200 defined and clarified requirements for LAAs, including new distinctions between mandatory and permissive LAAs, cross-teacher grading for verified credits, and completion in supervised, secure settings. These are new operational requirements for school divisions, and families will need clear information about what LAAs are, how they differ from statewide SOL assessments, and what role they play in their child's grades and graduation requirements. VDOE should develop plain-language guidance for divisions to use in communicating with families about LAA options and should coordinate this effort with the

Board's development of best practices for grading and scoring LAAs, required by SB 200 by September 1, 2026.

Sequencing and Responsibility

Communication and change management cannot be treated as an afterthought to procurement. The following four phases should guide the sequencing of communication activities, in coordination with the transition timeline in Section VI.A.

1. ***Pre-procurement:*** Establish internal communications capacity at VDOE, develop and test the "This is NOT a grade" messaging for the 0–100 scale with focus groups, recruit and begin training the ambassador cohort, and brief school board members and legislative staff on what is changing and why.
2. ***Core tier implementation (Year 1):*** Distribute "What to Expect" guides, deploy ambassadors for regional information sessions, publish audience-specific materials in multiple languages, and prepare score report prototypes for educator and family review.
3. ***Enhanced tier pilots (Years 2–3):*** Manage the score release as a planned communication event with rapid-response monitoring, proactive board briefings, and ambassador deployment in every region. Track KPIs including help-desk inquiry volume, training completion rates, and educator satisfaction.
4. ***Enhanced tier operationalization (Year 4 and beyond):*** Evaluate communication effectiveness against the aforementioned measurable outcomes, publish annotated student work samples, and transition from state-led explanation to teacher-led and district-led communication as the system matures.

VDOE may wish to include communication deliverables as a scored element in the RFP (see Section VIII.B), and VDOE should invest in dedicated communications capacity—within the assessment office or through a separate procurement—to ensure consistent messaging across the transition. The Non-Contract Costs section (Section VI.B) includes a recommendation for communications staffing as part of VDOE's internal capacity investment. The work group consistently identified communication as a priority across its recommendations, with 95% of members voting to prioritize modernizing data systems and reporting at the March 2026 meeting.

E. Legislative and Policy Considerations

Several legislative and policy questions remain unresolved as of this report's publication. Each has direct implications for procurement scope, implementation timeline, or system design. Table 10 identifies open issues and considerations that affect implementation planning.

Table 10. Open Legislative and Policy Issues and Considerations

Issue	Consideration	Deadline or Dependency
HB 1957 Subsections K and L reenactment	Subsections K and L were adopted as a first enactment and require reenactment by the 2027 General Assembly (SB 200, enactment clause 3). Subsection K would require the Board to make previous years' SOL assessments available as practice tools and to perform annual audits of mandatory and permissive LAAs for quality and consistency.	2027 legislative session for reenactment. Cost estimate to General Assembly due October 1, 2026.

Issue	Consideration	Deadline or Dependency
	Subsection L would require assessments to include items requiring application of knowledge, critical thinking, and logic, including open-ended questions and long-form writing. If these subsections are not reenacted, the governance infrastructure supporting the Transformative tier does not take effect. SB 200 also directs VDOE, to the extent permitted by federal law, to include these provisions as add-alternate components in the assessment RFP and to report their estimated costs to the General Assembly by October 1, 2026 (enactment clause 5).	
EOC assessment configuration	Virginia currently administers 13 EOC assessments across three windows, with multiple retake opportunities. SB 200 modified the EOC timing window and added exceptions for graduating seniors, retakes, and multiday assessments. Whether to reduce the number of EOCs, cap retakes, or modify the testing calendar affects item bank requirements, contract scope, and cost. The work group recommended a cost and value analysis.	Ideally a decision is needed before the RFP scope is finalized.
0–100 reporting scale implementation approach	SB 200 changed the HB 1957 requirement from "scored" to "reported" on a 100-point scale, preserving technical scoring flexibility. Virginia plans to map the existing scale to a 0–100 reporting metric rather than conducting a new standard setting. This approach avoids additional cost but creates a derivative metric that requires proactive communication. The choice between mapping and rescaling is a policy decision with psychometric and communication tradeoffs.	Communication materials should be developed and tested in advance of the first score release under the new system (see Section VI.D).
Delayed statutory provisions (subdivisions E 4 b and E 4 c)	SB 200 delays two provisions of HB 1957 until the beginning of the second full school year after pilot implementation of the new assessment system (enactment clause 2). The delayed provisions are: (1) the prohibition on additional EOC assessments beyond the required EOC SOL assessment or permissive LAA in any subject area (E 4 b), and (2) the requirement that each student's assessment score count for at least 10% of the final course grade (E 4 c). The 10% requirement interacts with local grading policies and will require Board guidance, particularly where school divisions use standards-based grading systems. VDOE must certify in writing to the Virginia Code Commission the date on which the pilot implementation condition is met.	Board guidance on grading policy interaction needed in advance of the second full school year after the new assessment system is piloted.
External scoring for verified-credit LAAs	SB 200 requires that teachers may not grade their own students' LAAs when used for verified credits. Small and rural school divisions with one or two	Board best practices due September 1, 2026. Cross-division scoring

Issue	Consideration	Deadline or Dependency
	teachers per subject area will need cross-division scoring coordination that does not currently exist. The Board must develop and make available best practices for grading and scoring both mandatory and permissive LAAs (SB 200, enactment clause 4).	coordination operational by 2027–28.
Federal flexibility window	The current federal administration has signaled support for state assessment flexibility, including Section 8401 waivers. If Virginia intends to pursue federal approval for Transformative tier features, the application timeline is relevant because the policy window may not remain open indefinitely. The Enhanced tier does not require federal approval.	Application timing depends on Virginia's needs.

F. Risk Mitigation

The transition to a modernized assessment system carries risks across the areas of procurement, technical design, communication, capacity, and legislative and federal compliance. Many of these risks are documented in the tier descriptions (Section V.B) and the communication strategy (Section VI.D). Table 11 consolidates them into a single reference so that VDOE and legislative stakeholders can assess exposure and track mitigation actions.

Table 11. Risks and Mitigation Strategies

Risk	Mitigation Strategy
Vendor transition disrupts 2027–28 testing	Contract extension secured through December 2027 provides a backstop. The RFP should require a detailed transition plan scored as an evaluation criterion.
Item bank insufficient for new standards	The current bank is concentrated at lower performance levels. The Core tier requires item development targets for the Proficient and Advanced ranges, specified as Year 1 contract deliverables with measurable annual benchmarks (see Table 7 in Section VI.A).
0–100 scale interpreted as percent correct	Mitigation should be through the proactive communication strategy described in Section VI.D, including the "This is NOT a grade" messaging, an educator ambassador cohort, and audience-tested materials in multiple languages.
AI scoring bias against English learners	Research cited in this report (Gregg et al., 2021; Beiting-Parrish & Whitmer, 2023) identifies English learner status as the highest risk demographic for automated scoring inaccuracy. The RFP should require disaggregated fairness validation for all demographic subgroups, with reporting specific to English learners.
VDOE internal capacity to manage vendors	A state data analyst, additional content staff, and communications capacity would likely improve VDOE's ability to oversee the contract independently and manage costs. Estimated cost: $400K to $600K per year (see Section VI.B).
Teacher burden from LAAs	HB 1957 LAA requirements create new scoring and administration demands on school divisions. The external scoring requirement in SB 200 compounds the burden for small and rural school divisions. Mitigation includes Board guidance by September 2026, cross-division scoring coordination, and phased

Risk	Mitigation Strategy
	implementation.
MST item pool depth insufficient for Enhanced tier	Enhanced tier MST routing requires sufficient items per content area. If Core tier item development targets are not met, Enhanced features cannot launch on schedule. Mitigation includes go/no-go checkpoints in the contract, and item pool depth specified as a prerequisite for the MST pilot.
Subsections K and L not reenacted in 2027	If Subsections K and L are not reenacted, the audit and prior-year release provisions do not take effect. This limits, but does not eliminate, the path to Transformative tier features. The Core and Enhanced tiers do not depend on these subsections.
Federal flexibility window closes	The current federal administration supports assessment waivers. A future administration may not. If Virginia wants IADA or Section 8401 authority, the application window matters. The Enhanced tier does not require federal approval; the Transformative timeline accounts for a multiyear application process.

VII. Emerging Technologies and Future Trends

Several technologies that are now receiving attention in education policy discussions have the potential to reshape how states design, deliver, and score assessments. Some of these technologies are already operational at scale. Others remain in early development despite generating considerable interest. This section distinguishes among what is proven, what is promising, and what is premature, so that Virginia's planning is grounded in evidence.

A. Current Innovations

Automated Scoring of Constructed Responses

Automated scoring of student writing and short constructed responses is the most mature AI application in state assessment. More than a dozen states, including Virginia, already use automated scoring engines in their operational summative programs. Pearson's automated scoring system currently scores constructed responses on Virginia's SOL assessments.

The technology works through a pipeline: student responses are preprocessed, flagged for anomalies, analyzed for features (grammar, usage, organization, content), and scored against rubric dimensions using trained statistical models. These systems are not generative AI systems. They use traditional machine learning and natural language processing approaches with more than two decades of validation.

Performance benchmarks from large-scale deployments of current automated scoring engines are strong. In Texas, where a 2023 redesign of the STAAR assessment increased constructed-response items six- to sevenfold, automated scoring engines matched or exceeded human-to-human agreement rates across all operational items during spring 2024 (Texas Education Agency, 2023). The hybrid model, in which machines score all responses and at least 25% are independently verified by human raters, has produced substantial cost savings. Across the Smarter Balanced consortium (more than a dozen states), 60% to 80% of constructed responses are scored by automated scoring engines, with the remainder routed to human scorers, based on confidence thresholds.

Fairness evidence is generally supportive but raises one significant concern. Results from the 2022 NAEP Reading Automated Scoring Challenge found that English learner status accounted for half of all biased scoring results, exceeding race, ethnicity, and gender as a source of differential performance (Beiting-Parrish & Whitmer, 2023). A study of the Iowa Statewide Assessment of Student Progress found that human scoring tended to favor English learners more than automated scoring did (Vo et al., 2023). Given Virginia's growing English learner population, any expansion of automated scoring for new performance tasks and constructed-response items should include dedicated fairness monitoring for this group.

Virginia should also request detailed technical documentation from its vendor on the scope, validation evidence, and fairness performance of automated scoring currently in use on SOL assessments. Publicly available documentation on specifics is limited.

Adaptive Test Delivery

Computer-adaptive testing is a well-established technology. Virginia was among the first states to deploy item-level CAT for its stand-alone state assessment program (not as part of a

consortium). Virginia's SOL program uses item-level CAT for Grades 3–8 in mathematics and reading, anchored by IRT-based item selection algorithms.

The active area of growth is MST, which presents items in preassembled testlets rather than one at a time. Maryland implemented a 1-2-2 MST design for ELA in Grades 6–8 and Grade 10 in spring 2024 and expanded to Grades 3–5 in spring 2026. The Digital SAT uses section-level adaptation. WIDA ACCESS employs MST for ELP assessment across more than 1.4 million students annually.

MST offers a practical advantage for content areas where items are tied to shared reading passages or stimulus sets: it preserves passage integrity while still adapting to student ability. For Virginia, this is relevant to ELA assessments and to the new IRW components. The draft procurement guidance already developed as part of this project (see Section VIII) anticipates that offerers may propose MST designs for stimulus-heavy content.

No operational K–12 assessment program uses AI or machine learning for real-time item selection. All deployed systems rely on standard IRT-based algorithms with content-balancing constraints.

Technology-Enhanced Items

TEIs are now standard across virtually all state summative programs. Virginia's SOL assessments include drag-and-drop, multi-select, hot text, constructed-response, and equation entry items. The Smarter Balanced consortium offers the most extensive catalog of other TEI types that Virginia could consider for the next procurement cycle, including evidence-based selected response, graphing grids, match interactions, and table entry formats.

Most operational TEIs use digital tools to present familiar item types in new ways. For example, drag-and-drop categorization items are functionally equivalent to multi-select items in what they measure. True simulation-based items, where students conduct virtual experiments or interact with complex scenarios, originated in NAEP's Science and Technology and Engineering Literacy assessments. Massachusetts is the only state to develop computer-simulated science tasks for use in a statewide assessment program. These tasks, developed under IADA authority, were piloted from 2021 through 2023 and are currently being field-tested alongside the existing MCAS STE test for Grades 5 and 8, with new operational STE tests anticipated for spring 2027. Massachusetts is currently the leading example for any state considering similar approaches.

As Virginia considers expanding item types to include performance tasks, AI-assisted scoring will be an important consideration. AI-assisted scoring uses a model in which automated engines produce initial scores that are then reviewed, verified, or overridden by trained human raters. This is distinct from fully automated scoring in which machine-generated scores stand as final without human review; no state currently operates such a model. The two models carry different implications for legal liability, public trust, and the evidentiary standard required to defend scores in an appeals process. Virginia's approach should be clearly defined and documented before performance task scores are incorporated into the accountability system.

Research has shown that complex TEIs can introduce construct-irrelevant variance related to computer familiarity, socioeconomic status, and disability status (e.g., Bridgeman, 2009; Strain-Seymour et al., 2013). A 2023 study of measurement invariance across paper and online formats found that while item-level invariance held, differential bundle functioning (DBF) favored online

groups on figure-related materials, suggesting technology-related construct-irrelevant variance for certain item clusters (Kim & Circi, 2023). Complex linguistic features in TEIs may also introduce construct-irrelevant variance for English learners (Abedi, 2006; Kopriva et al., 2007). Virginia's RFP should require vendors to provide differential item functioning (DIF) and DBF evidence for any new item types prior to operational deployment.

Through-Year Assessment Models

Through-year assessment, in which multiple test events across the school year collectively inform a summative determination, is the most actively funded and rapidly evolving area of state assessment innovation. The 2024 CGSA competition awarded a total of approximately $30 million to 10 states, the program's largest investment to date. Four of those grants directly support through-year models.

As of the 2025–26 school year, one state (Montana) has fully replaced its traditional end-of-year summative assessment with a through-year model for federal accountability, and another (North Carolina) has integrated through-year data into a redesigned end-of-year summative assessment used for accountability.

Montana's MAST is the most complete example of a through-year replacement. After piloting in 2022–23 and administering a large-scale field test with more than 30,000 students in 2023–24, MAST became the operational statewide assessment for Grades 3–8 in mathematics and ELA in 2024–25, fully replacing Smarter Balanced. Approximately 66,500 students participated in the first full statewide administration. Montana educators set new cut scores to establish a baseline, and testlet scores were aggregated into a summative determination used for federal accountability. The system is now subject to ESSA peer review, and there is tension between the peer review framework (designed around a single end-of-year test) and Montana's through-year design. How peer review resolves that tension will have implications for any state considering a similar path, including Virginia. Montana received approximately $7 million in CGSA funding across 2022 and 2024 grants.

North Carolina's NCPAT completed its statewide rollout in 2024–25, adding Grades 3 and 6 to the piloted Grades 4, 5, 7, and 8. The NC Check-Ins 2.0 interim assessments are available statewide in reading and mathematics at Grades 3–8 and science at Grades 5 and 8. They are designed to provide immediate formative feedback to educators to inform instruction and to generate data that can be used statistically to assign students to appropriate levels of the multistage adaptive end-of-grade tests. Beginning in 2025–26, the multistage adaptive summative serves as the regular statewide assessment, with results used for ISRs, state and local report cards, and the federal accountability system. An independent evaluation by the National Center for the Improvement of Educational Assessment (Brandt, 2023) confirmed that the North Carolina Department of Public Instruction (NCDPI) accomplished what it set out to do: create a more personalized assessment system that supports instructional decision-making while maintaining score comparability to the original end-of-grade tests. Of the through-year models now operational, North Carolina's may be the most relevant to Virginia because it preserves a rigorous end-of-year summative assessment for accountability while using through-year data to improve the testing experience for students and simultaneously providing formative feedback to educators.

Florida's FAST remains one of the longest running statewide through-year assessment systems, operational since 2022, with three annual administrations (fall, winter, spring) for voluntary prekindergarten (VPK) through Grade 10 in ELA and for VPK through Grade 8 in mathematics. In practice, only the spring administration counts for accountability, making it functionally a traditional end-of-year test preceded by progress monitoring windows. All three administrations test the full blueprint, including content not yet taught in earlier windows.

Indiana moved from a voluntary pilot (approximately 1,350 schools in 2024–25) to mandatory statewide implementation in 2025–26. All Indiana schools must now administer three ILEARN Checkpoints in Grades 3–8 for mathematics and ELA, with a shortened end-of-year summative assessment. The checkpoints provide through-year data to inform instruction, but the end-of-year test remains the accountability measure.

The Nebraska Student-Centered Assessment System (NSCAS) Growth is a computer-adaptive assessment administered in fall, winter, and spring for Grades 3–8 in ELA and mathematics. It is designed to provide multiple measures throughout the year, including instructional and growth information as well as summative interpretation. NSCAS Growth has been fully operational statewide since 2022–23, making it one of the longest running through-year systems, alongside Florida's. Nebraska's design is distinctive because it replaced both the state summative assessment and MAP Growth (the interim assessment that most districts in the state were already using) in Grades 3–8 ELA and mathematics, with a single adaptive through-year system administered in fall, winter, and spring, streamlining interim and summative testing within a single adaptive model. The spring administration produces summative proficiency scores for accountability, and the fall and winter administrations produce instructional data on the same scale. Nebraska's approach is relevant for Virginia because it directly addresses a problem that Virginia's work group identified: the perceived redundancy between state-administered growth assessments and locally chosen interim tools such as MAP. Nebraska eliminated that redundancy by design.

Georgia's withdrawal from IADA in February 2023 remains the most instructive cautionary tale. Both of Georgia's pilot consortia collapsed due to comparability requirements, the tension between formative and summative purposes, and COVID disruption. Georgia's experience illustrates that the bar for replacing established summative assessments with innovative models is high under current federal requirements, though Montana and North Carolina have since demonstrated that the bar can be met with sustained investment and careful design.

Psychometric challenges with through-year models remain, even as states move forward. How to aggregate scores from multiple administrations into a single defensible summative determination is a design problem that states are solving differently. Montana aggregates testlet scores; North Carolina uses through-year data only for routing, not for scoring; Florida and Nebraska count only the spring score; Indiana counts only the shortened end-of-year test. No consensus approach has emerged. Through-year models have also not reduced total testing time in practice. A study of four states (Croft et al., 2023) found that districts maintain existing local interim assessments alongside state through-year tests—adding to, rather than replacing, testing. Virginia should plan for this reality, rather than assuming that a through-year model will reduce the assessment burden on schools.

B. Near-Term Opportunities for Virginia

Virginia's existing assessment infrastructure provides a foundation for modernization in several areas. The opportunities described in this section are grounded in technologies that are operational elsewhere and feasible within the procurement and implementation timeline established by HB 1600.

Expanding and Documenting Automated Scoring

Virginia already uses automated scoring through Pearson. The near-term opportunity is twofold: first, obtain and publish documentation of the current system's scope, accuracy, and fairness performance, and second, expand automated scoring to new constructed-response and performance-task items being developed for IRW. The Texas and Smarter Balanced hybrid models (machine scoring with systematic human verification) provide tested implementation frameworks. Any expansion should include a dedicated fairness analysis for English learners, students with disabilities, and other subgroups, with results reported transparently.

Adopting MST for Stimulus-Heavy Content

Virginia's next assessment vendor should be expected to propose adaptive delivery designs that account for passage-dependent items. MST is well suited for ELA and science assessments where items share a stimulus. Maryland's recent MST implementation demonstrates feasible approaches. Virginia should require offerers to provide evidence of score comparability between any proposed MST design and its current program.

Piloting Through-Year Components with Caution

Virginia's current system (until July 1, 2026) already includes optional fall and winter growth assessments in Grades 3–8 reading and mathematics, providing some operational infrastructure for multiple assessment windows. Virginia law had mandated a statewide through-year growth system (originally enacted through HB 2027 and SB 1357 in 2021, with additional specifications in HB 1957); SB 200 repealed the through-year mandate, leaving through-year assessment as a design option rather than a statutory requirement. Lessons from other states suggest that Virginia should proceed carefully. A model that uses fall and winter administrations for instructional purposes and test personalization (similar to North Carolina's approach) while maintaining a rigorous spring summative for accountability would preserve instructional utility without the psychometric risks of score aggregation across windows. Virginia should avoid conflating instructional assessment with accountability measurement without rigorous validation and should plan for the reality that through-year models have not reduced overall testing burden elsewhere.

Strengthening Accessibility Infrastructure

Virginia's current accessibility features are comparable to most state programs. The Smarter Balanced consortium offers the most extensive multilingual support nationally, including translated glossaries in 13 languages and test directions in 20 languages. Virginia's new vendor should be expected to match or exceed this level of support. Native language assessments present both an equity opportunity and a significant implementation challenge. Stakeholders in QIP's engagement sessions repeatedly identified language complexity as a barrier for English learners, noting that math and science assessments often measure reading proficiency rather than content mastery. Translated or transadapted assessments can address this problem, but they carry substantial costs and technical requirements. The single RFI vendor that provided detailed

assumptions estimated translation costs at approximately $165K, and Massachusetts's contract, which includes translated assessments, runs approximately $37.9M per year for a system with fewer tests than Virginia's. Beyond cost, translated assessments require careful transadaptation rather than direct translation to ensure that items measure the same construct across languages, and each translated form requires its own field testing, equating, and comparability evidence. Scoring, reporting, and accommodation systems must also support multilingual delivery. Virginia should weigh these costs against the equity benefits, particularly given that listening sessions have identified native language access as a priority, and should determine during RFP development whether to require translated assessments in the initial contract scope or build in an option for phased implementation. Minnesota received $4 million in 2024 CGSA funding to develop accessible ELP assessment pathways for multilingual learners with significant cognitive disabilities, the first major federal investment specifically targeting this intersection. AI-powered accessibility features (e.g., neural text-to-speech, avatar-based sign language, real-time translation) are not yet validated for summative use and should not be considered for near-term solutions, but the RFP could require vendors to describe their developmental roadmaps for these capabilities.

Technologies Not Ready for Operational Use

Several widely discussed AI-powered technologies are not yet ready for operational summative assessment and should not drive procurement decisions. For a broader review of AI applications across the full assessment life cycle, see Lorié and Dadey (2026).

No state has adopted AI-generated items operationally for summative scoring purposes, though Hawaii has been experimenting with generative AI applications within its statewide field-testing program, including using large language models (LLMs) to predict item difficulty, simulate student responses, and support item authoring workshops (Kūkea-Shultz & Brockmann, 2025). The technology accelerates drafting but requires human review for content validity, distractor quality, and alignment. It may reduce item development costs over time, but it is not a substitute for the human expertise required in item writing and review.

LLM scoring of constructed responses has shown variable performance in research settings, with agreement indices ranging from 0.30 to 0.80 depending on context (Li et al., 2025), and has not been validated for high-stakes K–12 use. All operational automated scoring systems in state summative programs use traditional machine learning approaches, though vendors are increasingly incorporating LLM components into ensemble scoring models. LLM-based scoring may mature into a viable option during Virginia's next contract period.

AI-powered accommodations (neural text-to-speech with context-aware pronunciation, avatar-based ASL interpretation, real-time machine translation) remain in early research and development. No state assessment program currently deploys any of these accommodations in summative testing. Assessment platforms continue to rely on device-native text-to-speech engines and pre-recorded human ASL videos.

Cognitive diagnostic CAT, multidimensional CAT, and AI-driven item selection are active areas of psychometric research. These represent long-term possibilities, not procurement requirements.

VIII. Procurement Guidance Overview

This section describes procurement-related guidance on key decisions and strategies for VDOE's consideration, task components to be included in an eventual RFP, and next steps to allow VDOE to proceed with an effective procurement process.

A. Procurement Strategy Highlights

One key decision for Virginia's consideration is determining whether a single contract or multiple contracts (e.g., separating test administration from test development, or including/excluding alternate assessments) would best serve Virginia's needs. A single contract could streamline administration and ensure consistency, while multiple contracts could encourage broader vendor participation and innovation. In either case, it will be important to clarify which assessments are included.

Another key decision point is the length of the contract. Virginia will need to balance the need for stability with the ability to continue to improve the assessment system and to include new partners or methods over time.

The state will also need to decide how to best incentivize and manage strong vendor performance, such as identifying KPIs for key deliverables and potential penalties. Virginia should weigh its capacity to monitor performance in determining potential remedies.

Lastly, the state will need to determine its process for change management and innovation. Building in a structured approach can ensure that a modification or change process will support necessary adaptation and innovation over time.

To attract high-quality proposals and achieve a competitive and effective outcome, several procurement strategies should be considered. Virginia should carefully outline a timeline that will allow for a thorough process and reduce risk. Typical timelines range from 6 to 10 months from procurement release to decision-making, with additional time for final execution of contracts. This range does not include time for vendor transition, if required.

Developing criteria for the state to easily evaluate across vendors will also be important. This includes establishing minimum qualification thresholds in relation to financial and operational requirements and relevant experience. This process should also identify the items of highest priority, such as technical or management plans, to allow Virginia to consider the best-value proposal overall. Virginia may also want to consider incorporating demonstrations into the procurement process, to better evaluate vendor capabilities.

Other potential procurement strategies to consider include releasing a draft RFP to identify potential barriers to competition, clarify technical requirements, and strengthen the overall solicitation, and hosting a pre-proposal conference to gauge vendor interest, address questions, and adjust the RFP as needed.

B. Key RFP Components

A review of multiple state contracts and RFPs identified recommended RFP components by examining the most commonly included tasks across states. While states may organize these sections differently, including these components in Virginia's RFP is important to ensure that the assessment system contract addresses key elements.

Program/Project Management

This section will typically include requirements for the vendor's management team and staffing; kickoff/planning meetings; ongoing management meetings and documentation (e.g., agendas, minutes, issue/risk logs, decision logs); progress reporting; master schedules/workplans and milestone tracking; and clear expectations for how documents and data will be stored, versioned, and shared (including what systems/tools are required). They often also include formal change control, dependency management, governance/approval pathways (who signs off on what), performance standards/service-level agreements and remedies, and transition-in/transition-out planning (e.g., knowledge transfer, asset transfer, continuity of operations). These sections may also include requirements for stakeholder or advisory group engagement or TAC support (unless these requirements are covered elsewhere).

Transition Planning

This section typically includes minimum parallel operation periods with the outgoing vendor; acceptance testing gates that should be passed before the state commits to a go-live date; rollback plans if acceptance criteria are not met; escrowed source code and item bank exports in vendor-neutral formats; documented knowledge transfer sessions; pilot administration with a subset of districts before statewide launch; and psychometric replication requirements (e.g., the incoming vendor must demonstrate that it can reproduce IRT parameter calibrations and maintain the existing scale). If VDOE elects to use financial holdbacks, they should be tied to successful transition milestones and included in this section.

Item and Test Development

Sections on item and test development typically include expectations related to ownership and intellectual property; item imports and use of existing items; item development planning; required item types and formats; item banking/repository needs; test specifications/blueprints and item specifications; and the item development workflow, with defined quality criteria and required metadata/tagging. They often describe the review model (editorial, bias/sensitivity, content/standards alignment, and educator involvement), passage/stimulus development expectations (including readability targets and public domain vs. commissioned preferences), translation/transadaptation requirements, and development of accommodated forms and content (e.g., braille, large print, ASL). These sections also commonly address test form construction rules, practice test development, item exposure and release plans (and resulting item-refresh/field-test needs), and production QC steps such as proof review, final form verification, and coordination of scoring keys and rubrics (either in these sections or cross-referenced in the scoring section).

Platform/Technology (Online Testing and Related Systems)

These sections typically describe requirements for the end-to-end assessment technology stack, including test delivery, student authentication and session management, proctor dashboards, accommodations delivery, secure browser integration, test registration and session management, and online reporting portals. Common elements include performance and scalability expectations (peak concurrency, load testing, monitoring), reliability and recovery (pause/resume, handling intermittent connectivity, state recovery, disaster recovery targets), and usability/accessibility (WCAG-aligned design, assistive technology compatibility, multilingual needs). RFPs frequently include technology readiness expectations for districts (device/browser specifications, bandwidth

checks, readiness tools, practice environments), interoperability requirements (APIs/standards-based exchanges such as support for state/district imports or exports), and environment/release management (training sites, staging/user acceptance testing, version control, maintenance windows, and documented deployment/change procedures).

Test Administration

These sections typically spell out the vendor's responsibilities for supporting a smooth, secure, and standardized administration across all districts/schools and across modes (online and any required paper/accommodated formats). Common elements include development and annual refresh of test administration manuals and scripts (often with state review/approval cycles and timelines); comprehensive training for district and school staff (often role-based and delivered in multiple formats, with refresh training each year, accessible materials, and expectations for tracking participation/completion); and detailed operational procedures for administration readiness and execution (technology readiness checks, scheduling, session management, and required security steps). These sections also commonly include end-to-end materials management for any paper or accommodated materials (printing/production, kitting, shipping, barcode tracking and inventory controls, chain-of-custody procedures, overage/shortage processes, secure storage, return shipping, reconciliation of missing materials, and destruction certifications), along with clear delineation of responsibilities among the vendor, districts, and state. RFPs frequently specify customer support expectations before and during windows, including a dedicated help desk/call center with defined hours (often extended during test windows), response and resolution service-level agreements, tiered escalation paths, multilingual capacity where required, ticketing systems and auditable logs (including trend reporting to the state), and rapid incident response for time-sensitive issues that could disrupt testing. Finally, these sections typically define processes for handling irregularities and exceptions (documentation and reporting workflows, investigation support, make-up testing rules, and procedures for invalidations and reopens/resumes where allowed), plus expectations for accommodations and accessibility supports (configuration and delivery through the platform and/or special forms) and day-of-testing safeguards (e.g., secure browser and restrictions during testing).

Scoring

These sections typically lay out how the vendor should produce accurate, defensible scores across all item types and modes, and how the state will review/approve key scoring outputs before results are released. Common elements include scoring methods for machine-scored and constructed-response items (rubrics, scoring rules, and key verification); rater recruitment/training/qualification and monitoring; reliability targets and monitoring (including inter-rater reliability and validity reads); procedures for discrepant scores (second/third reads and adjudication rules); and QC/verification steps (answer key validation, roster/demographic checks, scoring model checks, and pre-release verification). These sections often also include requirements for automated scoring (if allowed), including model training data, human-in-the-loop QC, monitoring for drift, and documentation. They may require psychometric deliverables tied to operational scoring (calibration/equating/scaling), artifacts such as scoring tables and raw-to-scale conversions, and independent replication/verification. Finally, they typically specify timelines (including state review windows), rescore/recalculation processes, and how corrections are propagated to reports and data files.

Reporting

These sections typically spell out the vendor's responsibilities for producing, delivering, and supporting clear, accurate, audience-appropriate reports in both paper and electronic formats, with tight timelines and explicit state review/approval checkpoints. Common requirements include defining the full report suite (ISRs plus class/school/district/state summaries and rosters), ensuring federal/state reporting needs for aggregation and disaggregation, and specifying required elements (e.g., scale scores; achievement levels; conditional standard error of measurement; subscores/domain indicators; growth, where applicable; participation/validity indicators). RFPs frequently require both static deliverables (e.g., secure PDFs) and an online reporting portal that is role-/permission-based, supports filtering/drilldown and export/print-on-demand, and maintains archives across years. They also often require distinct audiences/portals (family/student vs. educator/administrator), accessibility and language requirements (e.g., translated parent-facing materials), and interpretive report content (narratives, graphics, explanations). Finally, they typically include report design review cycles (mock-ups, stakeholder feedback, state approval), rigorous QC and reconciliation against data files, and defined schedules and signoffs to support fast but defensible release.

Datafiles

These sections typically define the vendor's responsibilities for producing complete, accurate, secure, and state-usable data extracts after each administration (and often additional end-of-year comprehensive files), including student-, test-, and item-level records that support accountability, reporting, and research. Common requirements include agreed file formats and layouts (CSV/TXT/XLSX/XML or state-defined schemas), a complete codebook/layout with business rules, and clear version control so the state can track what changed between preliminary and final files. RFPs also routinely specify secure transfer mechanisms (portal/SFTP/SharePoint/FTP), naming/folder conventions, encryption expectations, and delivery checklists. They typically require correct merging of paper and online data, inclusion/exclusion rules (invalidations, no-tests, refusals), and pre-delivery QC (roster reconciliation, demographic completeness, accommodation flags, school/district codes, duplicates). Finally, these sections usually include delivery schedules and state review windows, processes for corrections and resubmissions, and clear handling of late updates (late invalidations, rescores, appeal outcomes) so that replacements are traceable and consistent across reports and files.

Other Analytic/Research or Psychometric Support

If these topics are not addressed elsewhere, some RFPs include specific sections describing ongoing psychometric, technical, and research responsibilities that support score validity, fairness, and comparability over time. In addition to calibration/equating/scaling, common requirements include operational analyses of item/test performance and item bank health (drift/stability, form statistics, conditional precision), reliability and classification accuracy/consistency analyses, and formal validity evidence (content/construct/criterion and consequential), including subgroup studies. RFPs often call for fairness/bias work (DIF using multiple methods plus content review), dimensionality/model-fit and invariance checks, and comparability studies across modes/devices/interfaces or accommodated vs. non-accommodated administrations. Many include research on accessibility tool use/impact/validity, field test or special study design and analysis, and growth/progress measure development and validation, where relevant. These sections also commonly require technical documentation deliverables

(technical reports, interpretive/technical manuals, year-over-year change documentation) and support for external review/peer review (assembling evidence packages and responding to findings), as well as independent replication/third-party verification of key psychometric outputs.

Communication

If this topic is not addressed elsewhere, some RFPs include specific sections on communications, including expectations for how the vendor will manage routine, stakeholder, and incident communications. Common elements include protocols for reminders/notices and escalation when deliverables, risks, or timelines are in trouble; descriptions of how communication with districts/local education agencies should be handled; and requirements for specific vendor-provided tools for communication and collaboration (e.g., toll-free conference line, online meeting services, secure electronic communications/transfer methods). Finally, some RFPs explicitly include external communications support: ensuring consistent messaging across audiences (legislature, media, stakeholders, the public, federal oversight) and requiring the contractor to help explain platform performance and assessment validity/reliability when needed, including communicating with districts about service interruptions or scoring/reporting errors.

Security

Some RFPs include stand-alone security sections; others embed security requirements throughout. Common elements of these sections include secure handling of confidential test content and personally identifiable information; role-based access controls and least-privilege administration; encryption in transit and at rest; audit logging and monitoring; secure development and vulnerability management; incident response and breach notification procedures; and subcontractor controls. These sections often also include operational test security requirements, such as secure browser/lockdown expectations, chain-of-custody for paper materials, irregularity investigation support, and, in some cases, social media monitoring and data forensics to detect compromised content or anomalous response patterns. Many also include privacy and data governance expectations (FERPA/state privacy compliance, retention/destruction schedules, access governance, and auditability of changes to key outputs such as data files and reports).

Federal Compliance Support

Depending on how responsibilities are divided, the RFP may include requirements for vendor support of federal reporting and oversight. This may include assembly of ESSA peer review evidence packages, participation rate monitoring, and documentation of accommodations use to support peer review findings. Some states also require the vendor to respond to federal findings or requests for additional evidence. If Virginia pursues through-year or competency-based approaches that require an IADA waiver or other federal approval, the RFP should require vendor support for waiver applications, including data collection, evidence assembly, and ongoing reporting to maintain waiver status.

C. Next Steps for VDOE

The major next step for VDOE is to finalize the key considerations outlined in Section VIII.A. These decisions will shape the overall procurement approach and provide a clear foundation for defining RFP requirements. As part of this process, VDOE will need to coordinate with internal departments, such as IT, procurement, legal, and program teams, to ensure that technical and compliance needs are addressed. This collaboration will support the development of a cross-functional RFP, positioning VDOE for a successful procurement process.

Appendix A. Virginia Innovative Assessment Work Group Members

Table A1 lists members of the Virginia Innovative Assessment Work Group. The table includes their role and/or title and corresponding agency or organization, as well as their current membership status as of April 2026.

Table A1. Virginia Innovative Assessment Work Group Members

Name	Role/Title	Status (as of April 2026)
Beth Ackerman	Senior Vice President, Rivermont Schools	Active
Chad Aldeman	Education Analyst, Center for Reinventing Public Education	Active
Kristen Amundson	Former Executive Director of NASBE; Former Member VA House of Delegates; Former Member of Fairfax County School Board	Active
Shelby Averette	Division Testing & Accountability Coordinator, Mecklenburg County Public schools	Active
Karissa Barnes	Instructional Technology Resource Teacher, Lancaster County Public Schools	Active
Sarah Beach	Director of Research, Education Rights Institute, University of Virginia School of Law	Active
Scott Brabrand	Executive Director, VASS	Active
Douglas Brown	Co-Chair, Virginia School Boards Association	Active
Ashley Chadwell Bruner	Middle School Math Specialist, Lee County Public Schools	Active
Wendy Chandler	DDOT, Augusta County Public Schools	Active
Jim Cowen	Executive Director, Collaborative for Student Success	Active
Beau Dickenson	President Emeritus, Virginia Social Studies Leaders Consortium (VSSLC); K–12 Social Studies Supervisor, Rockingham County Public Schools; Assessment Consultant, C3 Teachers	Active
Christopher R. Gareis	Professor of Education, William & Mary	Active
Victoria Gelbert	Principal, Cumberland Co.	Active
Amy Griffin	Deputy Executive Director, VASS	Active
Megan Hawley	Principal, Montgomery County Public Schools	Active
Andrea Herndon	Coordinator of Gifted Education and Advanced Studies, K–12, Hanover County Public Schools	Active
Christy Hovanetz	Senior Policy Fellow, ExcelinEd	Active
Victoria Hulick	Teacher, Lancaster County Public Schools	Active
Matt Hurt	Director, Comprehensive Instructional Program	Active
Anne Hyslop	Director of Policy Development, All4Ed	Former
Tara Jenkins	ESL Title III Coordinator, Chesterfield County Public Schools	Active
Chris Jones	Executive Director, Virginia Association for Teaching, Learning, and Leading	Active
Paula Mooradian	Title III Project Director/Data Manager, Virginia Community Colleges (VCCS)	Active
Shelly Norden	Executive Director, School Board Member Alliance (SBMA)	Former
Tian Olson	Virginia Council on the Interstate Compact on Educational Opportunity for Military Children member/Virginia K–12 Parent	Active

Name	Role/Title	Status (as of April 2026)
Clinton Page	Chief of Accountability and Research, Alexandria City Public Schools	Active
Alisa Jane Pappas	Advanced Academic Resource Teacher, Hampton City Schools	Active
Melanie M. Patterson	Principal, Virginia Association of Secondary School Principals (VASSP)	Active
Alan Seibert	Former State Board of Education Member; Former Salem City Schools Superintendent; Constituent Services and Gov. Relations Officer of Roanoke City Public Schools	Active
Daniel W. Smith	Superintendent, Stafford County Public Schools	Active
Kristy Somerville-Midgette	Superintendent, Lynchburg City Schools, Virginia Association of School Superintendents (VASS)	Former
Chad Stewart	Policy, Virginia Education Association	Active
Isabel Sullivan	Professional Learning/Grant Coordinator, Virginia Education Association	Alternate (for Chad Stewart)
Todd Truitt	Parent	Former
Kelly Tucker	Special Education/Structured Literacy Teacher, Arlington County Public Schools	Active
Miranda Turner	Partner, Plevin & Turner LLP	Active
Genevieve Kathryn Weaver	VA Teacher of the Year (Region 5), Adapted Special Education, Jefferson Forest High School Bedford	Active
Latrese Younger	Principal, Richmond City Public Schools	Active

APPENDIX B. EXEMPLAR STATE PROFILE BY KEY FOCUS AREA

Focus Area A: Alignment with Standards and Instructional Goals

Virginia's modernization agenda positions alignment as the foundation for both valid measurement and instructional usefulness, aiming for statewide assessments that clearly measure student learning aligned to the SOLs in a way that supports teaching and learning decisions. For a review of this focus area, WestEd examined how exemplar states document and verify alignment of assessments to standards, develop test blueprints, and communicate connections between assessment items and curriculum expectations.

Massachusetts

The Massachusetts Comprehensive Assessment System (MCAS) documents alignment in a way that is both technically defensible and usable for educators—linking standards, test structure, and item demands through a coherent set of public artifacts. The 2023 Next-Generation MCAS and MCAS-Alt Technical Report (Massachusetts Department of Elementary and Secondary Education [MDESE] & Cognia, 2024) describes alignment as an end-to-end process: it begins with test and item specifications and blueprints, proceeds through structured educator and expert reviews, and culminates in the public release of materials that show how operational content connects to the Massachusetts Curriculum Frameworks.

Appendix C of the MCAS and MCAS-Alt Technical Report ("Test Design and Blueprint Specifications") provides grade- and content-specific blueprint tables that specify the number of items and total points by item type and session, along with target percentage ranges of points by reporting category (see Figure B1 for an illustrative example from Grade 3 ELA). These blueprint specifications function as a primary alignment control by translating standards expectations into measurable design targets.

Figure B1. MCAS Blueprint Example for Grade 3 ELA

Distribution of ELA Common Items by Session and Item Type—Grade 3

Session	Number					Points				
	SR1	SR2	CR	ES	Total	SR1	SR2	CR	ES	Total
1	11	1	0	1	13	13		0	7	20
2	15	3	1	0	19	21		3	0	24
Total	26	4	1	1	32	34		3	7	44

Item Types: 1 SR1 = MC; 2 SR2 = 2-pt MC or TE, EBSR; 3 CR (hand scored); 7 ES (hand scored)

Reporting Categories	G3	
Language	25%	+/-5%
Reading	65%	+/-5%
Writing	10%	+/-5%
Totals	100%	

Beyond the blueprint tables, the technical report includes test design rules that further reinforce alignment. For example, ELA passages are selected to be grade-appropriate, support item development, and be free of bias/sensitivity issues; passage sets include single or paired passages drawn from published works and have grade-scaled word-count guidance. Massachusetts further operationalizes item-to-standards alignment through released-item documentation (MDESE, 2025d) that includes item-level tables identifying each item's reporting category, specific standard(s) assessed, and item type, making the standards connection transparent at the level of individual items.

Finally, the technical report situates alignment within a broader validity argument: it explains that validity evidence is documented across test design and development, administration, scoring, item statistics (including DIF), reliability, and reporting, and it provides summary tables to help readers locate these different forms of evidence. The report's content validity section (beginning around page 82) further underscores the role of expert review as part of the evidence base supporting test content. Appendix B of the MCAS and MCAS-Alt Technical Report also includes achievement level standards and descriptors that describe, by grade and content area, what students at each performance level are expected to know and be able to do, supporting consistent interpretation of results alongside the alignment documentation.

Smarter Balanced

Smarter Balanced employs ECD as its foundational framework for documenting how the assessment's claims, evidence, and task design are derived from the Common Core State Standards (CCSS) and translated into operational test content (Smarter Balanced Assessment Consortium [Smarter Balanced], 2023). The consortium's technical documentation describes how content specifications define the major claim structure and the assessment targets within claims. Item and task specifications are then built from these targets so that items can be coded to specific claims and targets and mapped back to the relevant standards.

Smarter Balanced operationalizes this alignment through a transparent public resource ecosystem. The Smarter Balanced Content Explorer (Smarter Balanced, n.d.-d) provides a user-friendly interface for examining how assessment targets relate to grade-level expectations and for accessing key test development materials. The consortium's summative test blueprints, available through the Content Explorer, specify the distribution of content across both the CAT and performance task components, and include grade-level guidance for sampling targets and content (see Figure B2 for an example from Grade 3 mathematics).

Figure B2. Target Sampling Example Based on Smarter Balanced Blueprints for Grade 3 Mathematics

Target Sampling Mathematics Grade 3						
Claim	Content Category	Assessment Targets	DOK[1,2]	Items		Total Items
				CAT	PT	
1. Concepts and Procedures	Priority Cluster	B. Understand properties of multiplication and the relationship between multiplication and division.	1	5-6	0	17-20
		C. Multiply and divide within 100.	1			
		I. Geometric measurement: understand concepts of area and relate area to multiplication and to addition.	1, 2			
		G. Solve problems involving measurement and estimation of intervals of time, liquid volumes, and masses of objects.	1, 2			
		D. Solve problems involving the four operations, and identify and explain patterns in arithmetic.	2	5-6		
		F. Develop understanding of fractions as numbers.	1, 2			
		A. Represent and solve problems involving multiplication and division.	1, 2	2-3		
	Supporting Cluster	E. Use place value understanding and properties of operations to perform multi-digit arithmetic.	1	3-4		
		J. Geometric measurement: recognize perimeter as an attribute of plane figures and distinguish between linear and area measures.	1			
		K. Reason with shapes and their attributes.	1, 2			
		H. Represent and interpret data.	2, 3	1		

Finally, Smarter Balanced's published alignment research includes a comprehensive alignment study report (HumRRO, 2016) and an additional external alignment study (WestEd, 2017), which examine the alignment across multiple connections in the design chain, from standards to specifications, blueprints, and operational forms, providing evidence of alignment to CCSS expectations and to selected state standards. These alignment studies largely provided evidence that Smarter Balanced items were well-aligned to standards in terms of both content and cognitive complexity across Grades 3–8 in ELA and mathematics. Finally, Smarter Balanced also supports instructional alignment through its Tools for Teachers platform (Smarter Balanced, n.d.-e), which provides standards-aligned formative resources and professional learning supports to help educators connect instruction to assessment evidence.

Tennessee

The Tennessee Comprehensive Assessment Program (TCAP) includes publicly available assessment blueprints that specify the intended distribution of operational items across reporting categories and standards (Tennessee Department of Education [TDOE], n.d.-a), including the number of items and points allocated by standard for each grade and subject (see Figure B3 for an example from Grade 3 ELA).

Figure B3. TCAP Assessment Blueprint Example for Grade 3 ELA

TCAP Assessment Blueprint
English Language Arts - Grade 3

This blueprint describes the content and structure of an assessment and defines the ideal percentage and number of operational test items by reporting category for the Tennessee Comprehensive Assessment Program (TCAP). For more information on this assessment, please see the Assessment Overviews here.

Reading: Informational Text — 32 – 34% of Items

Standard	Ideal Point Range
3.RI.KID.1	1 – 6
3.RI.KID.2	1 – 6
3.RI.KID.3	1 – 6
3.RI.CS.5	1 – 4
3.RI.CS.6	1 – 4
3.RI.IKI.7	0 – 4
3.RI.IKI.8	0 – 4
3.RI.IKI.9	0 – 4
TOTAL	9 – 14

Reading: Literature — 32 – 34% of Items

Standard	Ideal Point Range
3.RL.KID.1	1 – 6
3.RL.KID.2	1 – 6
3.RL.KID.3	1 – 6
3.RL.CS.5	1 – 4
3.RL.CS.6	1 – 4
3.RL.IKI.7	0 – 4
3.RL.IKI.9	0 – 4
TOTAL	9 – 14

Reading: Vocabulary — 13 – 18% of Items

Standard	Ideal Point Range
3.RL.CS.4	0 – 4
3.RI.CS.4	0 – 4
3.FL.VA.7a	1 – 7
3.FL.VA.7b	0 – 4
TOTAL	4 – 9

Conventions — 13 – 14% of Items

Standard	Ideal Point Range
3.FL.PWR.3	0 – 3
3.FL.WC.4	0 – 3
3.FL.SC.6	1 – 5
3.FL.VA.7c	0 – 3
TOTAL	2 – 7

Written Expression — 5 – 7% of Items

Standard	Ideal Point Range
3.W.RBPK.8	0 – 16
TOTAL	16

Summary

Operational Items	Operational Points
27 – 39*	52

*Please note that these numbers reflect operational items only. All ELA TCAP assessments also include embedded field test items.

As illustrated in Figure B3, the blueprint defines the ideal percentages and numbers of operational items by reporting category, and then lists the standards within each category, along with ideal point range values tied to those standards. This structure makes the blueprint directly usable for alignment review because it translates Tennessee Academic Standards into explicit design targets.

Tennessee supplements blueprints with grade- and subject-specific "Assessment Overview" documents that describe how the blueprint is realized in the operational test design, including how non–multiple-choice components (such as writing prompts) are incorporated and aligned to standards, helping educators to understand what the assessment is asking students to do in standards terms.

Tennessee also provides released item sets that allow educators to examine operational content and see how items are linked to Tennessee Academic Standards. The state releases test items through its LiveBinders platform (TDOE, 2023) and posts direct links to numerous TCAP item release PDFs, across grades and courses, which include metadata tables that tag each item to a Tennessee standard.

Although Tennessee's publicly available alignment study reports are not easily accessible in the same level of technical detail as is available for Massachusetts or Smarter Balanced, its blueprints, assessment overviews, and standards-tagged item releases provide a clear, public alignment record that educators and stakeholders can access and use.

Implications for Virginia

Virginia can strengthen its alignment documentation by combining the most transferable features of these exemplars into a coherent "alignment package." From Massachusetts, Virginia can adopt a more transparent technical reporting approach with validity evidence that also clearly documents blueprint intent and provides public-facing artifacts that make alignment inspectable. From Smarter Balanced, Virginia can adopt a claims-and-targets structure, grounded in ECD, that provides a particularly robust framework to define what proficiency and application of knowledge mean under the SOLs. Such an approach will ensure that item development, form construction, and reporting remain consistent with the defined proficiency expectations. From Tennessee, Virginia can adopt an implementation-facing transparency strategy that includes clear, educator-usable blueprints and standards-tagged released materials to reduce ambiguity during transitions and improve instructional usability. Taken together, these practices would help Virginia align standards, assessment design, and reporting in a way that is both technically defensible and readily interpretable by educators and the public.

Focus Area B: Non-Traditional and Performance-Based Approaches

Several states have received federal approval through IADA to pilot alternative approaches to traditional standardized testing. These programs explore competency-based assessment, curriculum-embedded tasks, and performance-based demonstrations of learning. WestEd's research on this focus area examined five exemplar programs representing diverse approaches to assessment innovation.

New Hampshire

New Hampshire's Performance Assessment of Competency Education (PACE) was the first federally approved innovative assessment pilot, initially receiving waiver approval in 2015 and subsequently approved under IADA in 2018 (New Hampshire Department of Education, 2018). PACE is grounded in competency-based education, ensuring that students demonstrate mastery of critical knowledge and skills before advancing.

The PACE system combines locally developed, locally administered performance assessments, tied to grade and course competencies, with common tasks administered across participating districts. The common tasks serve as calibration tools, providing evidence about comparability of proficiency judgments across schools and districts (New Hampshire Department of Education, 2020). Student proficiency determinations incorporate educator judgments at year-end, based on achievement level descriptors aligned with Smarter Balanced standards.

Key implementation features include a task bank of previously validated assessments that districts can access (available through the New Hampshire Learning Initiative [n.d.-a]), teacher leader networks for building local capacity, and a tiered participation system that allows districts to engage at different levels of implementation. Districts participating in PACE administer the state summative assessment only once per grade span (Grades 3–5, 6–8, and 10–12) rather than annually, reducing overall testing burden (see Figure B4 for an overview of the PACE system's grade and subject combinations for assessments).

Figure B4. New Hampshire PACE System Overview of Assessments by Grade and Subject Area

Grade	ELA	Mathematics	Science
3	Statewide assessment system (NH SAS)	Performance assessment system (PACE)	Local performance assessments
4	Performance assessment system (PACE)	Statewide assessment system (NH SAS)	Local performance assessments
5	Performance assessment system (PACE)	Performance assessment system (PACE)	Performance assessment system (PACE)
6	Performance assessment system (PACE)	Performance assessment system (PACE)	Local performance assessments
7	Performance assessment system (PACE)	Performance assessment system (PACE)	Local performance assessments
8	Statewide assessment system (NH SAS)	Statewide assessment system (NH SAS)	Performance assessment system (PACE)
9	Course-specific common performance assessments	Course-specific common performance assessments	Course-specific common performance assessments
10	Course-specific common performance assessments	Course-specific common performance assessments	Course-specific common performance assessments
11	SAT School Day	SAT School Day	NH SAS

Challenges have included scaling the program statewide, maintaining assessment quality across diverse districts, and the state's 2021 decision to withdraw formal support for PACE (educator-led implementation continues through the NH PLACE initiative [New Hampshire Learning Initiative, n.d.-b]). The program's reliance on local capacity and high-performing initial pilot districts raises questions about replicability in lower-capacity contexts.

Louisiana

In 2018, Louisiana received IADA approval to develop curriculum-embedded assessments aligned to the state's ELA Guidebooks curriculum (Louisiana Department of Education, 2018). This innovative approach aims for assessments to directly match the high-quality curriculum that teachers use and the instruction that students receive daily.

The Louisiana model features through-course assessments administered in three windows (fall, winter, spring) rather than a single spring summative assessment. Each unit ends with a culminating task that serves as a curriculum-embedded assessment (Louisiana Department of Education, 2025). This design allows for a more precise understanding of student learning on grade-level standards, texts, and content knowledge.

In 2022, Louisiana received a $2.9 million CGSA award to support development of mathematics and ELA curriculum-relevant through-year assessments connecting to each curriculum, scaling the approach, and helping the educator community interpret multiple administrations. In 2024, the state received an additional $3 million CGSA award to continue this work (United States Department of Education, n.d.).

The Louisiana Educational Assessment Program (LEAP) 2025 technical reports, such as the *LEAP 2025 Social Studies 3–8 Technical Report: 2023–2024* (Data Recognition Corporation [DRC] et al., n.d.), document comparability evidence across test forms, years, and languages (including Spanish). The state's vision extends to a system where multiple widely adopted curricula each have corresponding unit-based assessments that collectively meet ESSA requirements.

Colorado

Colorado's alternate assessment, CoAlt, provides a performance-based model for students with significant cognitive disabilities. The CoAlt technical reports (Colorado Department of Education, n.d.-a) detail the state's review process, including item development, field testing, data review, scoring protocols, and reliability evidence.

Colorado's IADA application focused on general education performance-based assessment, supported by philanthropic funding from the Hewlett and Gates Foundations for a multi-district performance-based accountability pilot (Colorado Department of Education, n.d.-b). Public documentation of this pilot is limited. However, the alternate assessment program demonstrates QC processes that are applicable to performance-based approaches more broadly.

Georgia and Missouri IADA Applications

Georgia submitted an IADA application in 2018 (Georgia Department of Education [GaDOE], 2018) but withdrew from the program in 2023 (see Section III for details). Missouri received IADA approval in 2025 for a modular assessment structure (Missouri Department of Elementary & Secondary Education, 2025). Since its program was recently approved, implementation evidence is not yet available, but the application itself may provide useful design insights for states considering similar approaches.

Implications for Virginia

Performance-based and competency-based approaches offer potential benefits including stronger instructional alignment, reduced end-of-year testing burden, and more authentic demonstrations of student learning. However, implementation challenges are substantial, including maintaining comparability across locally developed assessments, building educator capacity for high-quality task development and scoring, and ensuring equity across diverse district contexts. Virginia should carefully consider whether conditions support such approaches and what infrastructure investments would be required.

Focus Area C: Accessibility and Multilingual Supports

Ensuring that assessments are accessible to all students, including students with disabilities and English learners, is both a federal requirement and an ethical imperative. WestEd's research in this focus area examined how exemplar states implement UDL principles in their assessment design. States were also studied to understand how they provide communication and guidance on identifying and administering accommodations to facilitate student participation. Additionally, this work examined how states support multilingual learners to fully participate in assessments.

California

The California Assessment of Student Performance and Progress (CAASPP) system demonstrates comprehensive accessibility implementation aligned with Smarter Balanced guidelines. The California Assessment Accessibility Resources Matrix (California Department of Education [CDE], n.d.-b) provides detailed guidance on available supports, organized by type, and the assessments for which each support resource is available.

Key accessibility resources also include the CAASPP Student Needs Matching Tool (CDE, 2025a), designed to help educators and IEP teams select accessibility resources for individual students, and the Accessibility Resources Planning Tool (CDE, n.d.-a), which outlines resources,

grounded in UDL principles, for use during assessments. California's accessibility resources provide guidance on selecting appropriate accommodations and resources to address individual student needs during testing.

For multilingual learners, California provides translated test directions in multiple languages for the Smarter Balanced ELA and mathematics assessments, the California Science Test, and the English Language Proficiency Assessments for California (ELPAC) (CDE, n.d.-c). For the ELPAC assessments, which are administered to students whose primary language is not English, California provides the *ELPAC Information Guide* (CDE, 2025b), which provides guidance on the state's student identification, classification, and reclassification processes for English learners and guidance on understanding ELPAC results.

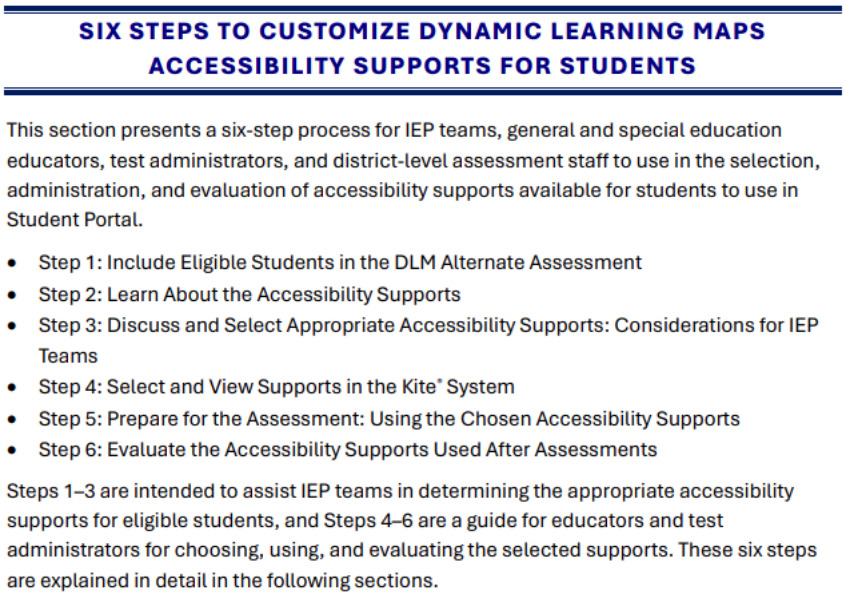
The state's annual implementation report from 2024 (CDE, 2025c) includes an overview of state supports and activities related to assessment development and administration training, as well as cost projections for accessibility research and further training.

Dynamic Learning Maps

The Dynamic Learning Map (DLM) Alternate Assessment System provides the leading model for assessing students with significant cognitive disabilities. The DLM system serves students in Grades 3–8 and high school across multiple participating states, with assessments in ELA, mathematics, and science.

The *Accessibility Manual 2025–2026* (DLM, 2025) provides guidance to state leaders, district staff, test administrators, and IEP teams on selecting and using DLM accessibility supports. This manual outlines a six-step process for selecting and implementing these supports (see Figure B5 for an overview of these steps).

Figure B5. Overview of DLM's Six Steps to Customize Accessibility Supports for Students



The manual provides guidance on confirming eligibility, selecting supports, administering assessments, and evaluating effectiveness. Supports include both online features (magnification, color contrast, overlay colors) and external accommodations provided by educators.

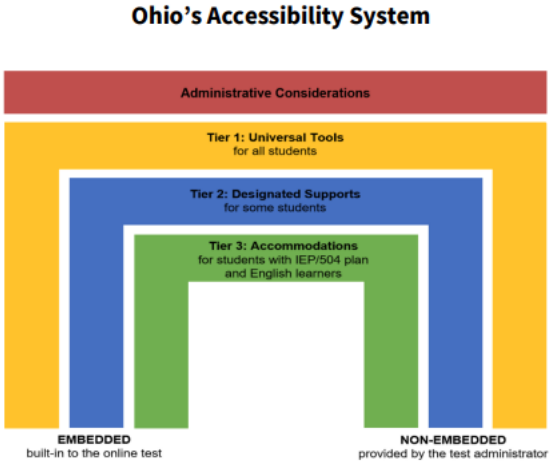
The system's *What Is a Learning Map Model?* resource (DLM, n.d.) shares links to its Essential Elements for ELA, mathematics, and science—grade-specific expectations related to college and career readiness standards. This approach ensures that students with significant cognitive disabilities access challenging, grade-level content, while allowing for multiple pathways to demonstrate knowledge.

The 2023–24 technical manual (DLM, 2024) provides extensive validity evidence, including item development processes incorporating bias, sensitivity, and accessibility reviews. Initial development was funded through a $22 million federal Office of Special Education Programs grant (2010–15) (DLM, 2010), with ongoing operational costs embedded in state consortium membership.

Ohio

The *Accessibility Manual for Ohio's State Tests* (Ohio Department of Education & Workforce, 2025) provides comprehensive guidance on accommodation options and a three-tier framework consisting of universal tools, designated supports, and accommodations (see Figure B6 for a graphic representation of Ohio's accessibility framework).

Figure B6. Ohio's Three-Tier Accessibility Framework



Ohio also provides an "Assessment Accessibility for Ohio's Students" resource page (Ohio Department of Education & Workforce, 2026a), which includes accommodation decision tools that help IEP teams and educators identify appropriate supports. Ohio also provides assessment training resources (Ohio Department of Education & Workforce, 2026b) for staff responsible for implementing accommodations.

In 2019, the state received an SLDS grant of $3.2 million, focused on "Using Data for Equity: Empowering Educators with Information, Tools and Training to Support Each Child in Achieving their Goals" (National Center for Education Statistics [NCES], 2019). This strategic plan is meant to align with state efforts to accelerate and expand data system use for educational purposes.

Implications for Virginia

Virginia should adopt comprehensive accessibility frameworks that incorporate UDL from the outset of assessment development. Decision-support tools such as California's CAASPP Student

Needs Matching Tool can help ensure that appropriate accommodations are consistently identified and implemented. For alternate assessment, the DLM model demonstrates how to maintain alignment to grade-level standards while providing accessible pathways for students with significant cognitive disabilities.

Focus Area D: Advanced Technology Integration and Infrastructure

Modern assessment systems increasingly leverage technology for adaptive testing, efficient delivery, and rapid score reporting. WestEd looked at state exemplars of CAT implementations, infrastructure requirements, and technology platforms, for insights that might inform Virginia's assessment updates.

Florida

The Florida Assessment of Student Thinking (FAST) system is one of the most technologically advanced state assessment systems currently operational. First administered in the 2022–23 school year, FAST is a progress monitoring assessment, administered three times per year (fall, winter, spring) for voluntary prekindergarten (VPK) through Grade 10 in ELA and for VPK through Grade 8 in mathematics, aligned with the state's Benchmarks for Excellent Student Thinking (B.E.S.T.) standards (Florida Department of Education [FLDOE], 2025a).

All FAST assessments are computer-adaptive, with items becoming progressively harder or easier, based on student responses, while maintaining blueprint coverage. The system uses IRT with 3PL and generalized partial-credit models (FLDOE, 2024b). Each progress-monitoring event covers the complete test blueprint for full grade-level content.

The *FAST Technology Guide* (FLDOE, n.d.-b) specifies infrastructure requirements for districts, including device specifications and network capacity. The system's rapid reporting capability delivers scores to the Family Portal within 24 hours of test completion (FLDOE, 2025a), a significant advancement over traditional end-of-year assessments that may take months to return results.

Florida contracted with Cambium Assessment for FAST delivery, with contract information publicly available through the state's financial transparency system (Florida Department of Financial Services, n.d.). The state received a 2009 SLDS grant of $10 million focused on expanding its reporting capabilities (United States Department of Education, 2009). Current operational costs are embedded in vendor contracts, rather than separately itemized.

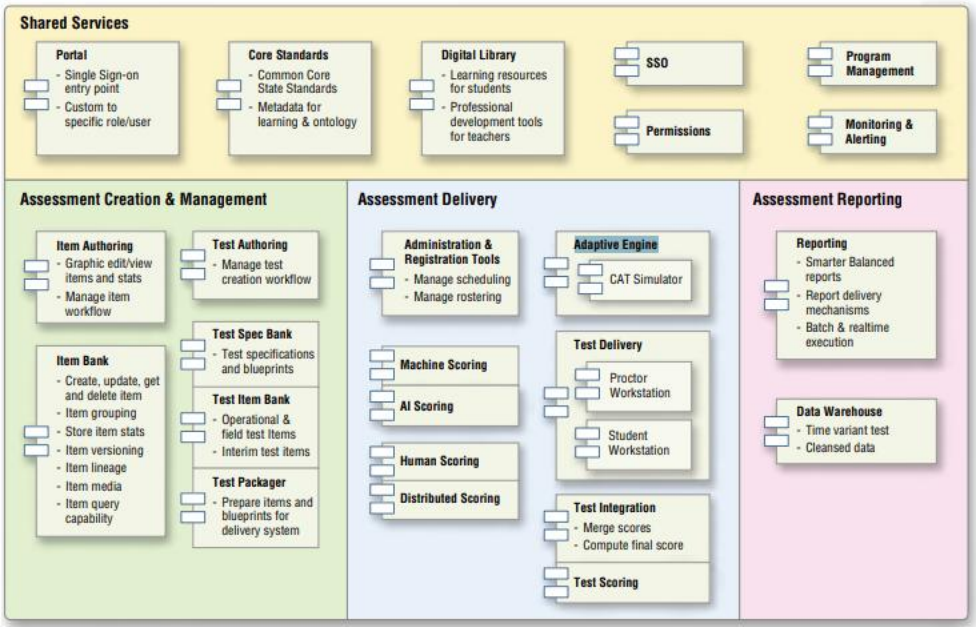
Smarter Balanced Adaptive Engine

Smarter Balanced developed a publicly available CAT algorithm (Cohen & Albright, 2014) that has remained largely consistent since the 2014–15 administration, and Cambium Assessment uses it to deliver Smarter Balanced assessments across member states.

Key features of the Smarter Balanced adaptive engine include blueprint satisfaction constraints that ensure content coverage; item exposure controls that prevent overuse of specific items; and precision optimization that maximizes measurement accuracy. The consortium's technical reports document blueprint fidelity analyses confirming that delivered tests meet content specifications.

The initial development of the adaptive testing system was funded through a Race to the Top Assessment Program grant (Smarter Balanced, 2010). The system architecture is documented in the *Smarter Balanced Architecture Report* (Smarter Balanced, 2014), and the open-source delivery system allows member states flexibility in implementation approaches (see Figure B7 for an overview of the Smarter Balanced logical components, including the delivery and reporting mechanisms).

Figure B7. Smarter Balanced Logical Components



Utah

Utah's Readiness, Improvement, Success, Empowerment (RISE) assessment system uses CAT delivered through Cambium's platform. The RISE test administration manual (Cambium Assessment, Inc., 2025b) documents the technical requirements for setting up and administering these assessments; however, its platform-specific documentation is less detailed than that of Florida or Smarter Balanced.

Utah provides resources to support RISE administration. Item Type Tutorials (Cambium Assessment, Inc., 2025a) help test administrators and students familiarize themselves with different TEI formats (e.g., evidence-based selected response, drag-and-drop, equation response). The *Quick Guide for Setting Up Your Online Testing Technology* (Cambium Assessment, Inc., n.d.) assists districts with infrastructure preparation, including guidance for setting up test administrator and student workstations and configuring networks and assistive technologies. The state's *The Future of Utah's State Assessment System* report (Utah State Board of Education, 2020) outlines its long-term vision for the assessment system, including affirming the state's focus on adaptive testing and incorporating non-summative assessments into the system.

Implications for Virginia

Adopting or enhancing CAT requires substantial infrastructure investment and ongoing maintenance. Florida's model demonstrates that through-year CAT administration with rapid reporting is technically feasible at scale. Virginia should assess current technology infrastructure

against requirements for desired assessment approaches, consider vendor partnerships (Cambium is the dominant provider), and develop realistic timelines and cost estimates for implementation.

Focus Area E: Authentic Growth Measures and Formative Feedback

Measuring student growth over time provides valuable information beyond proficiency snapshots. For this focus area, WestEd focused on innovative and systematic approaches to calculating student growth and using data to inform instructional adjustments. WestEd reviewed state documentation on growth models, interim assessment systems, and through-year approaches, which provide formative feedback to inform instruction.

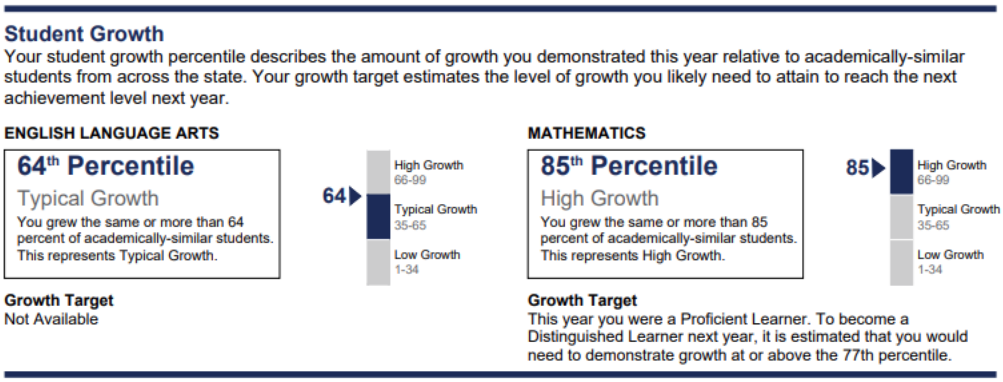
Georgia

The Georgia Student Growth Model (GSGM) provides a comprehensive approach to measuring student progress. The Georgia Milestones technical report (Data Recognition Corporation [DRC], 2025) includes a section on student growth percentiles (SGPs), which are calculated and reported in the Georgia Milestone Assessment student reports. This section of the technical report details how SGP values and targets are calculated and reported.

Researchers have evaluated Georgia's growth model, including Marion, Domaleski, and Lorié's (2025) overview of the GSGM. This research has found that GSGM aligns with best practices identified in *How Should States Choose a Growth Model?* (Marion et al., 2025).

Georgia provides extensive resources to help educators interpret growth data, including *A Guide to Understanding and Interpreting Outcomes from the Georgia Student Growth Model* (Georgia Department of Education [GaDOE], 2025a), which provides guidance on understanding and interpreting SGPs, and *Understanding Growth Targets* (GaDOE, 2025b), which supports educators and families in understanding and using growth targets. Sample ISRs for Grade 5 (GaDOE, 2024a) and Grade 8 (GaDOE, 2024b) demonstrate how growth information is communicated to families (see Figure B8 for an example of how student growth is communicated in a Grade 8 student report).

Figure B8. Student Growth Section of Georgia Grade 8 Student Report Sample



The state's SLDS (GaDOE, 2014) supports districts and schools in managing and using education data, including assessment data. Georgia received a 2009 SLDS grant of $8.9 million (United States Department of Education, 2009) and a 2023 grant for "My Georgia Insights" data platform modernization (NCES, 2023).

Montana

The Montana Aligned to Standards Through-Year (MAST) program represents an emerging model for growth measurement through multiple assessment windows. The program vision (Montana Office of Public Instruction, n.d.) and theory of action (Montana Office of Public Instruction, 2022) outline how through-year testing provides actionable data and a more meaningful model of student performance through the school year. The program provides testlet score reports (New Meridian, n.d.) that give educators ongoing feedback throughout the year. MAST is also designed to allow districts to flexibly align testing with local scope and sequence.

Montana received CGSA funding of $2.97 million in 2022 to develop math and ELA through-year assessments for elementary and middle school, and in support of expanding through-year assessments to include science assessments for Grades 6–8, Montana was awarded an additional $4 million in CGSA funding (U.S. Department of Education, n.d.).

North Carolina

North Carolina's IADA application (North Carolina Department of Public Instruction [NCDPI], 2018) proposed a through-year assessment design in which assessments are administered through the school year to improve timeliness and usefulness of test results. In its proposed model, North Carolina proposed to increase the number of items per content standard and report student-level performance on individually assessed content standards. This application produced the North Carolina Personalized Assessment Tool (NCPAT) system.

As part of NCPAT, the state administers the NC Check-Ins 2.0 (NCDPI, n.d.-c), a through-year formative component that provides interim feedback throughout the year. The NCPAT pilot report (NCDPI, 2024) documents progress on NCPAT implementation.

Technical documentation (NCDPI, n.d.-d) and accountability data sets and reporting (NCDPI, n.d.-a) provide a closer look at assessment specifications and student performance outcomes. As part of efforts to encourage data collection and use, North Carolina uses the SAS Education Value-Added Assessment System (SAS EVAAS) tools to capture data on student learning and analyze trends in subgroup performance (NCDPI, n.d.-b).

Implications for Virginia

Growth models provide important complementary information for proficiency measures. When considering using growth models, Virginia will need to consider how student learning growth is captured and how growth metrics are integrated into broader assessment system design. Virginia should consider whether current assessment designs support valid growth measurement; whether through-year approaches would provide faster, actionable feedback; and what educator capacity-building would be needed to effectively use growth data for instructional improvement.

Focus Area F: Application of Knowledge and Higher-Order Reasoning

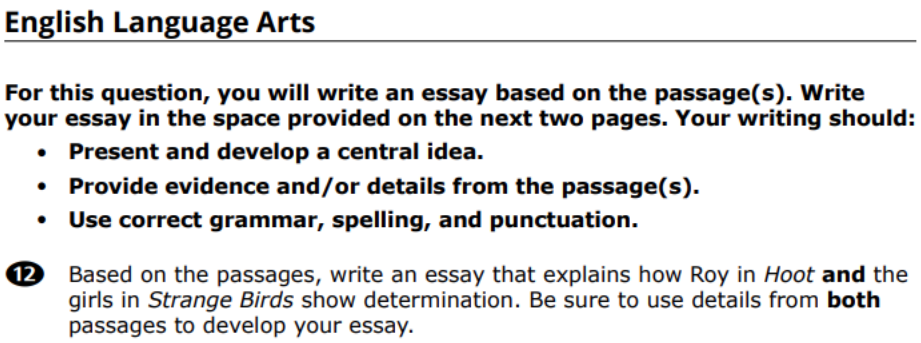
Assessments should measure not only students' recall of facts but also their ability to apply knowledge and engage in complex reasoning. WestEd looked for examples of assessment approaches to capturing higher-order reasoning in systematized ways. This section examines how exemplar states incorporate higher-order cognitive demands into their assessments.

Massachusetts

Massachusetts provides detailed cognitive complexity frameworks for its assessments. For example, the *Cognitive Skill Level Descriptions for ELA MCAS Items* (MDESE, 2023) and parallel descriptions for Science and Technology/Engineering (STE) (MDESE, 2025b) specify how items at different cognitive levels should function.

Released items demonstrate the range of cognitive demands. For example, the Spring 2025 Grade 7 Mathematics release (MDESE, 2025g) includes items requiring multistep reasoning, while ELA releases, such as the Spring 2025 Grade 7 ELA release (MDESE, 2025f), include analytic writing tasks requiring argumentation at Grades 7 and 8 (see Figure B9 for an example of a released analytic writing task from the Grade 7 ELA paper-based test). Science releases include items requiring scientific reasoning, such as items 1, 2, 3, 5, 9, 10, 13, 14, and 18 in the Spring 2025 Grade 5 STE release (MDESE, 2025e).

Figure B9. Released MCAS Grade 7 ELA Item, Focused on Analytic Writing



English Language Arts

For this question, you will write an essay based on the passage(s). Write your essay in the space provided on the next two pages. Your writing should:

- **Present and develop a central idea.**
- **Provide evidence and/or details from the passage(s).**
- **Use correct grammar, spelling, and punctuation.**

12 Based on the passages, write an essay that explains how Roy in *Hoot* **and** the girls in *Strange Birds* show determination. Be sure to use details from **both** passages to develop your essay.

Massachusetts also provides sample student work and scoring guide resources (MDESE, 2025a), showing how student responses at different quality levels demonstrate varying degrees of reasoning and application. The Office of Student Assessment Services provides webinars and recordings for assessment coordinators and administrators to support assessment administration and understanding of scoring expectations.

Smarter Balanced

During test development, Smarter Balanced conducted cognitive laboratories, summarized in the *Cognitive Laboratories Technical Report* (American Institutes for Research, 2013), to understand how students engage with assessment tasks. The consortium's claims/targets framework, accessible through the Content Explorer (Smarter Balanced, n.d.-d) and the Sample Items webpage (Smarter Balanced, n.d.-c), organize assessment content by cognitive demand for ready review.

Resources on scoring (Smarter Balanced, n.d.-b) outline the scaled scores and achievement levels that are reported for students who take Smarter Balanced assessments. The Tools for Teachers platform (Smarter Balanced, n.d.-f) includes lesson plans, professional learning resources, and assessment tools to help educators understand and promote higher-order thinking aligned with assessment expectations.

Implications for Virginia

Virginia should ensure that assessment blueprints specify target distributions of cognitive complexity levels, provide detailed cognitive skill descriptions for item writers and reviewers, and release items demonstrating the full range of cognitive demands that students will encounter. Scored student work samples are particularly valuable for communicating expectations for complex reasoning tasks and providing insight into how student responses will be scored.

Focus Area G: Transparency

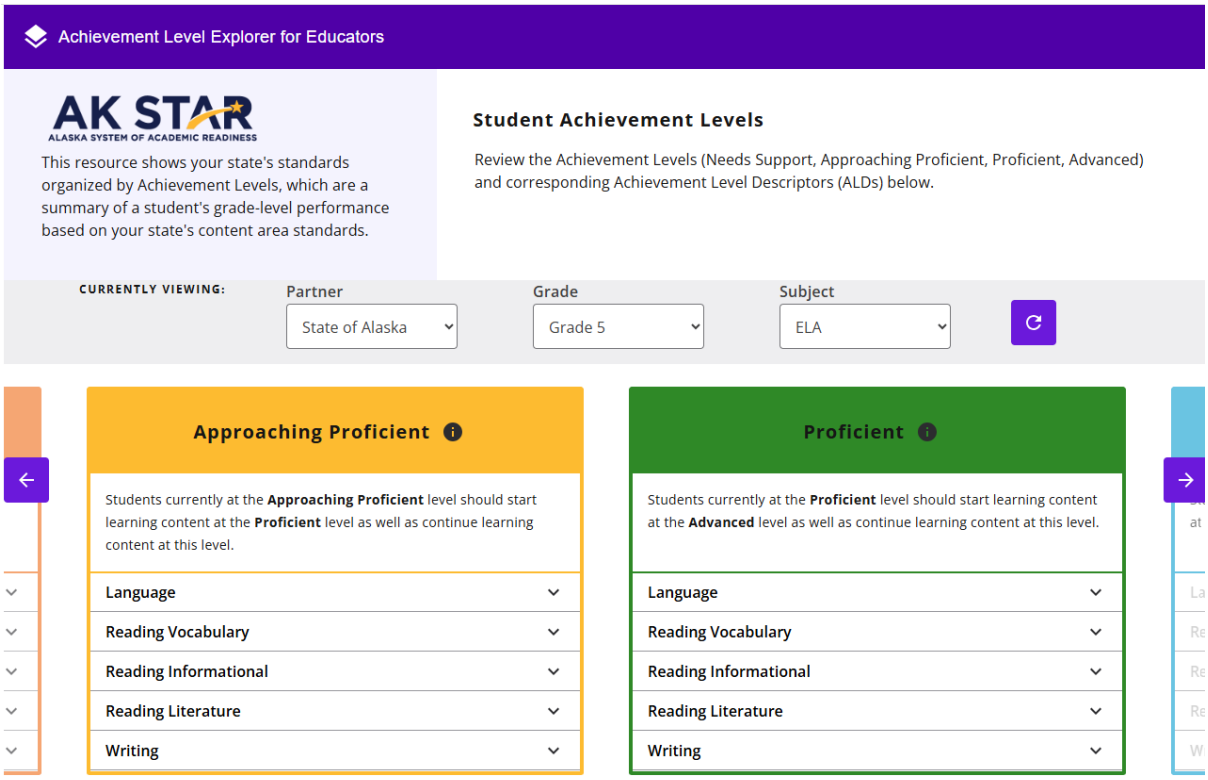
Transparent communication of assessment expectations—through rubrics, performance level descriptors, released items, and exemplars—builds public trust and supports instructional alignment. Transparency may also be demonstrated through reporting templates and processes that focus on ensuring that up-to-date information is consistently shared back in a timely manner. WestEd reviewed how exemplar states communicate assessment information to a wide range of audiences, including the different tools that states build to share current technical information.

Alaska

Alaska provides an exemplary model for assessment transparency. The *Educator Guide to Assessment Results* (Alaska Department of Education & Early Development [DEED], 2024a) explains how to access and interpret data from the Alaska System of Academic Readiness (AK STAR) portal, including details about instructional alignment. The *Family Guide to Student Reports* (Alaska DEED, 2025) communicates results in accessible language for parents and provides clear guidance on how to interpret each report section.

A notable feature of Alaska's model is its interactive Achievement Level Explorer (NWEA, 2022), which allows educators to explore achievement level descriptors by grade for ELA and mathematics. This interactive interface makes expectations easier to navigate than static documents. Figure B10 shows a snapshot of the Achievement Level Explorer interface.

Figure B10. Alaska's Achievement Level Explorer Tool, Grade 5 ELA



The state provides detailed achievement level descriptor documents for each grade band and subject area, such as Grades 3–5 ELA (Alaska DEED, 2023). Detailed scoring rubrics are not publicly available.

The spring 2025 practice tests (NWEA, 2026) provide students with a tool to familiarize themselves with assessment format. Alaska's *Educator Guide to Student Readiness* (Alaska DEED, 2026) also supports student preparation by providing educators with an overview of assessment tools and functions, as well as guidance for how to share assessment instructions with students. The state budgets $650,000 for statewide assessment support (Alaska DEED, 2024b), implemented through a vendor partnership with NWEA.

Massachusetts

Massachusetts provides extensive released materials. As previously described, released item sets are available for all grades and subjects, such as Grade 7 ELA (MDESE, 2025f), Grade 7 mathematics (MDESE, 2025g), and Grade 5 STE (MDESE, 2025e), including both paper- and computer-based versions, and are accessible through the MCAS Resource Center (Cognia, Inc., 2025). Published rubrics and a guide to using sample student work (MDESE, 2025h) outline expected student performance at different levels.

As part of efforts to develop assessment items that capture higher-order thinking, Massachusetts applied for and received CGSA funding for computer-simulated science performance tasks (MDESE, 2020b) and an IADA award to pilot innovative science assessments (MDESE, 2020a).

Smarter Balanced

Smarter Balanced emphasizes transparency as a core principle. Practice tests and sample questions (Smarter Balanced, n.d.-a) are publicly available for all grades. Achievement level descriptors and scoring information (Smarter Balanced, n.d.-b) communicate achievement expectations and provide guidance on how to interpret scores to understand student performance. Professional learning resources through Tools for Teachers (Smarter Balanced, n.d.-e) support educator understanding of assessment content and scoring.

Implications for Virginia

Investment in transparency supports both public trust and instructional improvement. Building resources that outline assessment design and what students are being asked to demonstrate increases confidence in assessment tools. Virginia should consider developing interactive tools for exploring expectations (like Alaska's ALD Explorer), maintaining comprehensive released item libraries with scored student work, and providing clear communication materials for multiple audiences (educators, families, students).

Focus Area H: Innovative, Reliable Scoring and Validation

Part of the complexity of assessment systems is consideration for scoring processes to ensure that scores are valid, reliable, and accurate. Effective scoring systems need to balance efficiency, cost-effectiveness, and measurement quality. This section presents examples of states using automated scoring implementations, human scoring protocols, and hybrid approaches.

Smarter Balanced Distributed Scoring

Smarter Balanced has implemented innovative scoring approaches including AI-assisted scoring. In collaboration with IBM, Smarter Balanced developed student-centric principles to guide how AI is used to inform large-scale assessment design (Smarter Balanced & IBM, 2025).

The consortium's Reliability, Precision, and Errors of Measurement section of the *2023–24 Interim Assessment Technical Report* (Smarter Balanced, 2024) provides evidence of scoring quality across the system. The hybrid human/AI approach demonstrates how technology can enhance, rather than replace, human judgment in scoring complex responses.

Louisiana Teacher-Moderated Scoring

Louisiana's LEAP 2025 technical reports document scoring procedures including inter-rater reliability evidence; for example, see Section 6, page 55, of the 2023–24 Grades 3–8 Social Studies technical report (DRC et al., n.d.). The scoring model includes QC processes generating reports that show response distributions to inform calibration.

Online training tools (DRC, 2026), available for ELA, mathematics, science, and social studies, provide educators with guidance and tools for scoring. Although detailed teacher moderation protocols are not publicly available, the technical reports indicate attention to scoring quality.

Florida Automated Scoring

Florida's CAT system enables automated scoring for selected-response items, supporting the rapid 24-hour reporting timeline. The K–12 *Statewide Assessments Guide* (FLDOE, 2026) references automated scoring procedures, though detailed validation documentation for automated systems is not prominently published.

Implications for Virginia

Virginia should consider how scoring approaches align with reporting timeline goals. Automated scoring enables rapid turnaround but requires validation evidence. Hybrid human/AI approaches may offer a balance of efficiency and quality for complex items. Any scoring system should have documented reliability evidence, QC processes, and validation studies.

Focus Area I: Reporting to Students, Families, and Educators

Timely, actionable reporting is essential for assessment results to be an actionable part of instructional responses. Reporting also supports state efforts to maintain transparency with partners. WestEd sought examples of state reporting systems designed to provide valuable and timely data to students, families, and educators.

Florida Family Dashboards

Florida's Family Portal provides exceptional reporting capabilities. Key features include scores populated within 24 hours of test completion during testing windows, box-and-whisker plots showing student performance relative to standards, and enhanced achievement level descriptions. Detailed individual student reports provide benchmark-level information that teachers can use to identify instructional needs. The *Family Portal Access Codes Quick Guide and FAQ* (FLDOE, 2024a) explains how families can access and interpret assessment results, while *Understanding Florida Statewide Assessments Reports for Families* (FLDOE, 2025c) provides interpretive guidance.

The *Understanding Florida Reporting System (FRS) Reports* guide (FLDOE, 2025b) helps educators interpret teacher-level reports by providing guidance for interpreting achievement levels and scale score ranges. Florida's 2009 SLDS grant ($10 million) focused on expanding reporting capabilities (United States Department of Education, 2009).

Massachusetts Next-Generation Reports

Massachusetts provides sample reports in its technical report appendices (MDESE & Cognia, 2024) and in its resources for parents/guardians (MDESE, 2025c) to provide them with guidance for interpreting the different sections of scoring reports.

The April 26, 2024 version of the News from DESE Student Assessment Services newsletter (MDESE, 2024) communicates reporting timelines. Massachusetts received a 2015 SLDS grant of $26 million to support data reporting and analytics (NCES, 2015).

Utah RISE Reports

Utah's Student Proficiency Results dashboard (Utah State Board of Education, n.d.) provides public access to assessment data. However, Utah's documentation of student, family, teacher, and principal reports is limited in public sources, compared to Florida and Massachusetts.

Implications for Virginia

Florida's model demonstrates what rapid, multi-audience reporting can look like, while Massachusetts and Utah provide examples of reporting that is responsive and accessible to different audiences. Efforts to communicate assessment data and how to understand implications related to student learning support broader transparency and communication efforts. Virginia should consider what reporting timeline is feasible, given assessment design; how reports can be

differentiated for various audiences (students, families, teachers, administrators); and what interpretive resources will help stakeholders use data effectively.

Focus Area J: Reporting to State Education Agencies

Assessment data must also flow back to the state education agency (SEA) in formats that support accountability, resource allocation, and targeted improvement. The following examples illustrate how states have designed data systems for agency use.

Florida

The Florida Automated System for Transferring Educational Records (FASTER) (FLDOE, n.d.-a) provides data formats and standards for statewide data collection and reporting. Detailed SEA dashboard documentation is not publicly prominent, but the state's through-year testing model with rapid reporting suggests that data are available to the SEA in near-real-time.

Smarter Balanced Reporting APIs

Smarter Balanced provides data formats and standards documentation including guidance on score types (Smarter Balanced, 2025). The consortium's multistate implementation means that reporting infrastructure must support diverse state systems; however, detailed API documentation is not prominently published.

Tennessee Education Information System

Tennessee's Education Information Systems (EIS) (TDOE, n.d.-b) is a state-level data system that collects and reports demographic and attendance data (Institute of Education Sciences, n.d.). The Education Information Systems website (TDOE, n.d.-b) provides guidance for districts to compile and transmit student data to EIS. Certica Solutions/Instructure supports the system.

Implications for Virginia

As part of logistical planning, Virginia will need to consider how assessment data are transmitted between districts and VDOE for collection and reporting purposes. These transfer processes will need to be responsive to inform accurate reporting. Virginia should consider how assessment data reporting is integrated into existing state data systems, what real-time or near-real-time data capabilities would support timely intervention, and how data infrastructure investments align with assessment modernization goals.

Appendix C. Federal Assessment Requirements Under ESSA

This Appendix summarizes the federal requirements under Every Student Succeeds Act (ESSA) that any modernized Virginia assessment system must satisfy. These requirements establish the compliance floor; no design option presented in this report may fall below them. State statutory requirements under HB 1957 and SB 200, which in several cases exceed the ESSA requirements, are described in Section II.B of this report.

Table C1. ESSA Assessment Requirements

Requirement	ESSA Provision	Implication for Virginia
Annual Testing	States must administer assessments in ELA and mathematics annually in Grades 3–8 and at least once in Grades 9–12, and in science at least once in each of three grade bands (3–5, 6–9, and 10–12).	Virginia currently meets this requirement through its SOL assessment program. Any redesigned system must, at minimum, maintain this testing schedule. Virginia exceeds this floor with additional EOC assessments in history/social science.
Alignment to State Standards	Assessments must be aligned to the state's challenging academic content standards and academic achievement standards.	Virginia's SOL assessments must align to the current SOLs. As revised standards are adopted (e.g., mathematics in 2023, ELA in 2024), assessment blueprints and items must be updated accordingly.
Technical Quality	Assessments must meet professional standards for validity, reliability, and technical quality. States must submit evidence through the federal peer review process.	Any new assessment system must be designed with peer review requirements in mind from the outset. Peer review evidence packages must be assembled and submitted following implementation. Through-year or innovative designs face additional scrutiny under peer review.
Accommodations	Assessments must provide appropriate accommodations for students with disabilities and English learners, including students with the most significant cognitive disabilities, through alternate assessments (capped at 1% participation).	Virginia must maintain its alternate assessment program (VAAP) and ensure accommodations for all student populations. Universal design principles should be embedded in new item and platform development.
ELP Assessment	States must annually assess the English language proficiency of all English learners in the four language domains (listening, speaking, reading, writing).	Virginia currently meets this requirement through participation in the WIDA consortium. This requirement is separate from the statewide summative assessment contract.

Requirement	ESSA Provision	Implication for Virginia
Disaggregated Reporting	Assessment results must be reported in a disaggregated and accessible format, including by race/ethnicity, gender, disability status, English learner status, migrant status, and economic disadvantage.	All score reporting, whether through the current system or through a modernized system, must support disaggregation across all ESSA-required subgroups. Reporting systems must be designed to produce these breakdowns at the student, school, division, and state levels.
Participation	States must assess at least 95% of all students and 95% of each subgroup.	Virginia must monitor and enforce participation rates. Assessment design decisions (testing time, accessibility, scheduling) affect participation compliance.

Table C2. ESSA Accountability System Requirements

Indicator	ESSA Provision	Implication for Virginia
Academic Achievement	The academic achievement indicator is based on state assessment results aligned to academic standards. Must meaningfully differentiate school performance.	SOL assessment results feed directly into the School Performance and Support Framework. Assessment redesign must preserve the ability to classify schools.
Other Academic Indicator (non–high schools)	For elementary and middle schools, the state must include either a measure of student growth, if determined appropriate by the state, or another valid and reliable statewide academic indicator that allows for meaningful differentiation in school performance. For high schools, student growth may be included at the state's discretion but is not required.	Virginia's choice of whether and how to measure individual student growth (see Section V.B, Enhanced tier) is both an assessment design question and an accountability question.
Graduation Rate	ESSA requires states to use a 4-year adjusted cohort graduation rate, with an option to include an extended-year rate.	Not directly affected by assessment redesign, though changes to verified credit requirements and EOC testing could have indirect effects on graduation rates.
ELP	The ELP indicator measures progress toward English language proficiency for English learners.	Measured through WIDA ACCESS. Separate from the summative assessment contract but relevant to how English learner performance is reported alongside SOL results.

Indicator	ESSA Provision	Implication for Virginia
School Quality or Student Success	At least one additional indicator of school quality or student success (e.g., chronic absenteeism, school climate, access to advanced coursework).	Virginia uses chronic absenteeism as this indicator. This indicator is not directly affected by assessment redesign, though performance task or portfolio components in higher tiers could eventually inform it or other school quality or student success indicators.

ESSA Peer Review Requirements

Federal peer review is the primary mechanism through which the U.S. Department of Education evaluates whether a state's assessment system meets ESSA requirements. Virginia's modernized system will be subject to peer review, and the design choices made during procurement will determine whether the system can pass peer review. Table C3 shows the areas that reviewers evaluate.

Table C3. ESSA Peer Review Requirements

Peer Review Area	What Reviewers Evaluate
Statewide System	Single statewide system with consistent standards, assessments, and technical quality across all districts. Coherent system of standards, assessments, and accountability.
Assessment Quality	Alignment to content standards, coverage of the full depth and breadth of standards, appropriate item types, and adequate reliability and validity evidence, including classification accuracy and consistency.
Fairness and Accessibility	Evidence that assessments are accessible to all students, that accommodations are appropriate and do not alter the construct measured, and that differential item functioning and bias reviews are conducted.
Technical Documentation	Comprehensive technical reports documenting psychometric properties, equating/scaling procedures, standard setting, and evidence of comparability across forms, modes, and administrations.
Scoring	Scoring procedures, QC, inter-rater reliability for human-scored items, and validation evidence for any automated scoring. Scoring must produce consistent, accurate results.
Reporting	Timely, accessible, disaggregated reporting to parents, educators, and the public. Reports must include individual student performance, achievement levels, and interpretive information.
Inclusion	Evidence of appropriate inclusion of students with disabilities and English learners, including documentation of accommodation policies and alternate assessment participation rates (1% cap).

Note on Innovative Assessment Designs

States pursuing assessment designs that depart from a single annual end-of-year summative assessment (such as through-year models, competency-based pathways, or locally administered performance assessments used for federal accountability) may need to obtain approval through IADA or a Section 8401 waiver. The Core and Enhanced tiers described in Section V do not require federal approval. The Transformative tier would require either IADA or waiver authority before any pilot results could replace the statewide summative for federal accountability purposes.

APPENDIX D. WHAT WE HEARD: STAKEHOLDER FEEDBACK ON VIRGINIA INNOVATIVE ASSESSMENT

WHAT WE HEARD: STAKEHOLDER FEEDBACK ON VIRGINIA INNOVATIVE ASSESSMENT

March 2026

This summary report highlights key themes and stakeholder feedback on a modernized Standards of Learning (SOL) assessment system. This report

- documents participation engagement activities;
- summarizes feedback to inform the modernized SOL system; and
- provides the Virginia Department of Education (VDOE) with a communications-ready executive summary for the Board of Education, legislative committees, and the public.

The goal of this broader statewide engagement is to build public understanding and identify implementation concerns early, when they can still inform design decisions, rather than derail implementation.

Prepared by

Quality Information Partners

This work was funded by the Virginia Department of Education as part of the Virginia Innovative Assessment Consultant contract (201-26-028) to WestEd

CONTENTS

Executive Summary ***105***
Key Findings and Cross-Cutting Themes Across Stakeholder Groups ***106***
Information Gaps and Communication Needs 106
Curricular and Developmental Appropriateness 108
Platform Operation Constraints 110
Innovation and Alternative Assessments 111
Systemic Barriers 114
Accountability Factors 116
Recommendations ***117***
Methodology ***118***

EXECUTIVE SUMMARY

Overview

As part of the Commonwealth's desire to analyze options for implementing an innovative new statewide assessment system to support high-quality teaching and learning, Virginia Department of Education (VDOE) contracted with WestEd and Quality Information Partners (QIP) to capture stakeholder perspectives on the modernization of the Standards of Learning assessment system. This outreach effort was designed to ensure that future assessments remain rigorous, fair, and instructionally useful.

Approach

The team conducted 13 focused engagement sessions with 65 participants across all regions of Virginia in January 2026. Respondents included state- and division-level leadership; school administrators and classroom educators; and parents and community-based organizations.

Key Findings

1. Strengthening Communication and Instructional Alignment: Stakeholders desire enhanced transparency about test design and mechanics. They want structured design documentation to better align daily classroom instruction with state expectations and reduce burden.

2. Enhancing Data Utility and Reporting Accessibility: Stakeholders emphasized the importance of quick data turnaround to inform summer interventions and start-of-year placements. Parents desire reports that contextualize a student's performance relative to their peers and growth trajectories.

3. Optimizing Developmental and Curricular Appropriateness: Participants noted that the length of testing sessions can exceed the physical stamina of younger learners and students with specific learning needs, potentially compromising their demonstration of academic knowledge. Language complexity can be a barrier for English learners or students with low reading skills.

4. Operational Efficiency and Technological Integration: While the transition to digital-only formats offers significant advantages, stakeholders identified a "digital divide" where device navigation skills impact performance. There is a call for better alignment between state assessments and external diagnostic tools to reduce test-taker confusion.

5. Balancing Innovation with Academic Rigor: There is a nuanced debate about the future of assessment formats. While there is interest in "tech-enhanced" items and simulations, any shift toward performance-based assessments should address concerns regarding administrative burden, local scoring validity, and the potential impact of artificial intelligence. Critical thinking and transferable skills should remain anchored in a solid foundation of factual knowledge and procedural fluency.

6. Systemic Stability and Accountability Frameworks: Stakeholders are concerned about the high-stakes nature of verified credits required for graduation. Division leaders seek a return to structured, multiyear assessment cycles to improve instructional alignment. There is a desire for accountability systems to be an incentive for resource allocation rather than a punishment for poor performance.

Recommendations and Next Steps

This candid feedback is valuable to VDOE's continuous improvement process. By improving communications and technical assistance to districts, VDOE can better support divisions in implementing any changes and improving student learning.

KEY FINDINGS AND CROSS-CUTTING THEMES ACROSS STAKEHOLDER GROUPS

As part of the Commonwealth's desire to analyze options for implementing an innovative new statewide assessment system to support high-quality teaching and learning, Virginia Department of Education contracted with WestEd and Quality Information Partners (QIP) to capture stakeholder perspectives on the modernization of the Standards of Learning (SOL) assessment system. This outreach effort was designed to ensure that future assessments remain rigorous, fair, and instructionally useful.

In December 2025, QIP contacted over 115 individuals across the Commonwealth, inviting them to participate in a 1-hour virtual focus group to share their feedback on the current SOL assessments. In January 2026, QIP hosted 13 focus groups with 65 participants. Key stakeholder groups included Virginia Department of Education (VDOE) state staff, division leaders, school administrators, teachers, assessment coordinators and administrators, parents, and community-based organizations (CBOs).

The following sections represent the direct perspectives of focus group participants.

Information Gaps and Communication Needs

Addressing Persistent Stakeholder Misconceptions

- During focus groups with teachers and administrators across the Commonwealth, stakeholders shared recurring misconceptions about SOL assessments. The following misunderstandings are detailed further in the report:
    - Concerns about developmental appropriateness.
    - Beliefs that tests are solely multiple choice.
    - Confusion about computer adaptability and grade-level content.
    - Beliefs that tests are not skills-based.

> "I work directly with the questions and use the standards to create quizzes, games, and questions to prepare for the tests. This is tricky because the state has released almost no content and the standards are difficult to interpret. Teachers and I find it hard to create helpful questions for students."
>
> **Founder, Community-based organization**

These misconceptions persisted across regions and roles, and could easily be addressed through clear, transparent communications with schools and divisions.

Instructional Clarity and Design Documentation for Teachers

- Teachers and school administrators need clearer, more predictable test design documentation to align instruction with state expectations. Many suggested adopting a structured model (some mentioned the New York Regents as an example) to provide a consistent testing schedule and clear content expectations, moving away from the current approach where major changes are rolled out with little preparation time. Educators also

need training on the specific "way of thinking" required by new assessment formats, such as integrated reading and writing.

The Impact of Data Timeliness and the "Data Lag"

- The timing of test result delivery is highly inconsistent across the Commonwealth, with turnaround times varying from 1 week in some divisions to several months in others.
- CBOs and parents highlighted a pervasive "data lag," where results often arrive after the end of the school year. This delay renders the data useless for immediate instructional adjustments and prevents parents from securing necessary summer tutoring or intervention when it is most impactful.

> "Results in ParentVue post fast, but if you don't look, you are not going to know it is there. We do get them by mail, but it is several weeks later. You would know if a kid didn't pass because they have to retest them. But if they pass you do not know."
>
> **PTA Member**

- This lag also affects the transition between grade levels, as the following year's teachers rarely receive data on their incoming students in a helpful or timely manner. As a result, schools often cannot use the previous year's end-of-term state data, preferring to rely on alternative fall benchmarks to determine student placement and support needs.

Parental Literacy and Accessibility of Reports

- Parents have gaps in understanding the purpose and mechanics of state assessments, particularly modern Computer Adaptive Tests (CATs). Many do not realize that adaptive tests present items above grade level to identify a student's "ceiling," which leads to confusion and missed opportunities for the state to communicate the benefits of these models.
- Current reporting methods are described by parents and CBOs as "wonky," overwhelming, and poorly timed. Parents want more visually intuitive reports, such as those used in Maryland, including graphs that show how a child scores relative to peers, rather than only a numerical score. There is also demand for reports and graphs to be translated into multiple languages so that all parents, especially those of English learners (ELs), can engage with their child's data.
- Stakeholders debated whether current scoring scales are intuitive, with many suggesting a 100-point scale would better align with familiar classroom grading.

> "We are a black hole... We need transparency for teachers, parents, and others to understand results. I support a 100-point scale. No one understands the current scale."
>
> **Teacher**

Curricular and Developmental Appropriateness

Test Length and the Measurement of Student Stamina

- Participants in almost every focus group, particularly teachers and school administrators, expressed concern that SOL assessments are too lengthy and complex for young learners' cognitive and physical abilities. Third grade assessments are seen as particularly inappropriate because students are often "defeated" by the visual size of reading passages before beginning their tests.
- Participants observed that testing has transitioned from a measurement of subject knowledge to a measurement of a student's "ability to focus for an entire day." For students with attention-deficit/hyperactivity disorder (ADHD) or executive functioning difficulties, "day-long" tests are seen as a major barrier. These students often receive daily instruction in small, manageable increments, but the standardized testing environment requires "stamina" for which they are not developmentally or instructionally prepared.

> "Four very long passages and 33 questions means it takes hours for a 3rd grader to take the test. This was an effort to increase rigor, but is not considered developmentally appropriate."
>
> **Elementary Assistant Principal and School Testing Coordinator**

- Stakeholders repeated misconceptions around the developmental appropriateness of reading tests for younger elementary students, specifically sharing concerns about the length of passages and the number of questions for each passage. Teachers seemed unaware that their fellow Virginia teachers and administrators review each item to ensure that it is developmentally appropriate and measures student knowledge fairly. While debate continues regarding test length, the individual items themselves have undergone careful vetting to ensure appropriateness.

Uneven Distribution of Testing Burden

- Stakeholders feel the testing load across grade levels is unbalanced, with 5th and 8th graders being disproportionately burdened. This high volume of testing creates a "grind" that disrupts the entire school environment, causing stress even for students in grades that are not being assessed.
- To manage uneven testing loads, some divisions have shifted tests like Virginia Studies to 4th grade to balance requirements across the elementary years. However, participants noted that this shortens the subject's available instructional time and does not resolve the underlying issue of "testing volume."
- Teachers and school administrators specifically described the 8th grade schedule of five separate SOL tests in a single year as "brutal" and "cruel." The structure of specific subject-area tests, particularly science, is a major driver of remediation needs and

instructional stress. In 8th grade, the science SOL covers 3 years of content, which many feel is too much material for a single high-stakes assessment.

> "Want to see more balance across grade levels. 4th, 5th, 8th, and 9th grade students get hit hard on assessments. 4–5 assessments [in one school year] is too many. It would be helpful to spread among grade levels"
>
> **Division Instructional Coach**

- High schoolers face a different but equal burden, as they may be required to take the same SOL test many times to satisfy graduation requirements, leading to student burnout. As a result, students are frequently pulled from multiple core class periods a day for past subject remediation, further limiting current instructional time.

Reading Skills as a Barrier to Subject Mastery

- Parents, teachers, and state leaders share a primary concern that math and science assessments are too "word-heavy," effectively assessing reading comprehension and English language proficiency rather than the intended content knowledge. Participants noted that non-mathematical words and complex names often distract students, impeding their ability to solve actual math problems. This creates a "no-win" scenario: a student may have mastered the scientific or mathematical concept, but fails the test due to complex language used in the test question.
- This structural barrier is particularly detrimental to students with low reading levels (including those with reading-related disabilities) and ELs, where assessments often end up measuring English proficiency rather than subject-matter mastery. As a result, these students are unable to demonstrate their knowledge, even when they have achieved mastery, because they lack the necessary reading skills. One teacher highlighted that past "plain English" math assessments were more effective because they isolated math skills from language proficiency. Their removal has majorly disadvantaged below-level readers, ELs, and special education populations.

> "Make sure it tests what it is supposed to test instead of vocabulary and English. Use simple names like "John" and "Sarah." More culturally responsive names can throw students off. Students get caught up in the vocabulary—they have no idea what a quiche is. Make it pizza. Stop putting stuff in there that the kids may never have heard of. We are taking it away from the assessment and making it about culture."
>
> **Teacher**

Platform Operation Constraints

Testing Format Burden and Technical Friction

- Digital-only formats introduce a "digital divide," where tests may inadvertently measure a student's "computer literacy" rather than subject mastery. Teachers and school administrators have to redirect significant instructional time toward teaching "device skills" rather than academic content. Online testing drives hardware dependency for students as young as 1st grade, which division leaders note disrupts early childhood learning environments.

> "The fact that all testing is on iPads drives so much technology use. It is the primary justification for using iPads heavily in the division. It takes so much time teaching students how to use the test and then it is not tied closely to teaching and learning. It is a big driver and a major concern because we now give every student an iPad starting in 1st grade."
>
> **School Board Vice Chair**

- The "technological interface friction" introduced by digital tests requires students to navigate long passages and multiple questions through lengthy scrolling and complex iPad interfaces, which can lead to disengagement and exhaustion. Some participants observed that the results of students with disabilities or from underserved backgrounds dramatically improve in paper-and-pencil format. For example, one teacher noted that students solve a problem correctly on scratch paper but fail to input the answer correctly into the device.
- The lack of standardization between different testing platforms, specifically the Measures of Academic Progress (MAP) and the SOL, confuses students. Interface mechanics (such as how to highlight text or perform "drag-and-drop" actions) vary, leading teachers to spend extra time re-teaching different navigation tools for each specific assessment.
- From an administrative perspective, "back-end" platform management is tedious and manual. Division staff expressed frustration with the Pearson system and identified uploading student data and sorting testing tickets as labor-intensive. State leaders identified a need for technical metadata such as "time on test" per item and "response change" data to help them identify whether students struggle due to knowledge gaps or testing fatigue.

Adaptive Design and Test-Taking Strategies

- Participants consistently preferred external diagnostic tools like MAP over the state-mandated Virginia Growth Assessments (VGA). While MAP provides actionable, "real-time" data that allows for immediate instructional shifts, the state-mandated VGA is often viewed as "ineffective" because it fails to provide a clear diagnostic of a student's academic standing.

- Teachers often misunderstood how CATs function. Because CAT and MAP assessments are designed to measure growth by finding a student's "ceiling," they intentionally include items above grade level. Some teachers mistakenly think that these assessments function differently, and they instruct students to use different strategies for each assessment type: guess on difficult items in growth assessments, and aim for accuracy when taking SOLs. Since both assessments have similar adaptive features and can include items above grade level, these conflicting strategies confuse and overwhelm students, compromising their performance on high-stakes grade-level testing.
- Educators identified a critical need for "gap analysis" reports tied to specific SOLs. The nature of CATs prevents teachers from viewing specific question stems, which makes it impossible to determine whether a student struggled with academic content or the question's language or technical structure. This lack of granularity hinders targeted remediation and reinforces the preference for MAP, which participants believe correlates better to standards and more accurately tracks a student's position on skill continuums.

> "It is hard to match assessment data to the standards. It is vague, so people don't use it. We need to also know what the question stem looks like. Looking at the wrong answers can tell misconceptions in thinking."
>
> **Division Reading Coach**

- The "opacity" of scoring systems and psychometric scales is very frustrating for school division staff and parents. Teachers described VGA scores (such as a raw number like "1453") as "essentially meaningless" because there is no accompanying scale or benchmark provided for interpretation. Further confusing the issue is the fact that students in the same grade take different versions of the same adaptive test, making it difficult for families to compare results or understand progress relative to peers. While MAP's Rasch UnIT (RIT) scores require some explanation, they are perceived as a clearer and more reliable indicator of academic standing.

Innovation and Alternative Assessments

Balanced Interest in Reliable Innovation

- There is potential interest in modernizing tests, but many stakeholders have "limited appetite" for radical redesigns, prioritizing fairness and scientific rigor over novelty. Division and state leaders warned that "innovation" can sometimes be a "euphemism for lowering academic standards" or a way to "pad stats" by passing students who have not met requirements.

> "We want assessments that measure what they are supposed to measure."
>
> **Assistant Superintendent**

- Many participants argued that the state should focus on remaining "equitable and fair" rather than simply "chasing a buzzword." Several groups were in consensus that any new assessment model should be "purposeful, cost-effective, and subject-appropriate" rather than universally applied, as high-tech simulations might be suitable for high school chemistry but unnecessary for 3rd grade math.

- For some, "innovation" is best realized through "tech-enhanced" item types (such as equation editors, graphing tools, and simulations), rather than a fundamental change to the assessment structure. These participants believe the current system is "not broken" but requires incremental "tweaking" and additions to remain on the cutting edge while minimizing "change fatigue."
- Across focus groups, parents, community members, and some educators characterized the SOLs as exclusively multiple-choice exams. This belief that tests are solely multiple choice overlooks the inclusion of several innovative item types already embedded in the assessments. In many cases, participants expressed a desire for question formats that already exist within the current assessments.

Tension Over Performance-Based Assessments (PBAs) and Portfolios

- Classroom-level stakeholders and state leaders hold disparate views on the use of PBAs.
    - Proponents, including many teachers and parents, view PBAs and portfolios as a "robust and impressive" way to show student growth through a "body of work" rather than a single-day snapshot. They argue these formats increase student engagement and allow those receiving special services more opportunities to demonstrate mastery. While teachers desire PBAs, they are currently "used to teaching multiple choice" and would require extensive professional development to implement alternative assessments effectively.
    - Conversely, division and state leaders have major reservations about the practicality and validity of PBAs. These leaders highlighted that local scoring of performance tasks creates a "significant administrative and labor burden" that many divisions are not equipped to handle. Furthermore, some describe state-level portfolios as a "mess" that is "not legitimate" for the purpose of state-wide standardized scaling. These leaders are more focused on having fair, reliable tests supported by sufficient preparation resources.
- Concerns regarding the authenticity of student work were also raised, particularly with the rise of artificial intelligence (AI). One participant cautioned that AI makes it increasingly difficult

> "Tests are not motivating and engaging for students. They hate them. All SOL tests should have some type of authentic and engaged assessments (like portfolios and project-based learning) to help with engagement and skills-based development. Allow students to submit these authentic assessments as part of their overall SOL grade as a way to get students to enjoy learning and make it relevant and meaningful for them."
>
> **Education Researcher**

> "My concerns with non-traditional assessment formats (like portfolios) is that it is very time-intensive. This results in more time assessing and preparing to assess rather than instructing. The portfolio process benefits English, but in social studies it is hard to manage within the class because it is labor-intensive."
>
> **Instructional Supervisor**

to verify student work in non-standardized formats.

The Debate of Factual Knowledge vs. Transferable Skills

- There is sharp disagreement over whether the primary goal of state assessments should be the recall of factual content or the demonstration of transferable skills. This was cited by community-based organizations as the "strongest point of disagreement" among their participants.
    - Some argued that assessments should track a student's movement along a "skill continuum" rather than their memory of specific historical facts.
    - On the other hand, many participants insisted that skills cannot be detached from facts, noting that "critical thinking is impossible without a foundation of factual knowledge." State leaders echoed this, suggesting that while critical thinking is a vital goal, it requires "procedural fluency" and a solid factual base to be meaningful.

> "Critical thinking is application of known facts to new situation; you cannot do that without solid facts."
>
> **School Board Member**

- Subject-specific needs further complicate this balance. Participants noted that math and science may be better suited for factual and process-based testing, while social studies and English might benefit more from skill-based evaluations. This tension suggests that a universal approach to the skills-vs-facts debate may be inappropriate across different academic disciplines.
- Some teachers shared the incorrect belief that tests are not skill-based and SOL questions assess only factual recall. However, the assessments include items that require applied skills—such as mapping tasks, science simulations, and multi-step items that require students to integrate and apply knowledge, not simply restate facts.

The Loss of Modified Assessments

- Participants across leadership groups identified a "big gap" left in the assessment system following the removal of the Virginia Modified Achievement Standards Test (VMAST). This assessment was designed for students who do not have "significant cognitive disabilities" and therefore do not qualify for the Virginia Alternative Assessment Program (VAAP), but who still struggle to demonstrate their knowledge on the standard SOL tests.

> "Special credit accommodations put the burden on the teacher to show what the students did in class. We want the burden on the students. Alternative tests cannot only include ACT IIs and AP; those don't help special education students."
>
> **Secondary Supervisor of Special Education**

- Participants described the VMAST as innovative because it provided supports such as formula boxes or highlighted variables to assist students with memory or recall issues without providing an unfair advantage. Without these modified options, these students with disabilities who do not qualify for

the VAAP are "forced" to take standard SOLs that do not meet their instructional needs, often leading to a "cycle of failure."

- The loss of this "2 percent test" (referring to the small but notable number of students served) is seen as a structural barrier to equity. Leaders argued that "innovation" should include the return of "specially designed assessment" tasks that mirror the "specially designed instruction" students receive in the classroom, allowing for a more accurate measure of their academic achievement.

Systemic Barriers

Resource Disparities and the "Uneven Playing Field"

- A sharp divide exists between high-resourced and under-resourced divisions, creating what participants describe as an "uneven playing field" for students. In affluent areas, students often enter school with more resources and a higher baseline of vocabulary exposure, whereas in high-needs schools, education often moves to the "backburner" as families struggle with basic needs and "burnout."

> “The disparities in resources across the Commonwealth is a barrier. Some schools have no coaches; others have 15. There is so much disparity. How can we give them all the same assessment? There is a whole ecosystem beyond just student and teacher that is influencing student outcomes.”
>
> **Former Elementary School Principal**

- Wealthier divisions have the financial and staffing capacity to fill state-level gaps by creating their own supplemental practice items, targeted instructional materials, and custom data reports. In contrast, smaller or lower-income divisions may struggle with fundamental requirements like consistent internet access or departments with a "team of one," leaving them entirely reliant on state-provided outputs.
- Disparities extend to community-level support, where wealthier Parent Teacher Associations can provide resources or professional advocacy. These advantages are unavailable to Title I schools with high percentages of children from low-income families, further widening the gap in how students experience high-stakes testing.

Equity Disparities and the "One-Size-Fits-All" Conflict

- Participants across all groups concluded that a one-size-fits-all assessment model fails ELs and students receiving special education services. In particular, there is a "deep worry” that these students may be unfairly impacted by rigid requirements, including for graduation.

- The mismatch between daily instruction and the standardized assessment format creates tension. While teachers are required to differentiate instruction all year, the state assessments are inflexible, offering little "wiggle room" for students who require hands-on interaction or simpler language to demonstrate their knowledge. Participants called for "specially designed testing" that aligns more closely with individualized supports provided in the classroom.

> "Lots of special education students really struggle, especially in reading. We have lots of accommodations in the classroom, but they are limited during assessments. Students with multiple disabilities are particularly struggling with assessments and end up with mental fatigue."
>
> **Special Education Teacher**

- Parents and teachers have observed conflicts between school accreditation requirements and the individualized needs of vulnerable populations, including students receiving special education accommodations. Division leaders reported that pressure from the Every Student Succeeds Act and accreditation often leads to a conflict of interest, where building-level administrators may override Individualized Education Program (IEP) team decisions regarding testing accommodations in hopes of securing a passing score.

Structural Barriers and Internal State Erosion

- Division leaders and state-level staff report that inconsistent and non-transparent communication from VDOE has undermined morale and created organizational silos. Division leaders report conflicting guidance from different offices, while state-level staff feel pressured to implement changes without adequate information, negatively impacting data quality and technical interpretation.

> "VDOE Offices don't always work together or talk to each other. One VDOE Office might tell us one thing, but another office says something completely different. People do not buy in because they know it will change. If something would hold still, it would be great."
>
> **Division Leader**

- Centralized, multi-layered approval processes have created timing constraints that hinder division staff's ability to support schools and students. For example, critical instructional tools and crosswalks are often released after testing windows close, rendering them "useless" for the teachers and students who needed them for that academic year. Educators noted a trend of resources labeled "forthcoming" indefinitely and requested firm release dates for resources so they can properly align instruction.
- Divisions have lost access to in-person and virtual professional development sessions, leaving them to have to "teach themselves" new systems and rely solely on manuals for complex implementation questions.
- Assessment quality is threatened by stagnant financial incentives for educators. Pay for teacher committees (composed of experts who develop and review test items) is generally

low, leading subject-matter experts to decline participation in the item development and review process.

Accountability Factors

The Burden of "Verified Credit" Graduation Requirements

- Stakeholders across multiple groups are deeply concerned about the "high-stakes" tie between state assessments and graduation. Virginia students must earn "verified credits" by passing SOL tests to graduate, a system described by state leaders as a "grind" for struggling students. One participant noted that Virginia is "on an island" because it is the only state requiring students to reach "Proficiency" levels on assessments to graduate, whereas other states allow for lower thresholds such as "Basic" or "Approaching."

> "At the high school level, the psychological piece is ignored. If they take earth science 4 times and biology 3 times, they're retesting plus doing all other SOL tests. It is damaging to a kid's mental and emotional health to tell them that they will not graduate if they do not pass."
>
> **High School Assistant Principal and School Testing Coordinator**

- Linking SOLs to "verified credit" requirements creates a cycle of failure, retakes, and high costs. This requirement consumes excessive instructional hours through constant remediation. Division leaders frustratingly noted that some students take a single test (such as Algebra I) six or seven times, a process that "primarily benefits the testing contractor rather than the student." For high school seniors, slow scoring processes create uncertainty, frequently leaving students and families unsure if graduation requirements have been met until the very end of the academic year.
- To address these concerns, some participants suggested moving to a system where the SOL counts for 10 percent of a course grade, allowing students to maintain "skin in the game" without the test becoming an overwhelming barrier to graduation.

> "There has been a long assessment cycle history. It used to be very predictable, coming every 7 years, and we had time to process it and figure it out. We had realistic timelines. Reading and math would never roll out simultaneously. Would love to see [a return to] that moving forward."
>
> **Assistant Superintendent for Instruction**

Systemic Instability and "Change Fatigue"

- Professional stakeholders are exhausted by frequent shifts in state standards and accountability models. Participants described the current culture of frequent legislative, policy, and political shifts as a "whirlwind" where staff are expected to implement major changes (such as new CAT formats or integrated reading and writing components) before staff are fully trained. Division leaders specifically noted that the state is on its third accountability system since the Every Student Succeeds Act was passed in 2015, which has left them feeling they are "building the airplane while flying it." This instability

prevents "apples-to-apples" data comparisons and makes it nearly impossible for divisions to track longitudinal data effectively.

- Established stakeholders would strongly prefer returning to a structured, multiyear assessment lifecycle. They noted that a predictable 7-year cycle previously provided a "reprieve" year that allowed sufficient time for VDOE and local divisions to align instructional materials and textbooks with new standards. Current months-long change cycles are viewed as "politicized" and lack the necessary "buy-in" from school staff who anticipate that rules will be replaced shortly after implementation. State staff strongly recommended that any new system have a 2–3-year development and testing cycle before field implementation to ensure the resulting data is stable, reliable, and valid.

Psychosocial Impact of Stress and Accountability

- The high-stakes nature of the current system drives both student and educator anxiety. Even gifted students who typically perform well experience stress due to the intense schoolwide emphasis on scores. Parents are concerned that end-of-course tests required for high school graduation damage student mental health.
- The emotional toll of "growth models" and "cut scores" (the minimum score needed to pass) was a recurring theme. Teachers noted that students often "cry" or feel like failures when they pass a test but do not meet high "growth" targets, negatively impacting their self-esteem as learners. Stakeholders have major concerns about planned increases to these cut scores; one division projected that a school's math pass rate could plummet from 81 percent to 12 percent if these changes are implemented immediately, which would be "devastating" for divisions with limited resources.
- Participants perceived the use of assessment data as a "stick" rather than a "carrot." Parents and community organizations have concerns that scores are used to justify cutting funding or penalizing disadvantaged schools, rather than directing resources to schools in need of remediation. This creates a high-pressure environment where teachers may prioritize test-taking strategies over deep learning, as the penalty for student opt-outs can unfairly burden a school's official accreditation rating.

> "I want our school system to be really good for every kid with very high rigor. But it is not fair to put that on ELs or special education students. VDOE should not use the scores as a stick to take away funding. We know which schools need help—remediate first. After that then take a stick. Schools already at a disadvantage will be at more of one. People may disagree with me because they want rigor. We all want the best schools possible, but we disagree on the process and who is ultimately accountable."
>
> **Special Education PTA Member**

RECOMMENDATIONS

Based on the stakeholder feedback, the most impactful actions by VDOE would be:

- ***Invest in communications and technical assistance to districts:*** Schools and divisions are asking for guidance and information from VDOE. They are requesting printed materials, trainings, and coaching to ensure that they understand potential changes and can best prepare for them.

- ***Develop tools to inform instruction:*** Because SOLs happen at the end of the year, the data are often not used to inform instruction. It would be helpful for VDOE to create actionable reports for teachers to use the following year on items such as the standards students in their class struggled with most and the strengths and weaknesses of incoming students.
- ***Simplify the language on non-English/Language Arts Assessments:*** Write non-English/Language Arts assessments at a reading level at least one grade level below the assessed grade to ensure that the test is accurately measuring student knowledge and not reading level.
- ***Consider how to reduce burden:*** Consider whether tests can be shortened or have fewer reading passages to minimize the time spent on each test, especially for younger grades.
- ***Provide more sample items:*** Teachers use sample items to help prepare students for the types of questions they may see on the exam. Providing timely and current questions can ensure that students are well prepared.
- ***Convey clear and accurate timelines:*** Divisions struggle with the pace and frequency of changes and would benefit from clear communication about what changes are being implemented, when they are implemented, and why. When divisions can articulate what to expect and why it is happening, teachers are more likely to buy into the changes and work to implement them.

METHODOLOGY

- **Engagement Approach:** The team conducted 13 focused engagement sessions with 65 participants across all regions of Virginia in January 2026. Respondents included state- and division-level leadership; school administrators and classroom educators; and parents and community-based organizations.
- **Regional Diversity:** Excluding VDOE staff, focus group participants came from across the Commonwealth. The QIP team contacted multiple interviewees per region to ensure geographic diversity, but had uneven participation.
    - Region 1: 7 participants
    - Region 2: 3 participants
    - Region 3: 1 participant
    - Region 4: 13 participants
    - Region 5: 9 participants
    - Region 6: 1 participant
    - Region 7: 3 participants
    - Region 8: 2 participants
- **Analysis:** Data were analyzed using a thematic approach, prioritizing the preservation of participant sentiment and terminology.

- **Limitations:** Findings represent the depth of experience of focus group participants and may not capture the breadth of every position or division in the state.
- **Objectives:**
    - Capture stakeholder perceptions of state assessments at the local level.
    - Provide detailed input on usability, accessibility, and practical implementation concerns from key user groups.

APPENDIX E. REFERENCES

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 377–398). Lawrence Erlbaum Associates.

Alaska Department of Education & Early Development. (2023). *Achievement level descriptors (ALDs) English language arts grades 3–5*. https://education.alaska.gov/assessments/akstar/AKSTAR_ALDs_ELA_grades%203-5.docx

Alaska Department of Education & Early Development. (2024a). *Educator guide to assessment results: AK STAR English language arts and mathematics*. https://education.alaska.gov/assessments/akstar/EdGuide_AssessmentResults_AKSTAR.pdf

Alaska Department of Education & Early Development. (2024b). *FY2025 significant budget items*. https://education.alaska.gov/state_board/october-2024/8.1%20FY2025%20DEED%20Budget%20Highlights.pdf

Alaska Department of Education & Early Development. (2025). *AK STAR family guide to student reports*. https://education.alaska.gov/assessments/akstar/FamilyGuide_AssessmentReports_AKSTAR.pdf

Alaska Department of Education & Early Development. (2026). *Educator guide to student readiness: AK STAR English language arts and mathematics*. https://education.alaska.gov/assessments/akstar/EducatorGuidetoStudentReadiness_Final.pdf

American Institutes for Research. (2013). *Smarter Balanced Assessment Consortium: Cognitive laboratories technical report*. https://portal.smarterbalanced.org/hubfs/44715956/cognitive-laboratories-technical-report.pdf

Beiting-Parrish, M., & Whitmer, J. (2023). Lessons learned about evaluating fairness from a data challenge to automatically score NAEP reading items. *Chinese/English Journal of Educational Measurement and Evaluation*, *4*(3), Article 5. https://doi.org/10.59863/NKCJ9608

Brandt, W.C., (2023). *Evaluation of North Carolina's Innovative Assessment Demonstration Authority (IADA): The North Carolina Personalized Assessment Tool (NCPAT)*. The National Center for the Improvement of Educational Assessment, Inc.: Dover, New Hampshire.

Bridgeman, B. (2009). Experiences across multiple studies of computerized testing with constructed responses. *Applied Measurement in Education, 22*(2), 105–126.

California Department of Education. (2025a). *CAASPP student needs matching tool*. https://www.caaspp-elpac.org/s/docs/caasppmatchingtool.pdf

California Department of Education. (2025b). *ELPAC information guide*. https://ca-toms-help.ets.org/elpac-info-guide/intro/

California Department of Education. (2025c). *Report to the governor, the legislature, and the Legislative Analyst's Office: 2024 California Assessment of Student Performance and Progress annual implementation update and five-year cost projection*. https://www.cde.ca.gov/ta/tg/ca/documents/caaspannualupdate2024.docx

California Department of Education. (n.d.-a). *Accessibility resources planning tool*. https://www.caaspp-elpac.org/s/docs/Accessibility-Resources-Planning-Tool-2022-23.pdf

California Department of Education. (n.d.-b). *California assessment accessibility resources matrix*. https://www.caaspp-elpac.org/resources/preparation/accessibility-resources/matrix

California Department of Education. (n.d.-c). *Translated test directions*. https://www.caaspp-elpac.org/resources/preparation/accessibility-resources/translated-test-directions

Cambium Assessment, Inc. (2025a). *Utah RISE assessments item type tutorials.* https://utahrise.org/resource-list/en/item-type-tutorials

Cambium Assessment, Inc. (2025b). *Utah RISE assessments test administration manual user guide and scripts.* https://utahrise.org/content/contentresources/en/RISE%20Test%20Administration%20Manual%202025-2026.pdf

Cambium Assessment, Inc. (n.d.). *Utah RISE assessments quick guide for setting up your online testing technology.* https://utahrise.org/content/contentresources/en/Quick-Guide-for-Setting-Up-Your-Online-Testing-Technology-2024-2025_Approved.pdf

Cognia, Inc. (2025). *MCAS resource center*. https://mcas.onlinehelp.cognia.org/

Cohen, C., & Albright, L. (2014). *Smarter Balanced adaptive item selection algorithm design report.* https://www.smarterapp.org/documents/AdaptiveAlgorithm.pdf

Colorado Department of Education. (n.d.-a). *CMAS and CoAlt technical reports*. https://ed.cde.state.co.us/assessment/cmas-coalt-techreport

Colorado Department of Education. (n.d.-b). *Collaboratively-developed, standards-based performance assessment*. https://ed.cde.state.co.us/postsecondary/perfassessment

Croft, M., Ali, T.T., and O'Keefe, B. (2023). *Testing the waters: Insights into parent perspectives on through-year assessment implementation*. Bellwether. https://bellwether.org/publications/testing-the-waters/

Data Recognition Corporation. (2025). *Georgia Milestones Assessment System 2024 operational technical report*. https://lor2.gadoe.org/gadoe/file/02f3f898-8d68-44e8-9efd-c96590676b40/1/GA_Milestones_2023_2024_Technical_Report.pdf

Data Recognition Corporation. (2026). *Online tools training*. https://wbte.drcedirect.com/LA/portals/la/ott1?display=LEAP+Grades+3-8&adminId=510520&index=1

Data Recognition Corporation, Pearson, & WestEd. (n.d.). *LEAP 2025 social studies 3–8 technical report: 2023–2024.* https://doe.louisiana.gov/docs/default-source/assessment/2024-leap-grade-3-8-social-studies-field-test-technical-report-public.pdf?sfvrsn=22609d5e_3

Dynamic Learning Maps. (n.d.). *About the consortium*. https://dynamiclearningmaps.org/consortium

Dynamic Learning Maps. (2010). *U.S. education department awards grants to improve assessments for students with disabilities*. https://dynamiclearningmaps.org/us-education-department-awards-grants-improve-assessments-students-disabilities

Dynamic Learning Maps. (2024). *2023–2024 technical manual update*. https://2024-ye-techmanual.dynamiclearningmaps.org/

Dynamic Learning Maps. (2025). *Accessibility manual 2025–2026*. https://dynamiclearningmaps.org/sites/default/files/documents/Manuals_Blueprints/Accessibility_Manual.pdf

Dynamic Learning Maps. (n.d.). *What is a learning map model?* https://dynamiclearningmaps.org/model

Florida Department of Education. (2024a). *Family portal access codes quick guide and FAQ*. https://faa.fsassessments.org/content/contentresources/en/FAA-Family-Portal-Access-Codes-Quick-Guide-and-FAQ.pdf

Florida Department of Education. (2024b). *Florida Assessment of Student Thinking (FAST), Benchmarks for Excellent Student Thinking (B.E.S.T.), and science & social studies statewide assessments*. https://www.fldoe.org/file/5663/BEST-2324-Vol4.pdf

Florida Department of Education. (2025a). *2025–26 Florida Assessment of Student Thinking (FAST) fact sheet*. https://www.fldoe.org/file/20102/2526FASTGrd310FS.pdf

Florida Department of Education. (2025b). *Understanding Florida Reporting System (FRS) reports*. https://flfast.org/content/contentresources/en/Understanding%20Florida%20Reporting%20System%20Reports_25-26_FINAL_080425_508.pdf

Florida Department of Education. (2025c). *Understanding Florida statewide assessments reports for families*. https://flfast.org/resource-item/en/understanding-fsa-reports-for-families

Florida Department of Education. (2026). *Statewide assessments guide 2025–2026*. https://www.fldoe.org/file/20102/K12SAG.pdf

Florida Department of Education. (n.d.-a). *Florida Automated System for Transferring Educational Records (FASTER)*. https://www.fldoe.org/faster/

Florida Department of Education. (n.d.-b). *Technology guide*. https://flfast.org/tech-guide.html

Florida Department of Financial Services. (n.d.). *Contract information: Cambium Assessment Inc.* https://facts.fldfs.com/Search/ContractDetail.aspx?AgencyId=480000&ContractId=22679

Georgia Department of Education. (2014). *SLDS district/school dashboard user guide*. https://www.liberty.k12.ga.us/pdf/TandL/SLDS_DistrictandSchoolDashboardUserGuide.pdf

Georgia Department of Education. (2018). *Georgia's application for the Innovative Assessment Demonstration Authority under Section 1204 of the Elementary and Secondary Education Act (ESEA)*. https://www.ed.gov/media/document/gaiadaappdec2018pdf-98190.pdf

Georgia Department of Education. (2024a). *Georgia Milestones grade 5 sample individual student report*. https://lor2.gadoe.org/gadoe/file/62957c85-709a-4253-9928-5cdcb989071e/1/GM_Gr5_Sample_ISR.pdf

Georgia Department of Education. (2024b). *Georgia Milestones grade 8 sample individual student report*. https://lor2.gadoe.org/gadoe/file/62957c85-709a-4253-9928-5cdcb989071e/1/GM_Gr8_Sample_ISR.pdf

Georgia Department of Education. (2025a). *A guide to understanding and interpreting outcomes from the Georgia Student Growth Model*. https://lor2.gadoe.org/gadoe/file/1515feae-3c09-4d73-8976-82698f17e86f/1/GSGM-Guide.pdf

Georgia Department of Education. (2025b). *Understanding growth targets*. https://lor2.gadoe.org/gadoe/file/1515feae-3c09-4d73-8976-82698f17e86f/1/Understanding-Growth-Targets.pdf

Gregg, N., Young, M., & Lottridge, S. (2021, June). *Examining fairness in automated scoring.* Paper presented at the annual meeting of the National Council on Measurement in Education, Online.

HumRRO. (2016). *Smarter balanced assessment consortium: Alignment study report.* https://portal.smarterbalanced.org/library/smarter-balanced-assessment-consortium-alignment-study-report/

Institute of Education Sciences. (n.d.). *Tennessee department of education longitudinal data system*. https://ies.ed.gov/use-work/awards/tennessee-department-education-longitudinal-data-system

Joint Legislative Audit and Review Commission. (2025). *Virginia's k–12 school accountability system*. https://jlarc.virginia.gov/landing-2025-virginias-k-12-accountability-system.asp

Kim, J., & Circi, R. (2023). Examining measurement invariance and differential bundle functioning between paper-based and online assessments. *Applied Measurement in Education, 36*(2), 73–92.

Kopriva, R. J., Emick, J. E., Hipolito-Delgado, C. P., & Cameron, C. A. (2007). Do proper accommodation assignments make a difference? Examining the impact of improved decision making on scores for English language learners. *Educational Measurement: Issues and Practice, 26*(3), 11–20.

Kūkea-Shultz, P., & Brockmann, F. (2025). *Bridging psychometric and content development practices with AI: A community-based workflow for augmenting Hawaiian language assessments* (arXiv preprint arXiv:2512.17140). https://doi.org/10.48550/arXiv.2512.17140

Li, H., Chen, C.H., Fan, K., Young-Johnson, C., Lim, S., & Feng, Y. (2025). Agreement between large language models and human raters in essay scoring: A research synthesis. *arXiv preprint*, arXiv:2512.14561. https://doi.org/10.48550/arXiv.2512.14561

Lorié, W., & Dadey, N. (2026). *Artificial intelligence in large-scale assessment programs: Applications and considerations for state education agencies*. National Center for the Improvement of Educational Assessment. https://www.nciea.org/wp-content/uploads/2026/02/AI-in-LSA-Final.pdf

Louisiana Department of Education. (2018). *Louisiana's application for Innovative Assessment Demonstration Authority under Section 1204 of the Elementary and Secondary Education Act (ESEA).* https://doe.louisiana.gov/docs/default-source/links-for-newsletters/la-iada-application_final_3-30-18.pdf?sfvrsn=43af931f_0

Louisiana Department of Education. (2025). *ELA guidebooks curriculum guide 3–5 (2022)*. https://ldoe-static.s3.us-east-1.amazonaws.com/3-5_Curriculum+Guide_2022.pdf

Marion, S., Domaleski, C., & Lorié, W. (2025). *How should states choose a growth model? Aligning your growth model with policy and technical values*. Center for Assessment. https://www.nciea.org/library/how-should-states-choose-a-growth-model/

Massachusetts Department of Elementary and Secondary Education. (2020a). *Application for Innovative Assessments Demonstration Authority*. https://www.ed.gov/sites/ed/files/2020/04/Massachusetts-IADA-Proposal-Final.pdf

Massachusetts Department of Elementary and Secondary Education. (2020b). *Application for new grants under the Competitive Grants for State Assessment program*. https://www.ed.gov/sites/ed/files/2020/10/Massachusetts-Dept.-of-Elementary-and-Secondary-Education_Redacted.pdf

Massachusetts Department of Elementary and Secondary Education. (2023). *Cognitive skill level descriptions for English language arts (ELA) MCAS items.* https://www.doe.mass.edu/mcas/tdd/ela-cognitive-skills.html

Massachusetts Department of Elementary and Secondary Education. (2024). *News from DESE student assessment services*. https://us14.campaign-archive.com/?u=d8f37d1a90dacd97f207f0b4a&id=2309335e13

Massachusetts Department of Elementary and Secondary Education. (2025a). *2025 MCAS sample student work and scoring guide*. https://www.doe.mass.edu/mcas/student/2025/grade8/ela-q12.pdf

Massachusetts Department of Elementary and Secondary Education. (2025b). *Cognitive skill level descriptions for STE MCAS items*. https://www.doe.mass.edu/mcas/tdd/cognitive_skills.html

Massachusetts Department of Elementary and Secondary Education. (2025c). *MCAS resources for parents/guardians*. https://www.doe.mass.edu/mcas/parents/default.html

Massachusetts Department of Elementary and Secondary Education. (2025d). *Release of spring 2025 MCAS test items from the grade 3 English language arts paper-based test*. https://www.doe.mass.edu/mcas/2025/release/g3-ela.pdf

Massachusetts Department of Elementary and Secondary Education. (2025e). *Release of spring 2025 MCAS test items from the grade 5 science and technology/engineering paper-based test*. https://www.doe.mass.edu/mcas/2025/release/g5-ste.pdf

Massachusetts Department of Elementary and Secondary Education. (2025f). *Release of spring 2025 MCAS test items from the grade 7 English language arts paper-based test*. https://www.doe.mass.edu/mcas/2025/release/g7-ela.pdf

Massachusetts Department of Elementary and Secondary Education. (2025g). *Release of spring 2025 MCAS test items from the grade 7 mathematics paper-based test*. https://www.doe.mass.edu/mcas/2025/release/g7-math.pdf

Massachusetts Department of Elementary and Secondary Education. (2025h). *Using the 2019 MCAS sample student work and scoring guides online tool*. https://www.doe.mass.edu/mcas/student/2019/about.html

Massachusetts Department of Elementary and Secondary Education & Cognia. (2024). *2023 next-generation MCAS and MCAS-Alt technical report*. https://www.doe.mass.edu/mcas/tech/2023-nextgen-tech-report.pdf

Missouri Department of Elementary & Secondary Education. (2025). *Missouri Assessment Program Success-Ready Student Assessment: Application for new authorities under the Innovative Assessment Demonstration Authority*. https://www.ed.gov/media/document/missouri-application-2025-110444.pdf

Montana Office of Public Instruction. (2022). *Assessment design and implementation considerations for the Montana Alternate Student Testing (MAST) pilot program*. https://drive.google.com/file/d/14SSR3gx1WMO-TjsdUhP952bQV-PJnz0f/view

Montana Office of Public Instruction. (n.d.). *MAST development & overview*. https://opi.mt.gov/Leadership/Assessment-Accountability/MontCAS/Required-Assessments/Montana-Science-Assessment-FAQ/Montana-Aligned-to-Standards-Through-Year

National Center for Education Statistics. (2015). *FY 2015 Massachusetts SLDS grant application abstract*. https://nces.ed.gov/programs/slds/pdf/MA_FY15_Abstract.pdf

National Center for Education Statistics. (2019). *FY 2019 Ohio state SLDS grant application abstract*. https://nces.ed.gov/programs/slds/pdf/2019OHabstract.pdf

National Center for Education Statistics. (2023). *FY 2023 Georgia SLDS grant application abstract*. https://nces.ed.gov/programs/slds/pdf/2023GAabstract.pdf

New Hampshire Department of Education. (2018). *Application for the new authorities under the Innovative Assessment Demonstration Authority*. https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/nhpaceapplication.pdf

New Hampshire Department of Education. (2020). *PACE innovative assessment and accountability system: Overview by grade and subject*. https://www.education.nh.gov/sites/g/files/ehbemt326/files/files/inline-documents/pacegradessubjects052720_4.pdf

New Hampshire Learning Initiative. (n.d.-a). *NH community guide for NH PACE*. https://nhlearninginitiative.libguides.com/c.php?g=887150&p=6375932

New Hampshire Learning Initiative. (n.d.-b). *NH PLACE: Performance Learning and Assessment Consortium for Educators*. https://nhlearninginitiative.org/empowered-learning/teaching-and-learning/nh-place/

New Meridian. (n.d.). *Montana Aligned to Standards Through-Year (MAST) program portal*. https://newmeridiancorp.org/montana-aligned-to-standards-through-year-program-portal/

North Carolina Department of Public Instruction. (2018). *North Carolina application for new authorities under the Innovative Assessment Demonstration Authority*. https://www.ed.gov/media/document/nciadaappdec2018pdf-98198.pdf

North Carolina Department of Public Instruction. (2024). *Report to the North Carolina General Assembly: North Carolina Personalized Assessment Tool (NCPAT) pilot*. https://webservices.ncleg.gov/ViewDocSiteFile/91344

North Carolina Department of Public Instruction. (n.d.-a). *Accountability data sets and reports.* https://www.dpi.nc.gov/districts-schools/accountability-and-testing/school-accountability-and-reporting/accountability-data-sets-and-reports#2024-25Reports-4468

North Carolina Department of Public Instruction. (n.d.-b). *EVAAS*. https://www.dpi.nc.gov/districts-schools/districts-schools-support/office-educator-preparation-licensure-performance/evaas

North Carolina Department of Public Instruction. (n.d.-c). *NC Check-Ins 2.0*. https://www.dpi.nc.gov/districts-schools/accountability-and-testing/state-tests/nc-check-ins-20

North Carolina Department of Public Instruction. (n.d.-d). *Technical information for state tests*. https://www.dpi.nc.gov/districts-schools/accountability-and-testing/technical-information-state-tests

NWEA. (2022). *Achievement level explorer for educators*. https://ald-explorer.nwea.org/pick-options

NWEA. (2026). *AK STAR practice tests*. https://akstar-assess.state.nwea.org/inbrowser/index.html

Ohio Department of Education & Workforce. (2025). *Accessibility manual for Ohio's state tests*. https://education.ohio.gov/getattachment/Topics/Testing/Accommodations-on-State-Assessments/Ohio-Accessibility-Manual.pdf.aspx?lang=en-US

Ohio Department of Education & Workforce. (2026a). *Assessment accessibility for Ohio's students*. https://education.ohio.gov/Topics/Testing/Accommodations-on-State-Assessments

Ohio Department of Education & Workforce. (2026b). *Assessment training resources*. https://education.ohio.gov/Topics/Testing/Assessment-Training-Resources

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater automated scoring engine and humans for demographically based groups in the GRE General Test. *ETS Research Report Series*, *2018*(1), 1–31. https://doi.org/10.1002/ets2.12192

Smarter Balanced Assessment Consortium. (2010). *Race to the Top Assessment Program: Application for new grants, comprehensive assessment systems* (CFDA No. 84.395B). Submitted by Washington State on behalf of the SMARTER Balanced Assessment Consortium. https://portal.smarterbalanced.org/hubfs/44715956/race-to-the-top-assessment-program-application-for-new-grants-comprehensive-assessment-systems.pdf

Smarter Balanced Assessment Consortium. (2014). *Smarter Balanced system architecture and technology report*. https://www.smarterapp.org/documents/SmarterBalanced_ArchitectureReport_07112014.pdf

Smarter Balanced Assessment Consortium. (2023). *2023–24 summative technical report*. https://technicalreports.smarterbalanced.org/2023-24_summative-report/_book/index.html

Smarter Balanced Assessment Consortium. (2024). *Chapter 2: Reliability, precision, and errors of measurement. 2023–24 interim assessment technical report*. https://technicalreports.smarterbalanced.org/2023-24_interim-report/_book/reliab-prec-err.html

Smarter Balanced Assessment Consortium. (2025). *Chapter 6: Reporting and interpretation. 2024–25 interim assessment technical report*. https://technicalreports.smarterbalanced.org/2024-25_interim-report/_book/report-interp.html

Smarter Balanced Assessment Consortium. (n.d.-a). *Practice tests and sample questions*. https://smarterbalanced.org/our-system/students-and-families/samples/

Smarter Balanced Assessment Consortium. (n.d.-b). *Reporting scores*. https://validity.smarterbalanced.org/scoring/

Smarter Balanced Assessment Consortium. (n.d.-c). *Sample items*. https://sampleitems.smarterbalanced.org/

Smarter Balanced Assessment Consortium. (n.d.-d). *Smarter content explorer*. https://contentexplorer.smarterbalanced.org/

Smarter Balanced Assessment Consortium. (n.d.-e). *Tools for teachers*. https://smartertoolsforteachers.org/

Smarter Balanced Assessment Consortium & IBM. (2025). *SmarterAI think tank: Student-centric design principles for responsible use of AI*. https://portal.smarterbalanced.org/hubfs/44715956/SmarterAI%20Think%20Tank.pdf

Strain-Seymour, E., Way, W., & Dolan, R. P. (2013). *Strategies and processes for developing innovative items in large-scale assessments*. Pearson.

Tennessee Department of Education. (2023). *Tennessee Comprehensive Assessment Program math grade 3 item release.* https://www.livebinders.com/b/2426642?tabid=43434b3f-bcc6-7469-ca5c-051fcaa03ab4

Tennessee Department of Education. (n.d.-a). *Blueprints for 2025–26*. https://www.tn.gov/content/tn/education/districts/lea-operations/assessment/tcap-blueprints.html

Tennessee Department of Education. (n.d.-b). *Education Information System (EIS)*. https://www.tn.gov/education/districts/lea-operations/education-information-system-eis.html

Texas Education Agency. (2023). *The State of Texas Assessments of Academic Readiness (STAAR) hybrid scoring study: Methods and results, spring 2023 items*. https://tea.texas.gov/student-assessment/reports-and-studies/2023-staar-hybrid-scoring-study.pdf

United States Department of Education. (2009). *Application for grants under the Statewide Longitudinal Data System Recovery Act grants*. https://nces.ed.gov/programs/slds/pdf/Florida2009-ARRA.pdf

United States Department of Education. (n.d.). *Competitive Grants for State Assessments*. https://www.ed.gov/grants-and-programs/formula-grants/school-improvement/competitive-grants-for-state-assessments#awards

Utah State Board of Education. (2020). *The future of Utah's state assessment system: Recommendations from the Utah assessment strategic plan work group*. https://schools.utah.gov/assessment/_assessment_/AssessmentStrategicPlan_Dec_20.pdf

Utah State Board of Education. (n.d.). *Data gateway: Student proficiency results*. https://datagateway.schools.utah.gov/Assessment/StudentProficiency/2025/StudentProficiencyPublic

Virginia Department of Education. (2024). *Revisions to the Virginia standards of learning summative assessments of proficiency. Recommendations from the HB 585 work group convened by the secretary of education and superintendent of public instruction*. https://www.doe.virginia.gov/home/showpublisheddocument/50172/638337158572970000

Virginia Department of Education. (2025). *Report on plan to implement a new state assessment system*.

Virginia Department of Education. (n.d.). *Virginia standards of learning assessments technical report 2024–2025 administration cycle.*

Vo, Y., Rickels, H., Welch, C., & Dunbar, S. (2023). Human scoring versus automated scoring for English learners in a statewide evidence-based writing assessment. *Assessing Writing*, 57, Article 100719. https://doi.org/10.1016/j.asw.2023.100719

WestEd. (2017). *Evaluation of the alignment between the common core state standards and the smarter balanced assessment consortium summative assessments for grades 3, 4, 6, and 7 in English language arts/literacy and mathematics.* https://portal.smarterbalanced.org/library/wested-alignment-evaluation/