Maryland State Department of Education Assessment and Accountability Task Force

DECEMBER 3, 2024

Scott Marion, Ph.D.
Chris Domaleski, Ph.D.
Cara Laitusis, Ph.D.
National Center for the Improvement of Educational Assessment

Maryland
STATE DEPARTMENT OF EDUCATION

TABLE OF CONTENTS

TABLE OF CONTENTS (CONTINUED)

EXECUTIVE SUMMARY

The Maryland State Department of Education convened the Maryland Assessment and Accountability Task Force to examine and recommend improvements to the state's accountability and assessment systems. The Task Force first studied the Maryland School Report Card and the state's accountability system and recommended ways to strengthen connections between school ratings and student achievement. The Task Force focused on improving the transparency of how school ratings (stars) are awarded and increasing the alignment between the *Blueprint for Maryland's Future (Blueprint)* and the ESSA accountability requirements across school systems and statewide while maintaining compliance with federal and state requirements.

The Maryland State Department of Education convened the Maryland Assessment and Accountability Task Force to examine and recommend improvements to the state's accountability and assessment systems.

The Task Force addressed accountability first in part because it is important for state assessment results. Thus, before recommending changes to the assessment system, it was important to understand how the assessment results would be incorporated into the accountability system. Once that was clear, the Task Force discussed how to improve the usefulness of the assessment results for multiple users and how to increase the credibility of assessments and its results.

Between May and November 2024, a broad range of education constituents from across the state participated in the Assessment and Accountability Task Force. Meetings were held in person and remotely and facilitated by experts from the Center for Assessment. This report documents the process and recommendations produced by the Task Force.

Accountability

The Task Force began by outlining the goals, purposes, and uses of an effective accountability system. They affirmed that effective systems should provide key information on valued outcomes and be integrated with improvement mechanisms that specify necessary conditions, resources, and supports to foster improved actions and results. The Task Force articulated goals for the system that emphasized the importance of providing equitable and inclusive learning opportunities for all students, promoting student achievement of Maryland's academic content standards, preparing students for post-secondary success, supporting educators, and fostering engagement from parents and the community.

The Task Force articulated goals for the system that emphasized the importance of providing equitable and inclusive learning opportunities for all students, promoting student achievement of Maryland's academic content standards, preparing students for post-secondary success, supporting educators, and fostering engagement from parents and the community.

Once the Task Force clarified the system goals, it identified the following principles to guide the development of an accountability system:

- Prioritize implementing changes to the system but preserve longitudinal comparability where possible.

- Support meaningful comparisons of school performance but explore ways to offer limited flexibility.
- Explore ways to streamline and simplify the system without sacrificing quality or comprehensiveness.
- Create a single coherent system that meets federal requirements and reflects state priorities.

Informed by the goals and design principles, the Task Force developed recommendations for five indicator categories and the overall design, which are summarized in the following table.

Table 1. Summary of Accountability Recommendations

COMPONENT	RECOMMENDATIONS
Academic Achievement Indicator	• The academic achievement indicator should be based exclusively on the proficiency rate in ELA and mathematics.
Growth Indicator	• Adopt Student Growth Percentiles or Value Tables configured in a manner that best supports the three prioritized criteria: 1. the extent to which the growth indicator is correlated with average prior achievement (lower correlations are preferred), 2. the precision of the growth scores for the full range of results, and 3. the degree to which results are sensitive to progress across the distribution. • Conduct analyses to evaluate models for prioritized criteria. • Make sure the methods used to produce growth scores are transparent and well-documented. • Ensure resources and supports are available to help constituents interpret and use results.
Graduation Rate Indicator	• Continue to include only the four-year adjusted cohort graduation rate and the five-year extended graduation rate. • Continue the current influence of each component. The four-year rate should have double the influence of the five-year rate.
Post-Secondary Readiness Indicator	• Include on-track, college and career readiness (aligned with the Blueprint) and post-secondary preparation in the school accountability model. • Continue to review and refine the accountability framework to ensure the named accomplishments are complete and appropriate, the performance expectations for similar outcomes are comparable in rigor, and the overall influence (i.e., points and weights) are appropriate.
Progress in Achieving English Language Proficiency Indicator	• Continue to use WIDA ACCESS with an exit standard of 4.5. • Conduct additional research on the conditions and time to exit to inform potential adjustments to the ELP indicator. • Supplement information from WIDA ACCESS with other sources of evidence to help support student success. • Focus on communication and support to help make information more actionable.

COMPONENT	RECOMMENDATIONS
Design Decisions for Aggregation and Determinations	• Establish common performance levels for indicators (e.g., 1-4) using a deliberative process with experts and other key constituents. • Conduct a small-scale study to determine whether accountability system users arrive at the intended interpretations when presented with reports derived from a profile method compared with those derived from a weighted average and overall rating. • Following this study, whatever decision-making process is endorsed, the Task Force recommended employing an accountability standard-setting process to guide the federally required determinations and to establish performance levels if overall performance levels are desired.

Assessment

The Task Force discussed key aspects of assessment design and implementation and offered recommendations to address the following critical questions associated with a state assessment program.

- **Accessibility and Fairness:** How can the MSDE help ensure assessments are fair and accessible to a broad range of learners?
- **Adaptive or Fixed Form:** Will the test be administered to students using a computer adaptive testing process or a "fixed form" approach?
- **Testing Time:** How much time should be required for state summative testing, and what types of items (questions) should be included on the test?
- **Score Reporting:** How should the system of score reports be designed to support high-quality and understandable information for various users in the educational system?
- **Non-Summative Resources:** Should the state procure non-summative resources (e.g., interim assessments, formative assessment tools) as part of the summative assessment RFP?
- **Communication, Outreach, and Advocacy:** How should MSDE design and execute a communication plan to enhance the credibility and usefulness of the state assessment system?

The assessment recommendations for each of these components are summarized in Table 2.

Table 2. Summary of Assessment Recommendations

COMPONENT	RECOMMENDATIONS
Accessibility and Fairness	• Design all MCAP tests using the most up-to-date research to ensure all students can demonstrate their knowledge and skills without barriers. • Evaluate all MCAP tests from the design process through the reporting of results to ensure the testing program is as fair as possible for all student groups and does not privilege any group. • Ensure the testing platform does not hinder students from demonstrating their knowledge and minimize the change in testing platforms throughout the K-12 testing experience. • Continue using the alternate assessments currently in place for students with the most significant cognitive disabilities and request that the technical advisory committee evaluate how best to integrate the results into the school accountability system.

COMPONENT	RECOMMENDATIONS
Adaptive or Fixed Form	• Develop a system that will allow MSDE to document the quality of every test form administered to students. Release a subset of test items each year to enhance reporting, credibility, and usefulness in helping educators and students understand the level of knowledge and skills required to perform successfully on the tests. • Encourage bids through the RFP process that rely on a multi-stage adaptive design. But allow offerors to propose an alternative design to meet the State's goals. In either case, the offeror must present evidence of the advantages and disadvantages of the proposed approach for the State.
Item Types and Testing Time	• Include a range of item types to ensure that the full breadth and depth of the standards are well-measured. Design open-response items/tasks to signal the types of tasks the Task Force and MSDE would like to see used as part of regular classroom instruction. • The total test length should be no longer than practically necessary to produce valid, reliable, and useful scores.
Score Reporting	• Support the development of a coherent system of score reports with a precise specification of each report's intended users and uses. • Commit to releasing score reports for both assessment and accountability as quickly as possible. • Create a report design process led by—or at least includes—communications experts. • Require report developers to present evidence (or a clear plan for collecting evidence) to evaluate claims of usefulness for each of the intended user groups. • Score information must be easily uploaded to district student information systems. • Support a comprehensive system of report interpretation and related assessment literacy professional learning opportunities for the various intended report users.
Non-Summative Resources	• Invite potential respondents to an assessment RFP to include the development of modular interim assessments as a cost option. • If MSDE exercises such a cost option, the state should support high-quality use through extensive professional learning opportunities and supporting materials. • Using these non-summative tools should be optional for school districts.
Communication, Outreach, and Advocacy	• Develop a comprehensive communication strategy to showcase positive stories about the assessment system and how schools and districts use the assessment results. • Conducting internal research and facilitating the use of Maryland assessment and related data for research uses to address policy-related and other important research and evaluation questions.

The MSDE Assessment and Accountability Task Force met regularly for over seven months in 2024 to deliberate and make recommendations to improve Maryland's assessment and accountability systems. The recommendations presented in this report provide meaningful guidance for MSDE as it prepares to release a Request for Proposals (RFP) for its next assessment system. The recommendations for improving the accountability system will provide valuable advice to MSDE as it creates the business rules to operationalize the new vision for school accountability in Maryland.

MSDE will implement many of the accountability recommendations for the 2024-2025 accountability results, contingent upon federal approval. Other recommendations, such as determining which growth model to use, will require study and analyses early in 2025 to have the information necessary to decide on the growth model by late spring 2025. The new or revised growth indicator will not be implemented before the 2025-2026 school year. Changes such as aggregation approaches and producing annual determinations will be on a similar timeline.

The recommendations presented in this report provide meaningful guidance for MSDE as it prepares to release a Request for Proposals (RFP) for its next assessment system. The recommendations for improving the accountability system will provide valuable advice to MSDE as it creates the business rules to operationalize the new vision for school accountability in Maryland.

The assessment recommendations will support the development of the next assessment RFP. The RFP and contracting process will occur during the first half of 2025 with hopes of awarding the next assessment contract by late summer 2025. Transitioning from one assessment system is a detailed endeavor that takes time to do well. MSDE plans to operationalize the next assessment system for the 2026-2027 school year but will examine prudent ways to accomplish this on a faster timeline.

The state assessment system and, to a lesser extent, the state accountability system are regularly reviewed by the MSDE Technical Advisory Committee (TAC). The Task Force recognized this critical function but recommended regularly convening a policy- and practitioner-oriented advisory committee to provide feedback on the implementation of these two systems. Further, MSDE, its technical advisors, and this type of policy/practice advisory committee should support a continuous improvement process to ensure that the accountability system meets the changing needs of the State of Maryland and its educational system.

MSDE, its technical advisors, and a policy/practice advisory committee should support a continuous improvement process to ensure that the accountability system meets the changing needs of the State of Maryland and its educational system.

INTRODUCTION

The Maryland State Department of Education (MSDE) sought to evaluate its state accountability and assessment systems to consider how both systems might need to be adjusted to better serve Maryland's students and education constituents and make recommendations for the future of each system. Toward this end, MSDE convened an Assessment and Accountability Task Force (the Task Force) comprised of key Maryland education stakeholders. MSDE partnered with the National Center for the Improvement of Educational Assessment (Center for Assessment), a non-profit, non-partisan consulting firm, to facilitate the Task Force and provide assessment and accountability expertise. MSDE held in-person and virtual meetings with the Task Force to deliberate on technical, policy, and practical issues associated with implementing improved state accountability and assessment systems.

The Task Force had two major goals. One goal was to provide recommendations to support the drafting of a new Request for Proposals (RFP) for Maryland's next statewide summative assessment system. MSDE's current assessment contracts extends through the reporting of 2025-2026 assessment results. Having a new (or continuing) assessment contractor in place for the 2026-2027 school year or sooner will require MSDE assessment staff to write a new RFP early in 2025. The Task Force's recommendations will greatly inform the technical specifications of the RFP. The Task Force's second goal was to support near- and long-term vision for state-led school accountability. The Task Force began with the accountability discussions because the assessment results are a major input into the accountability system. This way, Task Force members could understand the assessment needs to best support school accountability decisions.

This report presents the results of the Task Force's deliberations, recommendations to MSDE, and related considerations for the state's RFP for the next statewide summative assessment system. The recommendations in this report reflect the consensus of Task Force members. Where consensus was not reached, decisions were based on a majority of members. We noted throughout the report where consensus was not reached and did our best to outline the multiple perspectives.

Process

MSDE leadership recruited a representative collection of education stakeholders to form the Task Force. MSDE recruited school and district personnel from various Maryland communities and constituents from important organizations such as the Assessment Implementation Board of the Blueprint for Maryland's Future, the University of Maryland, and the State Board of Education. Twenty-seven education stakeholders constituted the Task Force, as seen in Table 3 below.

Table 3. Maryland Assessment and Accountability Task Force

TASK FORCE MEMBER	POSITION
James Allrich	Principal, Argyle Magnet Middle School
Jennifer Bell-Ellwanger	President and CEO, Data Quality Campaign
Deann Collins	Deputy Superintendent, MSDE
Clarence Crawford	Past President, Maryland State Board of Education
Tania Cunningham-Raycrow	Teacher (Special Education), Somerset Intermediate School

TASK FORCE MEMBER	POSITION
Melissa DiDonato	Chief Academic Officer, Baltimore County Public Schools
Cheryl Dyson	Superintendent, Frederick County Public Schools
Drew Fagan	Associate Professor, University of Maryland College of Education
Timothy Guy	Director of Assessment and Reporting, Howard County Public Schools
Zach Hands	Executive Director, Maryland State Board of Education
Millard House III	Superintendent, Prince George's County Public Schools
Thornell Jones	Education Chair, Caucus of African American Leaders (CAAL)
Cindy Lotto	Honors and AP US History Teacher, Gaithersburg High School
Maureen Margevich	Supervisor for Testing and Accountability, Washington County Public Schools
Josh Michael	President, Maryland State Board of Education
Jason Miller	Principal, Prince Street Elementary School
Maria Navarro	Superintendent, Charles County Public Schools
Ellen O'Neill	Executive Director, Atlantic Seaboard Dyslexia Education Center
Sharon Pepukayi	Superintendent, Talbot County Public Schools
Evelyn Policarpio	Teacher (Math, Grade 8), Benjamin Tasker Middle School
Alex Reese	Chief of Staff, MSDE
Geoff Sanderson	Deputy Superintendent, MSDE
Laura Stapleton	Chair, Department of Human Development and Quantitative Methodology, University of Maryland College of Education
Andrae Townsel	Superintendent, Calvert County Public Schools
Gerrod Tyler	President, Free State PTA
Darryl Williams	Associate Director, National Center for the Elimination of Educational Disparities, Morgan State University
Jennie Wu	Executive Director, Strategy & Continuous Improvement, Baltimore City Public Schools

The Task Force was led and facilitated by three professionals from the Center for Assessment, Drs. Scott Marion, Chris Domaleski, and Cara Laitusis. The first meeting was on May 2, 2024, and the process extended through November 22, 2024, with six full-day in-person meetings and two three-hour webinars. Four remote subcommittee meetings were formed to consider accountability indicators for English learners and college and career readiness.

The meetings were structured to guide the Task Force through a process built on foundational concepts in assessment and accountability, allowing the Task Force to deliberate over challenging design decisions. The Center for Assessment prepared a set of technical briefs and other materials that outlined critical issues associated with significant accountability and assessment topics. These materials allowed Center for Assessment facilitators and the Task Force members to address key design considerations more quickly. The Center for Assessment then solicited feedback from Task Force members via whole- and small-group discussions. Input from groups and individuals was captured in Google documents and related forms.

Table 4 below provides a list of meeting dates and focal topics.

Table 4. The Arc of the Task Force Work.

MEETING DATE	MAJOR DISCUSSION TOPICS
May 2, 2024	Orientation to the work, foundations of accountability, federal accountability and assessment requirements, review of Maryland's current accountability system, and describing intended purposes and uses of accountability results.
May 30, 2024	Accountability goals, uses, and design principles continued; review of state ESSA models and broader "measures that matter" for schools.
June 12, 2024 (remote)	Introduction to system design considerations
July 23, 2024	In-depth discussions and deliberations of growth models, college and career readiness indicators, and how they aligned with the Blueprint.
August 19, 2024* (remote)	Subcommittee meeting to discuss recommendations for the progress in English language proficiency indicator
August 30, 2024* (remote)	Subcommittee meeting #1 to discuss recommendations for the college and career readiness indicator
September 5, 2024	Introduction to the state assessment system, federal requirements, critical assessment decisions tied to desired uses and purposes, and focusing on some important technical considerations.
October 4, 2024* (remote)	Subcommittee meeting #2 to further develop recommendations for the college and career readiness indicator
October 15, 2024	Solidifying critical assessment and accountability recommendations.

MEETING DATE	MAJOR DISCUSSION TOPICS
October 31, 2024* (remote)	Subcommittee meeting three to refine recommendations for the college and career readiness indicator.
November 12, 2024	Review a draft of this report and make recommendations about aggregating the information from the multiple accountability indicators and making determinations about schools.
November 22, 2024 (remote)	Final report review

** Subcommittee meetings*

ACCOUNTABILITY FOUNDATIONS

System Goals, Purposes, and Uses

The Task Force emphasized that accountability systems are most effective when they 1) provide information about inputs and outcomes the state values the most and 2) integrate with improvement systems that specify the conditions, resources, and supports that can help promote improved actions and outcomes. Accordingly, the committee clarified the high-priority goals the system should support for students, educators, and leaders. These include:

- Support equitable and inclusive opportunities to learn for all students
- Promote student achievement of Maryland's academic content standards, focusing on literacy, numeracy, and critical thinking
- Prepare students for post-secondary success in college, careers, and community life
- Foster engagement of parents and community members
- Build support and capacity for teachers and leaders
- Promote safe and positive learning environments

Multiple constituencies rely on information from the accountability system to support these goals. For example, policymakers may use information to guide resource allocation. District and school leaders may use accountability data to monitor the effectiveness of interventions. Parents and community members leverage accountability results to inform decisions about engagement and advocacy. Task Force members emphasized that the accountability system is most effective when it provides clear and useful feedback in a timely manner that addresses the wide range of factors associated with student success.

More broadly, supporting these ambitious goals requires more than collecting and reporting information on valued outcomes. The system must be designed to help leaders and educators specify the practices that can support school improvement efforts. Ultimately, claims about how assessment and accountability work within a larger system to support the intended outcomes should be represented in a comprehensive theory of action.

Design Principles

Following the development of system goals, the Task Force worked to identify design principles to guide the development of an accountability framework. The committee discussed a revised system's desired characteristics and features, addressing some trade-offs associated with competing priorities. Ultimately, the committee identified the following design principles.

1. **Prioritize implementing changes to the system but provide longitudinal comparability where possible.**
   While longitudinal comparisons are helpful, changing the system to better reflect priority goals and uses may be more important. The Task Force sought to maintain continuity in selected areas to retain a basis to compare performance over time, such as using the same academic performance measure as the legacy system.

2. **Support meaningful comparisons of school performance but explore ways to offer limited flexibility.**
   The Task Force affirmed the importance of a school accountability system that allows constituents to meaningfully compare school performance. However, the Task Force was open to flexibility that minimally impacts comparability and maintains appropriate expectations. Targeted flexibility, such as offering choices to demonstrate college and career readiness, may help include various indicators and support equity and fairness.

3. **Explore ways to streamline and simplify the system without sacrificing quality or comprehensiveness.**
   Many Task Force members noted that the current system is not sufficiently understandable, which hinders its utility. For this reason, the system should be as streamlined and simple as possible while maintaining the necessary technical defensibility and breadth. Moreover, the design should not add burdensome new requirements for districts and schools.

4. **Create a single coherent system that meets federal requirements and reflects state priorities.**
   The state accountability system should meet federal requirements but should not be unnecessarily constrained by these requirements. Achieving the breadth necessary to represent the *Blueprint for Maryland's Future (Blueprint)* may require initiatives beyond the federal system. For example, Maryland may build a robust reporting system that is much broader than the federal system or may include criteria and ratings beyond what the Every Student Succeeds Act (ESSA) requires. However, accountability initiatives outside the federal system must be coherently linked to avoid sending different signals about priorities and performance.

ACCOUNTABILITY INDICATOR RECOMMENDATIONS

Indicators describe the data in the model and provide information about school performance. They should be valid, reliable, fair, and well-suited to meaningfully differentiate Maryland's schools' performance.

Indicators are combined in some way to support a larger system of Annual Meaningful Differentiation (AMD), which is used for school identification. The Every Student Succeeds Act (ESSA) requires the following indicators:

1. **Academic achievement** is measured by proficiency on the annual reading or language arts and mathematics assessments in grades 3-8 and one high school grade.
2. **Other academic indicator** as measured by student growth for elementary and middle schools or another valid and reliable statewide academic indicator that allows for meaningful differentiation. At the state's discretion, growth may also be included for high schools.
3. **Graduation rate** for high schools, as measured by the four-year adjusted-cohort graduation rate (ACGR) and, at a state's discretion, one or more extended-year ACGRs.
4. At least one **indicator of school quality or student success (SQSS)** that meaningfully differentiates between schools and is valid, reliable, statewide, and comparable.
5. **Progress in achieving English-language proficiency (ELP)**, as defined by the state and measured by the statewide ELP assessment.

The following sections outline the Task Force's recommendations for each of Maryland's accountability system's required indicators and components.

Academic Achievement

Under ESSA, the Academic Achievement Indicator is a measure of proficiency collected through the administration of mathematics and reading/language arts exams in grades 3-8 and once in high school to no less than 95% of enrolled students in those grades. States can include other tested grades and subjects, but those other than mathematics and reading/language arts would be included in the Other Academic or School Quality/Student Success indicator. While states are required to report percent proficient, ESSA allows states to determine, within some constraints, how to use assessment results in their systems of differentiation.

States commonly use one of two approaches to compute an achievement indicator. The most common approach, used by about two-thirds of states, is **percent proficient** on the state ELA and mathematics assessment. This is simply a ratio of all students who earn level 3 or 4 on the MCAP divided by the number of examinees. Proponents of using percent proficient as the academic achievement indicator note that it is straightforward to calculate and interpret.

A second approach involves creating a **performance index** using information from each achievement level. There are multiple ways to produce an index. Typically, it involves assigning point values to each performance level, which are averaged for all students to get a group or school value. Decisions about allocating points in a total index score reflect a value judgment about what achievement patterns will effectively distinguish schools. Approximately one-third of states currently use a performance index approach for the academic achievement indicator in their ESSA school

accountability system. Proponents of using a performance index cite the benefits of awarding partial credit for students in level 2 and incentivizing students to earn advanced performance.

Maryland's current accountability system uses a composite of percent proficient and performance index, with each approach equally weighted.

Academic Achievement Summary

The Task Force did not support a performance index or a composite of the two approaches, which can obscure low proficiency rates for some schools or groups by offsetting lower performance with higher performance. In contrast to the current approach, proficiency is clear and easy to understand and supports the design principle of simplifying and streamlining the school accountability system.

The Task Force emphasized the importance of rewarding academic progress but noted that growth is addressed prominently elsewhere in the model.

Academic Achievement Recommendation

The academic achievement indicator should be based exclusively on the proficiency rate in ELA and mathematics.

Growth

ESSA requires that the state's accountability system include another academic indicator for elementary and middle schools beyond academic achievement. The other academic indicator may include either a measure of student growth or another valid and reliable statewide academic indicator that allows for meaningful differentiation.

Different views of performance (see Carlson, 2001 or Castellano & Ho, 2012) can provide a more complete portrayal of academic performance to support improvement efforts, as shown in Table 5. The academic achievement indicator addresses status or performance at a single point in time, while growth examines the progress of individual students over time.

Table 5. Four Views of School Performance.

ACHIEVEMENT (in relation to standards)	**Status** **What performance is required on the selected assessment(s)? For example, percent proficient or mean scale score.**	**Improvement** **Is the performance of successive groups increasing from year to year? For example, has the percentage of students scoring proficient changed?**
EFFECTIVENESS (in relation to past performance)	**Growth** Are students making expected progress as they move from one point in time to another? For example, gain score or growth percentile.	**Acceleration** Is the school or group becoming more effective or improving more rapidly? For example, are growth rates for schools or groups increasing over time?

The Task Force members reviewed growth models commonly used in state accountability models to inform their deliberations. Table 6 summarizes these models and the central questions they address.

Table 6. Common Academic Growth Models.

MODEL	KEY QUESTION
GAIN SCORE	What is the magnitude of progress on a vertical scale?
Growth-to-Standard	Is the student's progress on track to a significant target?
Categorical (Value Table)	Has the student transitioned from one performance category to another?
Growth percentile	How does the student's performance this year compare to his or her academic peers?
Regression or Value-added	Statistically controlling for selected factors, has the student grown more or less than expected?

The models presented in Table 6 are not mutually exclusive. For example, it is possible to implement a growth-to-standard approach with growth percentiles. However, this categorization scheme was useful for exploring the characteristics, relative advantages, and limitations of different approaches.

Acknowledging that no gold standard exists for evaluating growth measures, the Task Force developed the following criteria for growth models used in Maryland's school accountability system:

- The model should correlate only weakly with school characteristics, demographics, and average prior achievement.
- The results should not systematically favor high or low-performing schools.
- The model should be technically strong and provide meaningful and sufficiently precise growth estimates across the full achievement scale.
- The approach should be relatively easy to communicate and invite few misconceptions.

These criteria address the most important policy, technical, and practical considerations for growth that are consistent with the goals and design principles for the overall accountability system.

The first two criteria reflect the value placed on ensuring growth is picking up on a distinct aspect of student progress rather than simply amplifying the influence of status (i.e., proficiency rates) already in the accountability model. By so doing, the model will produce more fair results. For example, schools that serve a greater proportion of economically disadvantaged students or English learners should not have dramatically different growth distributions compared to schools with fewer students in these groups. All schools should have access to favorable growth scores when students demonstrate academic progress.

An emphasis on technical defensibility ensures that the growth model can be meaningfully compared within and across years and that growth is as precise as possible throughout the distribution. Moreover, the model should be sensitive to detecting progress even for scores among the lowest or highest in the state.

The final criterion emphasizes the importance of ensuring results are clear and actionable for a wide range of users, including educators and parents. When models are overly complex, constituents may not fully benefit from the information. Moreover, a high degree of complexity can erode trust in the model.

Growth Summary

Based on the growth criteria, the Task Force members determined that Student Growth Percentiles and Value Tables were most likely to support their criteria. The Task Force was polled multiple times to determine which model was preferable, and the results were consistently split. Without a clear directive for either model, the Task Force acknowledged that the emphasis should be placed on ensuring the model specifications and implementation plan are focused on supporting the prioritized criteria. This led to the recommendations in the subsequent section.

Growth Recommendations

- *Adopt Student Growth Percentiles or Value Tables configured in a manner that best supports the prioritized criteria.*
- *Conduct analyses to examine these prioritized criteria:*
  - *The extent to which the growth indicator is correlated with average prior achievement (lower correlations are preferred)*
  - *The precision of the growth scores for the full range of results.*
  - *The degree to which results are sensitive to progress across the distribution.*
- *Make sure the methods used to produce growth scores are transparent and well-documented.*
- *Ensure resources and supports are available to help constituents interpret and use results.*

Graduation Rate

States have limited flexibility in operationalizing the graduation rate indicator for their state accountability systems under ESSA. All states must use the four-year adjusted cohort graduation rate (ACGR) and may also use, at their discretion, one or more extended-year ACGRs as the measures for the indicator (ESSA, Section 1111(c)(4)(B)). The ACGR is calculated as the percent of students in a ninth-grade cohort that graduates with a regular high school diploma in a specified number of years or less (i.e., four-year or, at a state's discretion, one or more extended years) consistent with the definition of the four- and extended-year ACGR in ESEA section 8101(25). The required ACGR calculation is shown in Figure 1.

$$\frac{\text{4-year cohort graduates in Year X}}{\text{(First time 9th graders in year X-4) + (Transfers in) – (Verified transfers out) – (Exclusions)}}$$

Figure 1. The Adjusted Cohort Graduation Rate

Moreover, federal requirements stipulate that the graduation rate must be based on students earning a regular high school diploma. Alternative accomplishments such as a certificate of completion, a modified diploma, or a general equivalency diploma are prohibited from counting toward the graduation rate under ESSA.

School graduation rates must be part of the state's accountability system for high schools, which is used to identify schools for CSI, TSI, and ATSI. Additionally, states must separately identify any school that graduates fewer than 67 percent of its students as CSI.

Maryland currently includes a composite of the four and five-year graduation rates in the state accountability system. Including a five-year extended graduation rate is a common practice used in more than two-thirds of the states. Including a six-year or longer extended graduation rate in state accountability is possible. However, this is very uncommon in state accountability systems. Among other concerns, data show that extended year rates beyond five years rarely contribute much system influence.

The four-year rate in Maryland has double the influence of the five-year rate. Specifically, graduation contributes 15 points in the current model, with 10 points coming from the four-year rate and five points from the five-year rate. The Task Force discussed whether it is desirable to increase the influence of the extended-year rate. However, participants noted the importance of ensuring that the emphasis on four-year rates is not obscured.

Graduation Rate Summary

The Task Force acknowledged the importance of incentivizing student persistence beyond four years while keeping the primary focus on graduating on time. Maryland should continue using the four- and five-year rates but not other extended ones. Moreover, the weight of the four- and five-year rates should be consistent with current practice.

While graduation rates are important, they provide limited information about the range of competencies that signal that a student is ready to thrive in post-secondary college and career pursuits. For this reason, they support a separate indicator of college and career readiness.

Graduation Rate Recommendations

- *Continue to include only the four-year adjusted cohort graduation rate and the five-year extended graduation rate.*
- *Continue the current influence of each component. The four-year rate should have double the influence of the five-year rate.*

Post-Secondary Readiness

Under ESSA, states are permitted to include at least one **indicator of school quality or student success (SQSS)** that meaningfully differentiates between schools and is valid, reliable, statewide, and comparable. The Task Force considered multiple candidate indicators for SQSS and ultimately focused on indicators associated with post-secondary readiness. An emphasis on post-secondary readiness is appropriate given the prominent emphasis that college and career readiness receive in the *Blueprint*. For this indicator, college and career readiness is a component of the large category of post-secondary readiness.

Members acknowledged that the Task Force should prioritize efforts to ensure the school accountability system was aligned with the priorities in the Blueprint, which include the following:

- Meeting or exceeding the English Language Arts and Math performance standards on state assessments
- Earning credits in advanced courses such as AP, IB, or dual credit or completing Career and Technical Education (CTE) opportunities such as an apprenticeship or industry certification

- Completing required courses for graduation and earning a high school GPA of 3.0 or better
- Demonstrating 'on-track' to readiness by earning sufficient credits in required courses in each year of high school
- Cultivating and exhibiting a wide range of success skills, such as collaboration and healthy work habits

In particular, the Maryland State Board of Education adopted a definition of readiness that must be prominent in the state's school accountability system. The Board set a standard for readiness by the end of grade ten that requires students to earn a high school GPA of 3.0 or higher and either earn a C or higher in Algebra I or score proficient on the Algebra I MCAP. Alternatively, the standard for readiness can be achieved by scoring proficient or above on the ELA 10 and Algebra I MCAP assessments. This standard is illustrated in Figure 2.

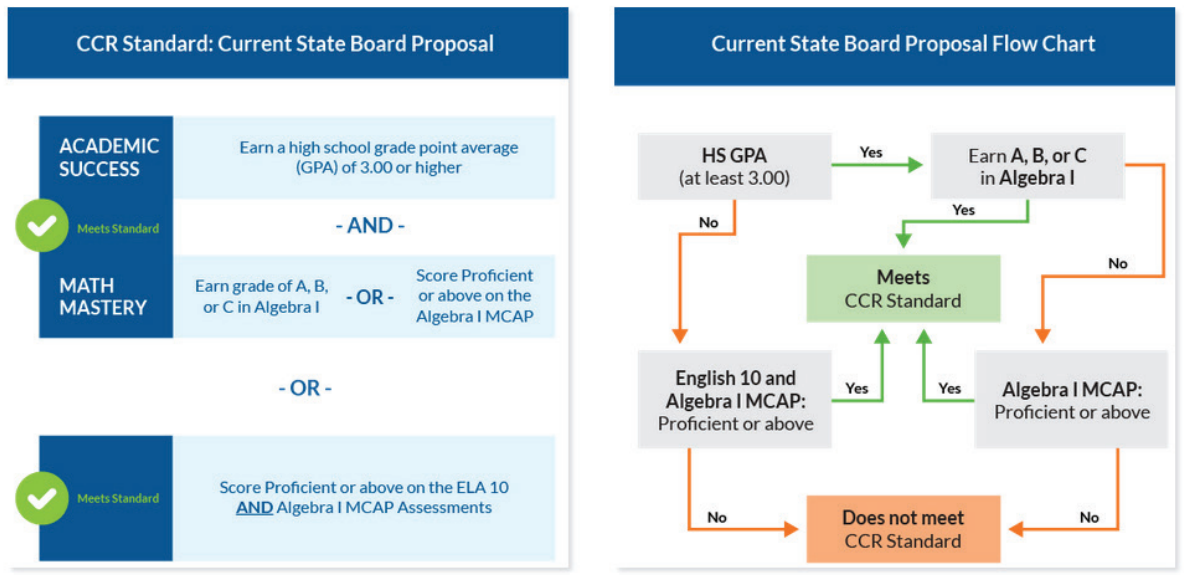


Figure 2. Maryland State Board of Education Readiness Standard

With the *Blueprint* and the Board's definition of readiness as a foundation, the Task Force reviewed multiple state models of readiness to better understand what components were included and how performance thresholds were defined.

During their review and discussion, the Task Force developed key design priorities for the CCR indicator. The indicator should:

- provide a choice among options both to account for differences in students' post-secondary interests and to account for variations in school course or program offerings
- produce some 'along-the-way' information about readiness – not just a single rating or score at the end of high school
- be sensitive to degrees of difference instead of being an 'all or nothing' designation

- incentivize both participation and performance
- incentivize schools to engage students in multiple post-secondary opportunities
- reward schools for helping students meet essential performance requirements associated with these post-secondary opportunities.

The Task Force developed a framework for the post-secondary indicator represented in Table 7 after discussing it at multiple full Task Force meetings and three separate subcommittee meetings.

Table 7. Proposed Post-Secondary Readiness Framework

INITIAL COLLEGE AND CAREER READINESS (Tied to Maryland's Blueprint)		POST HIGH SCHOOL PREPARATION OPPORTUNITIES	
ON-TRACK	**College Career Ready**	**Post-Secondary Engagement**	**Post-Secondary Performance**
Percent of 9th grade students 'on-track'	Percent of 12th-grade students who have met the CCR blueprint criteria as determined by the State Board of Education.	Complete a qualifying CTE pathway **OR** Complete at least two advanced courses (i.e., AP, IB, dual credit)	Meet performance threshold on at least two qualifying advanced courses: AP 3+, IB 4+, DC B+ **OR** Earn an industry-recognized certification or complete an apprenticeship. **OR** Meet readiness threshold on ACT, SAT, or WorkKeys **OR** Earn a Seal of Biliteracy

This framework includes four components. Schools are awarded varying amounts of credit (points) for:

- The percentage of students who earn sufficient credits at the end of ninth grade to graduate on time
- The percentage of students who have met the state's standard for CCR established by the State Board
- The percentage of students who have participated in opportunities for postsecondary preparation based on a flexible menu of qualifying options
- The percentage of students who have achieved key post-secondary preparation accomplishments associated with post-secondary success in college or careers by the end of high school

This model provides an approach for students to demonstrate a variety of accomplishments associated with the state's priorities for college and career, and post-secondary preparation. Schools receive credit/points for student accomplishments in the four columns.

The Task Force emphasized that the performance expectations in the post-secondary performance column should be comparable in rigor. For example, the WorkKeys standard should be relatively comparable to the ACT or SAT standard; otherwise, one test could be emphasized over others. The Task Force stressed the importance of additional study to establish the points and performance thresholds.

Moreover, there are likely achievements not reflected in the framework that should be added. The framework described for this indicator should be regarded as dynamic. The Task Force supports ongoing investigation to ensure the framework represents prioritized experiences and accomplishments. The Task Force noted that the specific amount of credit for each column must be worked out with the technical advisory committee and other key constituents.

Post-Secondary Readiness Summary

The Task Force developed a high-level framework for incorporating post-secondary readiness in the state's revised school accountability model. The framework is designed to incentivize on-track readiness in grade 9, achieving the Blueprint CCR standard in grade 10, and promoting ongoing preparation for post-secondary opportunities and performance throughout high school. By doing so, the framework provides information about readiness at multiple points in high school and provides credit for students at different levels of readiness.

More work is needed to refine and implement the framework, especially related to completeness and comparability.

Post-Secondary Readiness Recommendations

- *Adopt the framework developed by the Task Force to include on-track, readiness, and post-secondary preparation in the school accountability model*
- *Continue to review and refine the framework to ensure the accomplishments included in the indicator are complete and appropriate, the performance expectations associated with similar outcomes are comparable in rigor, and the overall influence (i.e., points and weights) are suitable.*

Progress in English Language Proficiency

Progress toward English language proficiency (ELP) is another required accountability indicator under ESSA. States can determine the definition of English proficiency, its statewide ELP assessment, and how ELP progress is included in its accountability system. The inclusion of ELP in the accountability system is intended to prioritize support for the development and acquisition of the English language skills necessary to succeed in K-12 and beyond.

In addition to discussing the ELP indicator during full Task Force meetings, a subcommittee met separately to examine the state's approach in more detail and identify strengths and areas of improvement.

The Task Force agreed that the state's ELP assessment, WIDA's ACCESS for ELLs, is appropriate and defensible. ACCESS is an established assessment with a strong evidence base, and it addresses all four important domains for language learners. Moreover, the performance threshold was recently adjusted from 5.0 to 4.5, representing an exit standard that should be maintained.

Regarding challenges, the Task Force pointed out that it is difficult for older language learners new to U.S. schools to demonstrate proficiency in a limited time frame. Similarly, obtaining the exit standard is very challenging for students who experience interrupted learning opportunities. It is important to better understand all the factors that influence the trajectory and time frame for language learning and include that in the model.

Another challenge is that the current ELP indicator is difficult to understand. The state should consider opportunities to streamline the indicator and provide more support to educators so that they can understand and act on the results.

The Task Force also discussed whether it would be possible to include additional sources of evidence for developing and attaining language proficiency apart from ACCESS results. It is unlikely that including additional evidence will meet ESSA requirements, but it may be beneficial to identify strategies outside of formal accountability.

Similarly, the state should consider other strategies to expand its support of language learners beyond accountability. These may involve sharing research, curating information about promising practices, and supporting professional development.

Progress in English Language Proficiency Summary

The state should continue using WIDA ACCESS to provide information about developing and attaining English language proficiency. The current ACCESS exit standard of 4.5 is appropriate. The state should also look for ways to streamline the indicator and support appropriate interpretation and use.

Additionally, the state can support English language proficiency outside of formal accountability, including for older language learners and students with interrupted learning opportunities. This includes identifying additional sources of evidence to indicate student progress toward and attainment of proficiency and sharing promising practices to help support student success.

Progress in English Language Proficiency Recommendations

- *Continue to use WIDA ACCESS with an exit standard of 4.5*
- *Conduct additional research on the conditions and time to exit to inform potential adjustments to the ELP indicator*
- *Supplementing information from WIDA ACCESS with other sources of evidence will help support student success*
- *Focus on communication and support to help make information more actionable.*

School Quality and Student Success in Grades 3-8

ESSA requires that state accountability systems include one or more indicators of school quality and student success (SQSS) for grades 3-8 and high school. In high school, SQSS is addressed through the post-secondary readiness indicator described in a previous section. However, time constraints prohibited the Task Force from addressing the SQSS indicator in grades 3-8 at a similar level of detail. Instead, the Task Force briefly reviewed the existing elementary and middle school SQSS indicators and shared feedback to inform the next steps. That feedback is summarized below.

One current SQSS indicator is *not chronically absent,* which reflects the percentage of students who are not absent 10% or more of the school days. The Task Force affirmed that this indicator is crucial as it communicates the value of attendance. However, chronic absenteeism alone is relatively

coarse insofar as students are classified as chronically absent or not in one of two conditions. Task Force members proposed investigating approaches to provide more fine-grained information, such as factoring in attendance and chronic absenteeism rates. Moreover, the state should consider ways to reward progress in improving attendance rates. For example, the indicator could be structured to reward attaining high attendance rates or demonstrating substantial improvement in attendance. Finally, Task Force members acknowledged the importance of clear communication and support to help districts and schools implement best practices for supporting attendance.

Task Force members also briefly discussed the well-rounded curriculum indicator, which measures the percentage of students in grades 5 and 8 who are enrolled in selected courses. Some members expressed concern that this indicator was of limited value, reflecting required course-taking practices and providing little to no differentiation of school performance. Others suggested the indicator is valuable to further ensure schools enroll students in important courses. The Task Force agreed that more study is needed to determine if or how this indicator should continue. In particular, it's essential to identify appropriate courses to represent a 'well-rounded curriculum.'

Finally, the Task Force discussed the *Maryland School Survey*. The survey is administered to students and educators, providing feedback on safety, environment, community, and relationships. Feedback on using the survey as a SQSS indicator was mixed. Some noted that the feedback was useful. Others expressed concern that the measure is redundant with other surveys and questioned whether the sample of respondents was sufficiently large and representative. Additionally, some Task Force members suggested that reporting could be improved to ensure the results are presented more clearly and are provided more quickly.

Lastly, the Task Force noted that the current 3-8 SQSS indicators draw heavily on results from grades 5 and 8. They advised exploring alternatives that better represented performance across grades 3-8.

ACCOUNTABILITY DESIGN DECISIONS

There are at least two levels of meaning-making associated with the accountability system. Users must be able to understand and use the information associated with each indicator (e.g., achievement) and make sense of the system overall. Further, federal education law requires that the state use the information from the indicators to make three main types of decisions:

- Comprehensive Support and Improvement (CSI), which are the lowest performing 5% of schools receiving Title I funds and high schools with graduation rates less than 67%
- Targeted Support and Improvement (TSI) identifies schools with consistently underperforming student groups
- Additional Targeted Support and Improvement (ATSI) identifies schools chronically underperforming student groups.

Indicator Reporting

The Task Force first discussed how they wanted to report the indicator values. The indicator values are all on different scales. For example, mean SGPs tend to range from 30 to 70, percent proficient could range from 0 to 100, and graduation rate generally ranges from 50 to 100. We can use algebra to combine them into a total score, but such approaches may be difficult to understand and inhibit confidence and trust in the system.

The current Maryland system uses an approach to put the indicator values on a somewhat common scale by dividing the points earned by the school for that indicator by the total points available for that indicator. Theoretically, these proportions can be compared across indicators, but users must still determine whether 50 percent of the total available points are good, average, or bad.

The Task Force discussed another approach used in many states, whereby each indicator's values (scores) are converted into a common scale. Many states using this approach have adopted a 1-4 score scale. In this case, users do not have to guess whether a particular score is good or bad. Instead, it is easy to understand that a 4 indicates good performance and a 1 indicates poor performance. Establishing these levels requires convening a group of content experts and other key users to engage in a deliberative process to establish the scores on the indicator values that divide the distribution into performance levels (i.e., cutscores). These performance levels help users quickly make sense of the indicator values and understand the strengths and weaknesses of a particular school's performance. An example of this type of approach is seen in Table 8 below.

Table 8. An Example of Converting Indicator Scores into Indicator Performance Levels

ACHIEVEMENT		ELP	
LEVEL	Score Range	Level	Score Range
1	0.0–2.10	1	< 50
2	2.11–2.59	2	50–59
3	2.60–3.00	3	60–69
4	3.01–4.00	4	> 69
GROWTH		**EQUITY**	
Level	Score Range	Level	Score Range
1	1–40.00	1	< 45
2	40.01–49.99	2	45–54
3	50.00–60.99	3	55–65
4	61.00–99.99	4	> 65

Overall Determinations

As noted above, MSDE must produce at least three types of overall school determinations: CSI, TSI, and ATSI. Most states believe they must calculate an overall score based on multiple indicators. However, this is not true. The state is not required to calculate a total score to produce these determinations.

Before delving into methods for producing overall determinations, the Task Force deliberated what and how it wanted to communicate. The facilitators asked Task Force members to produce rough

sketches of accountability home pages to support the discussion. In other words, the Task Force was trying to envision what constituents would see when they first looked at a school's accountability report. Would they see an overall school grade or other designation (e.g., stars) or indicator reports?

Some group members suggested that the first view should draw readers' attention to the school's performance on a limited number of critical indicators tied to valued state initiatives such as early literacy and those related to the Blueprint. One of the Task Force members pointed to Indiana's website as an example of this approach.

While the Task Force members did not rule out producing an overall score or performance designation, they emphasized that the first view into a school's performance should focus on a limited number of indicators, not an overall grade.

Combing Multiple Measures

There are several general approaches for combining multiple measures or indicators to arrive at an overall inference or decision. These four approaches are described in Table 9 below. Disjunctive approaches are not permitted under federal law because that would mean if a school performed well on any one indicator, it would receive a positive overall rating. Therefore, the Task Force discussed compensatory, conjunctive, and profile methods.

Table 9. Methods for Combining Multiple Measures

METHOD	DESCRIPTION	EXAMPLE
Compensatory	Higher performance on one indicator can offset lower performance on another.	Index or weighted composite, such as GPA or most course grades.
Conjunctive	The overall score can be no higher than the lowest indicator score, meaning that performance on ALL indicators counts equally.	NCLB methods (i.e., all groups must be proficient in all grades and content areas)
Disjunctive	Performance on ANY indicator provides the overall decision (highest score counts).	Menu approach (i.e., show college or career readiness by demonstrating one of the following…)
Profile	Define specific patterns regarded as sufficient for entry or exit into a classification.	Determining the patterns of indicator performance that demonstrate sufficient overall performance, such as done with the National Board of Professional Teaching Standards

To help ground the discussion, Task Force members viewed examples from states that illustrated the various approaches. These examples are shown below.

Compensatory

The compensatory example presented below from Illinois portrays the weights assigned to the indicators when combined into an overall average for each elementary/middle and high school.

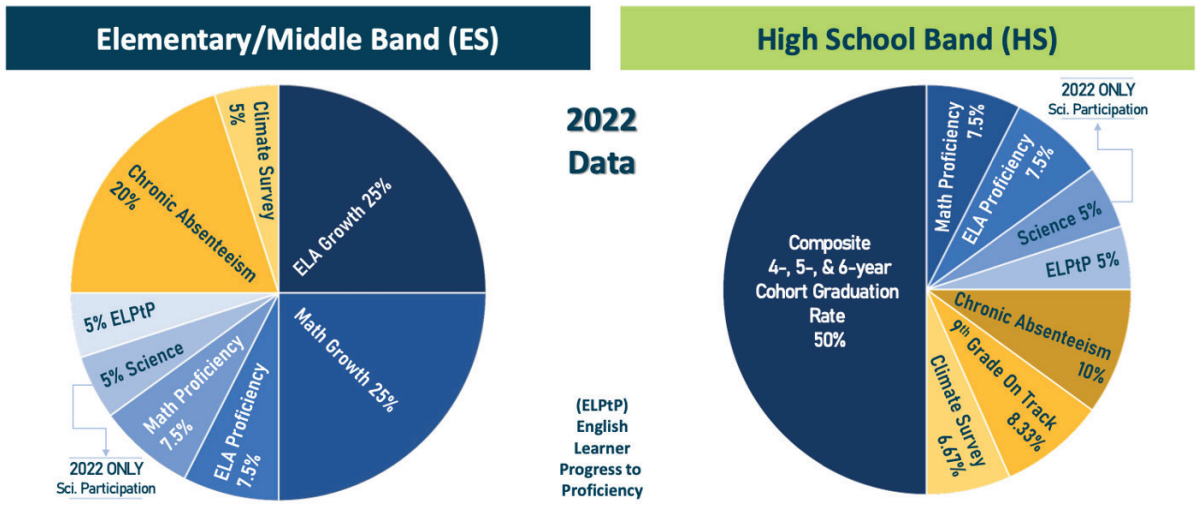


Figure 3. Illustration of Illinois' Compensatory Method for Combining Indicators

Conjunctive

The example below from Rhode Island appears to be a profile method but is a conjunctive model. The overall rating provided to schools is no higher than the number of stars associated with the school's lowest-rated indicator. For example, suppose the school received 1 point on the English language proficiency indicator. In that case, its overall rating cannot exceed two stars no matter how well the school performed on the other indicators.

School Classification Rules*							
ELA Achievement, Math Achievement, and Science Proficiency (Max. 11 points)**	**Growth: ELA and Math (Max. 6 point)**	**English Language Proficiency (Max. 4 points)**	**Graduation Rate (HS Only) (Max 5 points)**	**Commissioner's Seal and Post-Secondary Success (Max 6 points)*****	**Exceeds (ELA/Math) Absenteeism (Student/Teacher) and Suspension (Max. 15 points)********	**Targeted Support and Improvement: Subgroups**	**School Rating**
9 or more points (3 or 4 points each)	4 or more points (2 or 3 points each)	3 or more points	4 or more points	5 or more points	12 or more points	None identified	★★★★★
7 or more points (2-4 points each)	4 or more points (2 or 3 points each)	2 or more points	4 or more points	4 or more points (2 or 3 points on each)	10 or more points	1 identified subgroup maximum	★★★★
9 or more points		2 or more points	3 or more points	3 or more points	7 or more points	Could have multiple identified subgroups	★★★
6 or more points		1 or more points	2 or more points	2 or more points	5 or more points	Could have multiple identified subgroups	★★
3 or more points (1 point each)	2 or more points (1 point each)	1 or more points	1 or more points	2 or more points	5 or more points	Could have multiple identified subgroups	★

Figure 4. Illustration of Rhode Island's Conjunctive Method for Combining Indicators Profile

The example from New Hampshire below shows how the profile system is used to produce determinations for comprehensive support and improvement (CSI), targeted support and improvement (TSI), and additional targeted support and improvement (ATSI). CSI requires the state to identify the lowest performing 5% of schools that receive Title I funds. In this case, a deliberative body determined that school scoring at Level 1 for all indicators would be identified, which is a relatively easy decision. If 5% of the schools were not identified in this step, the accountability team would move through the steps until 5% of the schools were identified. Notably, a deliberative body engaged in a standard-setting activity to determine the specific constellation of indicator levels that would lead to identification.

Step 2 — **Identify the final step for each school by using the indicator levels and the decision matrices below**

ES/MS

Step	Achievement	Growth	ELP	Equity
1	Level 1	Level 1	Level 1	Level 1
2	Level 1	Level 1	Level 2	Level 1
3	Level 1	Level 1	Level 1	Level 2
4	Level 1	Level 1	Level 2	Level 2
5	Level 2	Level 1	Level 2	Level 2
6	Level 2	Level 2	Level 2	Level 2

HS

Step	Achievement	Graduation Rate	ELP	CCR
1	Level 1	Level 1	Level 1	Level 1
2	Level 1	Level 2	Level 1	Level 1
3	Level 1	Level 2	Level 2	Level 1
4	Level 1	Level 2	Level 1	Level 2
5	Level 2	Level 2	Level 2	Level 2

Figure 5. Illustration of New Hampshire's Profile Method for Combining Indicators

Considerations for Compensatory Systems

Compensatory approaches can be more complex than either conjunctive or profile approaches. Therefore, the Task Force discussed some additional considerations associated with compensatory systems.

There are several ways to determine the weights for each of the indicators. One way is based on policy priorities (e.g., we want growth to count 50% of the weight). Another way is statistically based to maximize the reliability of the system. Reliability, in this case, refers to the consistency with which schools would be identified or not identified if we knew the "truth." Like many other issues, designers may choose to balance the identified policy and statistical goals. In any case, these choices must be explicit and transparent.

Designers must consider the difference between nominal and effective weights. Nominal weights are assigned to each indicator, usually due to a policy decision. For example, we might assign 40% each to achievement and growth, 10% to ELPA, and 10% to chronic absenteeism. These are called nominal or intended weights.

However, effective weights are what actually happens when the various indicators are combined into an overall score. The effective weight is highly related to the variance associated with each indicator. The more variance associated with an indicator, the more weight it will have in the overall score. Let's look at an extreme example. Assume there were two indicators, growth, and achievement, that we intended to weigh 50-50. Also, assume that every school in the state had the same growth score. Effectively, 100% of the determination would be based on achievement because growth would simply be adding a constant.

Accountability Performance Levels

The Task Force endorsed creating performance levels/designations for indicators and overall if a compensatory approach is used. They thought it would be better to communicate the results to a broad range of constituents rather than providing decontextualized scores.

Performance levels should be used to define stars, numerical levels, or grades used in many compensatory systems. Unfortunately, many states have converted numerical averages into grades or stars by assigning points to grades arbitrarily by treating school performance like student test scores (e.g., 90% = A, 80% = B, etc.). Performance standards are the more appropriate way to answer the question, "What's good enough to achieve a designated score or rating?"

Many well-developed approaches to setting assessment standards have been applied to accountability systems. When done well, accountability standard-setting reflects policy priorities, is informed by the judgments of a broad group of experts and constituents, is guided by relevant information, including consequences, and is transparent and well-documented. The defensibility of performance standards is strongly linked to the process.

Aggregation and Determinations Summary

The Task Force spent considerable time discussing what should be reported to various accountability audiences and how they want it reported. The Task Force started with indicator scores, such as for achievement and growth. They considered several ways to report indicator values, including the current system that reports indicator performance as the percentage of total points earned for each indicator. The Task Force also discussed converting indicator values (e.g., percentage of students scoring proficient in a school) into performance levels. They viewed an example from another state

where all the raw indicator scores were converted into one of four performance levels (see Table 8 and Figure 5 above). The Task Force appreciated that users are not left wondering if a particular score is good or bad, as can be the case with the current system.

The Task Force members had mixed preferences for producing overall determinations. Many members preferred a profile approach, like that used in New Hampshire and New York, while others preferred a compensatory method that uses a weighted average, like the example from Illinois. However, Task Force members indicated that if a compensatory method is used, the state should convert the resulting scores into a performance level, such as grades or numeric levels similar to the indicator levels. The Task Force opposed using stars because they thought breaking from the current system was vital. Several Task Force members indicated they wanted a compensatory approach "with some profile sprinkled in." After elaboration, this could be addressed by reporting total scores using the same 1 through 4 scale as the indicators, where, for example, a school's average performance could be a 3.2.

There was considerable discussion and debate about how various users would interpret a profile compared with an overall determination approach. After the discussion went on for some time, it was suggested that the Task Force should recommend a small-scale study to understand how users interpret profile reports compared to something like school grades.

Aggregation and Determinations Recommendation

Establish common performance levels for indicators (e.g., 1-4) using a deliberative process with experts and key constituents.

Conduct a small-scale study to determine whether the accountability system users arrive at the intended interpretations when presented with reports derived using a profile method compared with those derived using a weighted average and overall rating.

Following this study, whatever decision-making process is endorsed, the Task Force recommended employing an accountability standard-setting process to guide the federally required determinations and to establish performance levels if overall performance levels are desired.

ACCOUNTABILITY IMPLEMENTATION GUIDANCE

The recommendations outlined in this report provide a framework for Maryland's accountability system. Moving from design to implementation, MSDE should consider the following:

Establish Operational Definitions and Business Rules

The Task Force's recommendations address features, priorities, and acceptance criteria associated with the indicators and system design but do not establish the operational definitions. For example, more work is needed to define the final set of accomplishments in the post-secondary readiness indicator and to determine if adjustment to the timeline for exit should be adjusted for the ELP indicator. Additionally, business rules for each indicator need to be reviewed and defined (e.g., the minimum number of students required to report an indicator). This is understandable, given that the Task Force was formed as a policy advisory group, not a technical advisory group. In subsequent phases, MSDE should work with subject matter experts, technical advisors, practitioners, and other constituents to further specify and implement the system.

Establish Aggregation Rules and Performance Expectations

As noted in the design decisions section, some essential ongoing steps are needed to determine how indicators should be combined to inform overall designations. Moreover, MSDE must finalize rules for reporting overall and indicator performance. Some of these decisions are more technical, such as determining whether and how to scale indicators so they honor the intended weights. The state's Technical Advisory Committee (TAC) should be consulted for these decisions. In other cases, these decisions are more policy-focused, such as determining "good enough" performance for classification categories. For these decisions, an accountability standard-setting process will help ensure that performance expectations are associated with meaningful criteria and not based on arbitrary norms. Finally, before the reporting methods are finalized, MSDE should conduct a small-scale study with constituents to ensure they can effectively arrive at the intended interpretations when presented with reports.

Address Exceptions

Every accountability system must address exceptional circumstances and conditions. For example, how are schools with unusual grade configurations (e.g., K-2), special student populations, and/or small schools addressed? Determining business rules for these and other exceptional circumstances is vital to the development and implementation process. The Task Force could not have an in-depth discussion about exceptional circumstances and conditions that affect some schools. These conversations could be part of a subsequent implementation phase.

Examine and Refine

Once additional specifications have been established, MSDE and its partners should examine performance on indicators and overall classifications to better understand the extent to which the system supports the intended interpretations and uses. Research questions might include:

- Are indicators and classifications sufficiently reliable (stable) and accurate?
- Do indicators and overall results meaningfully and appropriately differentiate school performance? For example, are accountability results in sync with other sources of credible evidence regarding school performance?
- Are indicators and overall scores fair to all schools? For example, are scores correlated with factors that should not be associated with performance (e.g., school size)? Can schools from different regions or that serve demographically diverse students access favorable outcomes?

The MSDE TAC and other partners may be able to help explore these and other questions to inform system refinements and continuous improvement.

The state assessment system and, to a lesser extent, the state accountability system are regularly reviewed by the MSDE Technical Advisory Committee (TAC). The Task Force recognized this critical function but recommended regularly convening a policy- and practitioner-oriented advisory committee to provide feedback on the implementation of these two systems.

ASSESSMENT RECOMMENDATIONS

The Task Force discussed key aspects of assessment design and implementation and offered recommendations to address the following critical questions associated with a state assessment program.

- **Accessibility and Fairness:** How can the MSDE help ensure assessments are fair and accessible to a broad range of learners?
- **Adaptive or Fixed Form:** Will the test be administered to students using a computer adaptive testing process or a "fixed form" approach?
- **Testing Time:** How much time should be required for state summative testing, and what types of items (questions) should be included on the test?
- **Score Reporting:** How should the system of score reports be designed to support high-quality and understandable information for various users in the educational system?
- **Non-Summative Resources:** Should the state procure non-summative resources (e.g., interim assessments, formative assessment tools) as part of the summative assessment RFP?
- **Communication, Outreach, and Advocacy:** How should MSDE design and execute a communication plan to enhance the credibility and usefulness of the state assessment system?

ACCESSIBILITY AND FAIRNESS

Before discussing any design decisions, the Task Force clarified that all assessments must be designed so that all students can show what they know. While Task Force members included individuals with expertise in teaching students with the most significant cognitive disabilities, no concerns were raised about the alternate assessments used in Maryland (Dynamic Learning Maps and WIDA Alternate ACCESS). Therefore, this report's primary focus has been on recommendations for the general assessments taken by most Maryland students. Features of the assessment (e.g., extra wording in the questions) or the administration platform should be designed so that all students can access the assessment without any barriers to their performance. The Task Force expressed frustration with the current assessment system where students needing accommodations must take a single fixed-form version of the test while all other students participate in an adaptive test. There is almost no excuse these days not to include all students in a universally designed testing experience to the fullest extent possible. Therefore, the Task Force issued the following recommendations:

- *All MCAP tests must be designed using the most up-to-date research to ensure that all students can demonstrate their knowledge and skills without barriers.*
- *All MCAP tests must be evaluated from the design process through the results to ensure the testing program is as fair as possible for all student groups and does not privilege any group over others.*

COMPUTER ADAPTIVE OR FIXED FORM TESTING

Tests are administered and scored in a variety of ways. The general approaches discussed by the Task Force were fixed-form tests and two types of computer adaptive testing.

1. **Fixed form:** These are the types with which most people are familiar. All students in each subject/grade/course are administered essentially the same test items in the same order. That is, the test looks the same for every examinee. There are variations on the "same" because field-test and sometimes linking items are embedded differentially in the test forms, but, in general, the students in a grade/subject are administered and scored on the same set of items.

2. **Computer adaptive testing (CAT):** These tests rely on the power and speed of modern computers (even though CAT is decades old) and Item Response Theory (IRT) to adjust the test questions presented to each student or group of students. The adaptation is based on students' performance on previous items. There are two main types of adaptations: item-by-item or in stages.

    a. **"Item-level CAT"** tailors each student's test to their level of achievement as determined by their performance of all prior test items they answered after the first item. In an item-level CAT, each item presented to a student is based on the student's performance on the previous item(s) and the difficulty of those items. For example, if a student responds incorrectly, the following item will be easier than the last. The next item will be more difficult when the student answers an item correctly. Figure 6 illustrates the operation of this adaptive principle.

    b. **Multi-stage testing (MST)** involves administering a pre-determined set of items to students in various stages, usually two or three. All students in a specific grade/course and content area complete the same Stage I form. Based on their performance on Stage 1, they are then routed to easier or harder (or moderate) forms in subsequent stages to best measure a student's actual achievement level as precisely as possible. See Figure 7 below.

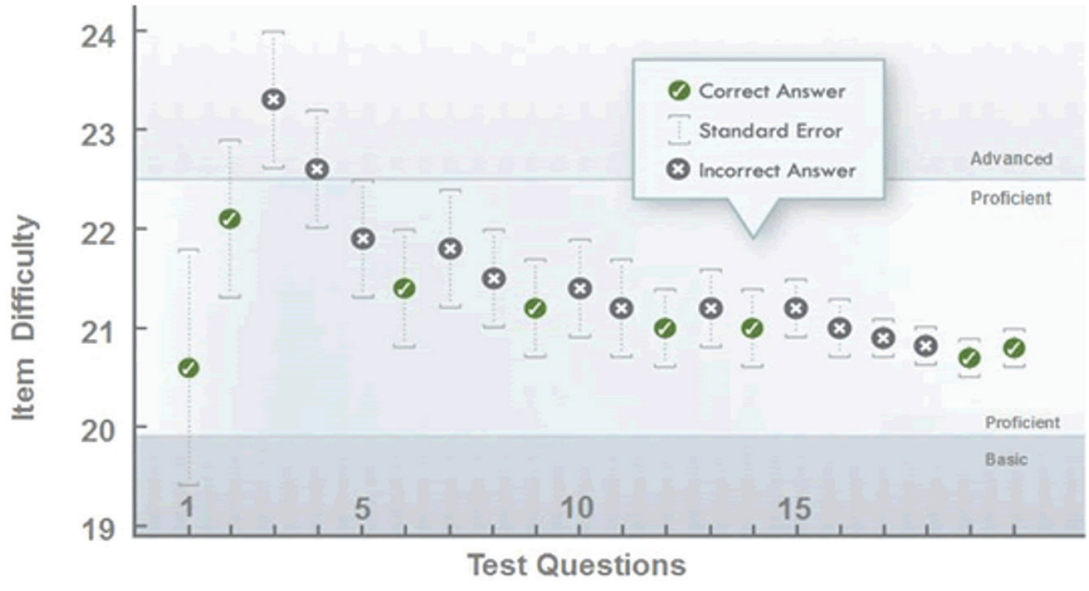


Figure 6. An example of item-level computer adaptive testing.[1]

[1] From http://www.ascd.org/publications/educational-leadership/mar14/vol71/num06/The-Potential-of-Adaptive-Assessment.aspx

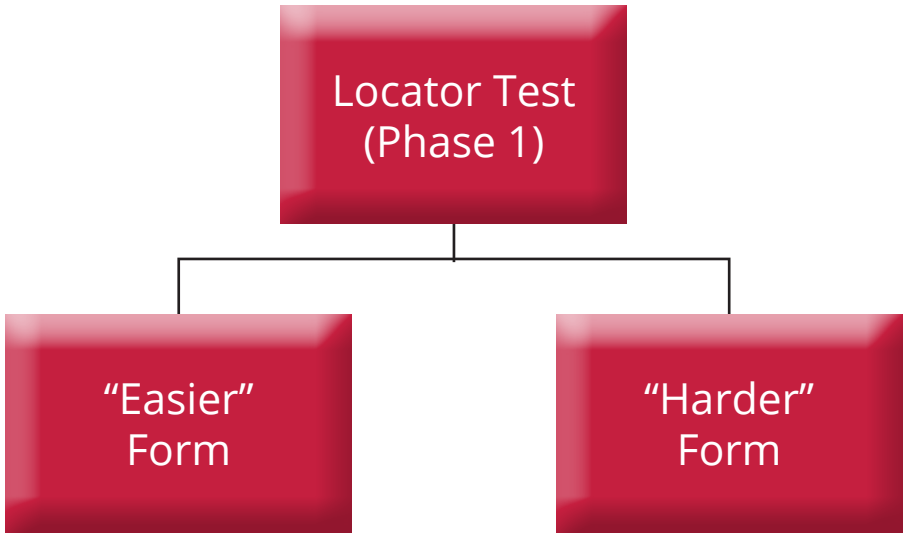


Figure 7. An example of multi-stage computer adaptive testing.

The Task Force discussed each approach's potential advantages and disadvantages, which are presented below.

Fixed Form

Advantages

Fixed-form tests offer many advantages.

- The test form assembly is tightly controlled, and content experts and test developers can ensure that every item on the test is aligned to the appropriate learning targets.
- Quality is assured and is transparent because test developers can see every item students will experience.
- Fixed-form tests need a considerably smaller item bank than adaptive tests. Since item development is a major cost driver of testing contracts, fixed-form tests are the least expensive option. Further, Maryland already possesses a relatively healthy item bank, which could help reduce these costs considerably.
- Fixed-form tests generally allow for a greater variety of item types than adaptive tests, particularly those that require human scoring or allow students to respond in less constrained ways than what is available in computer-based environments.
- Fixed-form tests allow for efficient use of released-item reports because every student experiences the same test items for a given test. A released-item report is one in which a subset of the test items is released each year, and teachers get information on how the students in their class performed on each released item. Seeing actual test items and how their students responded helps clarify the grade-level learning expectations. There is also the risk of encouraging teachers to "teach the test," but this can be ameliorated with clear guidance and professional learning opportunities.

Challenges

Several challenges associated with fixed-form tests help explain the growing popularity of computer adaptive testing.

- The primary challenge associated with fixed-form tests is that students may face a fair number of items that are too hard or too easy for them. This can be frustrating for students and dampen motivation. This occurs because test developers focus on ensuring that the most information (i.e., the number of test items) is associated with the most critical score points on the test, generally around the cutscore separating Level 2 from Level 3 performance (i.e., proficient), as well as around the score distribution associated with the Level 1-2 cutscore.
- Since measurement error—the uncertainty associated with every measurement activity—is associated with the amount of test information at a particular point along the score scale. Since information is generally associated with the number of items, it stands to reason that fixed-form tests contain considerably more error (uncertainty) at the upper and lower ends of the scale than adaptive tests.
- Student longitudinal growth measures have more uncertainty than single point-in-time "status" measures as a function of the error on each test. Therefore, because fixed-form tests have more uncertainty associated with high and low scores than adaptive tests, the growth measures for the lowest- and highest-scoring students will be less reliable (more uncertain) than adaptive tests.
- Finally, because all students are tested with the same items, a security breach (e.g., items or an entire form is exposed) can threaten the entire program for that year. Of course, testing programs using fixed-form approaches have several "breach" forms in waiting should such a breach occur. Further, the measurement field has developed many processes and tools to minimize these threats. Nevertheless, the risk is greater for fixed-form compared to adaptive tests.

Item Adaptive

Computer adaptive testing, or CAT, has generally been operationalized as item-level CAT in the many years it has been used. CAT has proliferated over the past 25 years but has expanded considerably since 2014 due to the Smarter Balanced Assessment Consortium, funded through the Race-to-the-Top program. CAT offers advantages that have led to its growing use, but it also carries some challenges. The advantages of adaptive tests offset many of the disadvantages of fixed-form tests, but the converse is also true. The Task Force discussed both, as described below.

Advantages

- The major advantage of item-level CAT is that the testing experience is tailored to each student's achievement level. This helps offset several of the notable challenges associated with fixed-form tests.
- Importantly, item-level CAT provides relatively precise (i.e., lower levels of uncertainty) score estimates for high and low-achieving students, which helps when tests are used for measuring student longitudinal growth.
- Generally, item adaptive tests have high levels of test security since each student is theoretically taking a unique form of the test.
- Theoretically, item adaptive testing allows for a more efficient (e.g., faster) testing experience than fixed-form testing. However, when these tests are part of federally required state

assessment systems, they must meet strict alignment requirements, ensuring that items are included to measure the full range and breadth of the state's content standards. This requirement essentially eliminates any potential efficiency differences between adaptive and fixed-form tests.

Challenges

Many challenges are associated with item adaptive testing, some of which are fairly significant.

- Item adaptive tests are item hungry. That means they require a large item bank to have the system work as intended (i.e., adapting to each student). Since item development is one of the major cost drivers for testing programs, more items means higher costs.
- Reading tests are designed to have students respond to questions based on several reading passages. A set of specific questions is tied to each passage, which is the right way to test reading, but it also limits the item adaptivity of reading tests.
- Since each student theoretically completes a unique set of items, determined by their responses to items throughout the test, item adaptive tests have a degree of obscurity since the items a student completes are not visible to anyone other than the student.
- Item adaptive tests require items that can be scored very quickly (instantly) to decide what item the student will see next and avoid having the student wait too long. This may limit the types of items available for the test, especially precluding items or tasks that require human scoring.
- Some might perceive item adaptive tests as unfair because not all students can try the most challenging items. This is true but is a crucial part of adaptive testing design.

Item adaptive testing produces the most precise scores throughout the achievement distribution and potentially the shortest test. If done well, it minimizes the exposure of items more than other types of adaptive testing. However, it requires the most investment in up-front item development and the largest pool of items with appropriate ranges of difficulty and complexity.

Stage Adaptive Testing

As discussed above, stage adaptive testing carries many advantages over item adaptive and fixed-form testing, but it also has to address several challenges.

Advantages

- The stage adaptive testing experience is somewhat tailored to each student's achievement level. It does so by using multiple stages of discrete sets of items rather than adapting on an item-by-item basis.
- Stage adaptive testing allows for tightly controlled form assembly, which aids with transparency and quality assurance.
- Despite not adapting following each item, stage adaptive allows for relatively precise score estimates for high and low-achieving students, which helps when measuring student longitudinal growth.
- Stage adaptive testing can be more secure than fixed-form testing, especially if multiple forms are available at each stage, but it still will not be as secure as item adaptive testing.

Challenges

Like the advantages, the challenges associated with stage adaptive testing fall between the challenges for item adaptive and fixed-form testing.

- Stage adaptive tests require a larger item bank than fixed-form tests but considerably fewer than item adaptive tests.
- There is a chance that students close to the cutscore after the first stage may get incorrectly routed and have a relatively more challenging time getting high test scores compared to those routed correctly.
- Like item adaptive tests, stage adaptive tests may preclude some types of human-scored items. However, stage adaptive testing has the space at the end of each stage to include the types of rich items available for fixed-form testing.

Computer Adaptive or Fixed Form: Summary

The Task Force spent considerable time discussing the options for the type of testing platform they would recommend for Maryland's next assessment system. First, the Task Force acknowledged the advantages of fixed-form tests for the high school end-of-course testing system. The Task Force preferred stage adaptive testing for the grades 3-8 English language arts and mathematics tests. Notably, they opposed item-level adaptive testing for the ELA and math tests in grades 3-8. The Task Force viewed fixed-form favorably but did not think they carried all of the potential benefits of stage adaptive tests.

Computer Adaptive or Fixed Form: Recommendation

The Task Force recommends a system that allows MSDE to document the quality of every test form administered to students. Further, the Task Force recommends releasing a subset of test items each year to enhance reporting, credibility, and usefulness in terms of helping educators and students understand the level of knowledge and skills required to perform successfully on the tests. Therefore, the Task Force recommends that MSDE encourage bids through the RFP process that rely on a multi-stage adaptive design. However, the Task Force recommends allowing offerors to propose an alternative design to meet the State's goals. In either case, the offeror must present evidence regarding the advantages and disadvantages of the proposed approach for the State of Maryland.

TESTING TIME AND TYPES OF ITEMS INCLUDED ON THE TEST

The types of test questions (items and tasks) included on the test are closely related to the amount of time students will need to complete it. For example, if the test required students to complete three writing prompts or similar performance-based tasks, 2-3 hours would be added to the time necessary for the rest of the test.

The discussion of item types does not start with "shopping" for different types of items or tasks. Rather, the Task Force first wrestled with essential questions that can be used to guide recommendations for the types of items to include on the test.

The most important question is, "What are you trying to measure?" The answer to this question is not as simplistic as "5th grade mathematics," for example. Rather, content experts must specify the

nature of the knowledge and skills students will be expected to demonstrate. Once this is done, test designers and item development experts identify the items best suited to elicit the targeted knowledge and skills. Again, these decisions interact with testing time, so test designers must decide what mix of item types will help the state meet its goals.

These decisions are not made in isolation. The state needs to consider the available resources and the resources required to support the desired item development. For example, items that require human scoring—of which there are not many anymore—will cost more than items that can be machine-scored. Similarly, technology-enhanced items cost considerably more than conventional items to develop and validate.

Another critical decision is considering the number of items the state will need to support its program. Avoiding item adaptive testing reduces the number of items needed and could free up resources to develop richer or more innovative items. Right now, it appears that Maryland possesses a robust item bank. Still, the existing item bank must be evaluated against potential new content standards and the cognitive demands envisioned for the next testing programs.

Testing Time Summary

The Task Force favored including a range of item types on subsequent MSDE assessments to measure the full depth and breadth of the standards. However, some group members expressed concern about technology-enhanced items because teachers would have to spend more time getting students ready for the format rather than the substance of the item. This is particularly true for younger students with less experience with these sorts of items than older students. Additionally, items that require fine motor skills (e.g., drag and drop) may disadvantage younger students and students with disabilities. Finally, while Task Force members recommended including open-response tasks on the tests to signal the types of instruction and learning the state wants to see in classrooms, they cautioned against including too many of these types of items because of its impact on testing time.

Testing Time Recommendation

The Task Force recommends including a range of item types to ensure that the full breadth and depth of the standards are well-measured. Open-response items/tasks should be designed to signal the types of tasks the Task Force and MSDE would like to see used as part of regular classroom instruction. However, this should be balanced with ensuring that the total test length is no longer than practically necessary to produce valid, reliable, and useful scores.

SCORE REPORTING SYSTEM

The compressed timeline of the Maryland Task Force process limited the time we could devote to discussing score reporting. Nevertheless, score reporting is one of the most important aspects of assessment design. As the late Ron Hambleton, a leading measurement expert, liked to remind us, "Score reports are the main way that we communicate with the public about our tests, but they are the last thing we attend to in the test design." The Task Force agreed with this sentiment and pushed for MSDE to continue creating a strong framework for a coherent reporting system.

All state assessment systems contain multiple score reports for various users, including all of the following and often more:

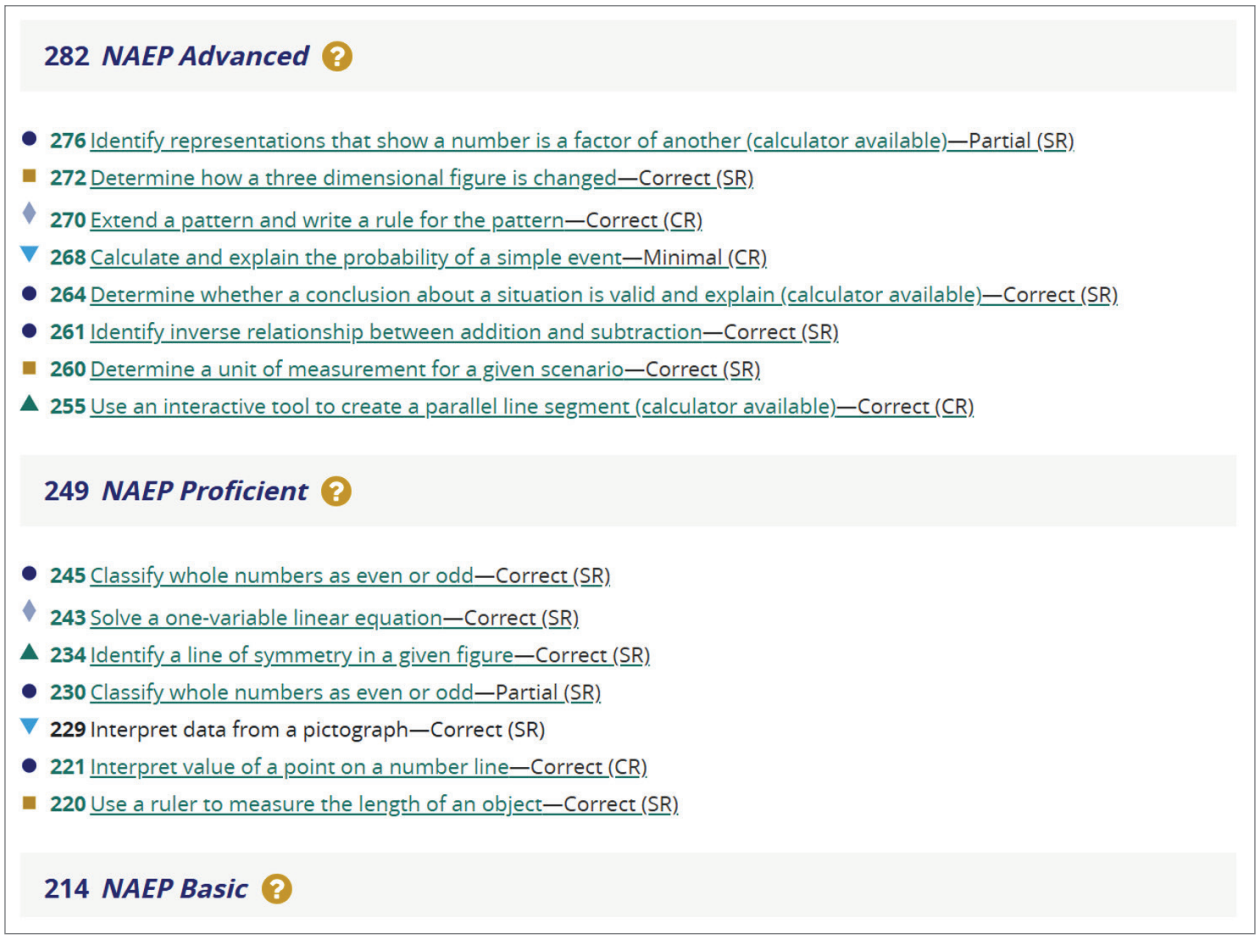
- Individual score reports (for students and parents)
- Classroom reports (for teachers)
- School reports (for school leaders, teachers, school improvement teams, and community members)
- District reports (for district leaders, school boards, and community members)
- State reports (for state leaders and state policymakers)
- Public dashboards for multiple levels of the system

Federal law (ESSA) requires many of these reports, including public reports and individual score reports, and it also requires the presence of critical elements in these reports. However, simply having all of these reports does not ensure coherence. Unless designed intentionally, the reports may provide incoherent messages. The odds are against producing contradictory messages since all the reports are derived from the same data. However, there is still a good chance that unless it is done thoughtfully, the system of reports may not be as coherent as possible.

Score reporting has undoubtedly improved over the last twenty years, but we still have a long way to go to make it understandable and actionable for each intended user group. Score reports are often designed by measurement experts who try to pack as much information as possible into the report. Bringing teachers and other users into the report-design process would help, but there are ways to do even better.

The Task Force favored including released item reports in the score reporting system. Ideally, this would be done so teachers could see the performance of each of their students on a subset of the test items. This has cost implications because replacing released items can be costly. The Task Force recommended that MSDE consider the cost when deciding how many items can be released. Still, they urged MSDE to release enough items so teachers can gain a solid understanding of the types of knowledge and skills students are expected to demonstrate.

The Task Force also discussed using item maps to enhance the public's understanding of the test and student expectations. Item maps help illustrate what students know and can do in tested subject areas by positioning descriptions of individual assessment items along the test scale at each grade level. An item is placed at the point on the scale where students are more likely to respond successfully. Figure 8 displays an excerpt from the 2022 Grade 4 National Assessment of Educational Progress (NAEP) item map.



Some of the newer reporting systems offer some "default" interpretations based on the data in the report instead of just presenting teachers with tables of numbers, even with some nice graphics. The rapid advances in artificial intelligence (AI), especially generative AI, offer considerable promise for enhancing the interpretability of score reports. MSDE and its technical advisors should continue exploring the potential of AI for enhancing score reporting. Communications experts should be closely involved in or even lead the report design process, along with assessment and content experts. This way, score reporting systems would be more apt to be designed with the user in mind. Critically, all potential report designs must be evaluated iteratively with some type of cognitive-laboratory methodology. These approaches generally ask the user or examinee to think aloud as they navigate the report, which enables designers to gain insight into how users make sense of the information in the reports and where they struggle.

The Task Force also discussed the importance of the score reporting system producing data files that can be efficiently and accurately uploaded into districts' student information and/or learning management systems. The Task Force recommended that MSDE create a comprehensive system of report interpretation and related assessment literacy professional learning opportunities for the various intended report users. These should be regularly evaluated to ensure they are supporting the intended learning goals.

Score Reporting Recommendation

The Task Force recommends that the state support the development of a coherent system of timely score reports with a clear specification of each report's intended users and uses. The report design process should be led by or at least include communications experts. The Task Force recommends that the report developers present evidence or a clear plan for collecting evidence to evaluate claims of usefulness for each of the intended user groups. The Task Force also recommended that the score information be easily uploaded to district student information systems. Finally, the Task Force recommended that MSDE support a comprehensive system of report interpretation and related assessment literacy professional learning opportunities for the various intended report users.

NON-SUMMATIVE RESOURCES

States and assessment consortia (e.g., Smarter Balanced) have been attempting to support local leaders and teachers with assessment tools and supports they can use throughout the school year to help support learning and teaching. Such tools range from conventional interim assessments administered 2-3 times a year to assessment literacy supports that teachers can use to enhance their daily formative assessment practices. Modular interim or benchmark assessments are one of the more common sets of resources currently supported by states. These relatively short tests (e.g., 8-15 items) are tied to defined knowledge and skills, often represented by a single or just a few content standards.

There are many good reasons for a state to procure these resources. They see this as a way to support the development of more balanced assessment systems than is the case with a single, end-of-year accountability test. States also envision the benefit of providing a lower-cost and more coherent option for districts in place of all of the commercial interim assessments they purchase.

However, these hopes have not always come to fruition. District leaders appear reluctant to give up their commercial interim assessments. If they also encourage teachers to use state resources, it could lead to a considerable increase in overall testing time. Additionally, because state resources are seen as part of the state testing regime, there is early evidence that they are used more as test preparation tools rather than as tools for instruction throughout the year. With this framing in mind, the Task Force discussed the following three options for recommendations associated with non-summative resources.

1. MSDE should not, at least at this time, pursue non-summative assessments (e.g., block interims) as part of a revised MCAP.
2. MSDE should include non-summative assessments (e.g., block interims) as part of a revised MCAP but make their use optional.
3. MSDE should include non-summative assessments (e.g., block interims) as part of a revised MCAP and require their use.

Non-Summative Resources Summary

Task Force members expressed a variety of opinions and preferences. Option 2 received the most support from participants. However, participants expressed concerns about whether it was worth spending money on this option because they wondered if it could be better spent enhancing the

summative assessment. Some members also raised the issues associated with less obvious costs associated with retraining staff to use a new system effectively. That said, the Task Force participants favored including a cost option in the RFP for modular interim assessments as part of the state assessment procurement.

Non-Summative Resources Recommendation

The Task Force recommends that MSDE invite potential respondents to an assessment RFP to include the development of modular interim assessments as a cost option. If MSDE exercises such a cost option, the state should support high-quality use through extensive professional learning opportunities and supporting materials. However, using these non-summative tools should be optional for school districts.

COMMUNICATION, OUTREACH, AND ADVOCACY

The importance of MSDE having a well-developed and executed communication system cannot be overstated. Many Task Force members raised concerns about the actual and perceived credibility of the statewide assessment system. They noted that MSDE's apparent lack of a comprehensive communication approach hindered sharing positive stories about the assessment system and the results. The Task Force emphasized that the communication should not just be an attempt to "sell" the assessment system but should focus on sharing interesting and helpful uses of the assessment results. Also, the Task Force suggested that MSDE should share research and evaluation results that use the state assessment results as outcomes or as another important variable in the research studies.

Communication, Outreach, and Advocacy Recommendation

The Task Force recommends that MSDE develop a comprehensive communication strategy to share positive stories about the assessment system and how schools and districts use the assessment results.

In addition to conducting internal research, the Task Force recommends that MSDE facilitate the use of Maryland assessment and related data for research to address policy-related and other vital research and evaluation questions.

SUMMARY

The MSDE Assessment and Accountability Task Force met regularly for over seven months in 2024 to deliberate and make recommendations to improve Maryland's assessment and accountability systems. The recommendations presented in this report provide meaningful guidance for MSDE as it prepares to release a Request for Proposals for its next assessment system. The recommendations for improving the accountability system will provide valuable advice to MSDE as it creates the business rules to operationalize the new vision for school accountability in Maryland.

MSDE will implement many of the accountability recommendations for the 2024-2025 accountability results, contingent upon federal approval. Other recommendations, such as determining which growth model to use, will require study and analyses early in 2025 to have the information necessary to decide on the growth model by late spring 2025. The new or revised growth indicator will not be

implemented before the 2025-2026 school year. Changes such as aggregation approaches and producing annual determinations will be on a similar timeline.

The assessment recommendations will support the development of the next assessment RFP. The RFP and contracting process will occur during the first half of 2025 with hopes of awarding the next assessment contract by late summer 2025. Transitioning from one assessment system is a detailed endeavor that takes time to do well. MSDE plans to operationalize the next assessment system for the 2026-2027 school year but will examine prudent ways to accomplish this on a faster timeline.

The state assessment system and, to a lesser extent, the state accountability system are regularly reviewed by the MSDE Technical Advisory Committee (TAC). The Task Force recognized this critical function but recommended regularly convening a policy- and practitioner-oriented advisory committee to provide feedback on the implementation of these two systems. Further, MSDE, its technical advisors, and this type of policy/practice advisory committee should support a continuous improvement process to ensure that the accountability system meets the changing needs of the State of Maryland and its educational system.

Maryland State Department of Education
Assessment and Accountability Task Force

Maryland
STATE DEPARTMENT OF EDUCATION