Bellwether

# Multiple Choices

*Weighing Updates to State Summative Assessments*

By Michelle Croft, Bonnie O'Keefe, Marisa Mission, and Juliet Squire

JUNE 2024

# CONTENTS


**3** GLOSSARY OF TERMS

**4** INTRODUCTION

**7** BACKGROUND AND ASSUMPTIONS

**9** POLICY GOAL 1: REDUCE TESTING FOOTPRINT

- Reducing Test Length
- Matrix Sampling of Items
- Sampling of Students
- Grade-Band Testing

**16** POLICY GOAL 2: IMPROVE INSTRUCTIONAL RELEVANCE

- Performance Assessment
- Through-Year Assessment

**21** OTHER AREAS OF INTEREST FOR STATES

**23** BARRIERS TO CHANGE

**24** POLICY RECOMMENDATIONS

**27** CONCLUSION

**28** APPENDICES

- A. ESSA and Peer Review Guidance Requirements
- B. Sample of Statewide Student Assessment Testing Times

**30** ENDNOTES

**36** ACKNOWLEDGMENTS

ABOUT BELLWETHER

ABOUT THE AUTHORS

# Glossary of Terms

**Comparability**: Ability to compare scores across test forms and/or across time.

**Computer adaptive testing**: A system in which test items vary in difficulty or in content based on the student's responses to earlier items, instead of all students receiving the same test items.

**Growth measure**: A means of answering questions about the student's academic progress by using a student's assessment data over different time periods.[1]

**Items**: The questions or tasks the students are asked to answer.[2]

**Performance levels**: A label that classifies a test taker's competency or proficiency in a particular content area. Often, states will use classifications such as "Below Basic," "Proficient," and "Advanced."[3]

**Reliability**: The consistency of scores.[4] For instance, a test has high reliability if a student were to test again and receive roughly the same score.

**Sampling**: A process for selecting a smaller number of students to represent a larger student population.[5]

**Subscores**: Scores representing more specific information about a student's knowledge, skills, and abilities in a particular area.[6] For instance, "Integration of Knowledge and Ideas" may be a subscore within an English language arts (ELA) test.

**Testing window**: The dates during which districts must administer an assessment.

**Validity**: The degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.[7]

## MORE FROM BELLWETHER

For more on these terms, refer to Bellwether's Demystifying Statewide Standardized Assessments series.[8]

# Introduction

Since the passage of No Child Left Behind (NCLB) in 2002, each year across the country students in grades 3-8 and high school take their state's standardized tests in math, reading, and, in certain grades, science.[9] These tests, also known as statewide summative assessments, are designed to measure what a student has learned relative to their state's grade-level standards.[10] The scores are used in different, and sometimes overlapping, ways by different groups of people.

For **policymakers**, the scores provide information about how well schools and districts serve students across a state so that the state can better direct resources and intervention efforts toward schools with the greatest need. For **educators**, the scores identify strengths or gaps in learning and inform adjustments to instructional programs moving forward. For **families**, the scores provide a useful indicator of their child's learning, serving as an objective check against grades and report cards,[11] as well as understanding their school's performance.

Despite these use cases, there is growing concern about whether the value of data from statewide summative assessments outweighs the disadvantages, as seen in recent examples of assessment reform efforts overpromising and underdelivering. For instance, tests promising more instructional or parental relevance have regularly delivered scores far too late in a given school year to be useful.[12] Tests promising less administrative burden have been cumbersome and confusing in the early years of implementation.[13] And tests promising to spur improvement in student outcomes have done so indirectly or not at all.[14]

Although conversations about changing summative assessments have been ongoing for years, now is an important time for policymakers and advocates to think through the goals and trade-offs associated with annual summative assessments (see Factors Contributing to the Call for Change).

Today, two broad goals for assessment reform have emerged among policymakers, advocates, families, and educators as the most important: 1) reducing the footprint of summative assessments and 2) increasing summative assessments' instructional relevance.

- Advocates of **reducing the testing footprint** argue that because the summative assessments provide only high-level information, overall testing should be reduced — either by reducing testing time or by reducing other testing-related activities such as test preparation or educator administrative tasks. Some advocates' interest in reducing the footprint of summative assessments is grounded in a more fundamental desire to deemphasize the use of standardized test scores in schools.

- Advocates of **increasing instructional relevance** are interested in redesigning assessments to better support instruction. By structuring tests as opportunities for intervention or progress checks, teachers and families can access information before an end-of-year look back. Some proponents also hope that more instructionally relevant state tests could encourage more effective curricular choices and supplant the interim assessments districts often choose to purchase.

Many stakeholders would like to achieve both these goals; however, there are tensions that may make success on both counts unlikely. But several possible assessment models could advance each goal. Some are federally allowable for states now; some would require changes in federal law to be fully feasible.

# *Factors Contributing to the Call for Change*

Momentum is building in many individual states and on a national level to rethink state summative assessments. This momentum is influenced by several overlapping factors.

## The Elementary and Secondary Education Act (ESEA)

Reauthorization of ESEA, overdue as of school year (SY) 2020-21, has prompted advocates and education leaders to think about what might come next,[15] although it is unclear when Congress might take action.[16]

## Erosion of the Bipartisan Coalition

Reauthorization may be challenging due to the erosion of the bipartisan coalition at the state and federal levels in favor of mandated statewide summative assessments.[17] This has manifested in state lawmakers across a variety of political contexts exploring ways to reduce time and emphasis on standardized tests.[18]

- Some Democrats think standardized tests do not accurately represent students' aptitude and thus may perpetuate and deepen inequities, and believe tests put unfair pressures on overburdened teachers, students, and school systems.[19]

- Some Republicans are less enthusiastic about accountability than their predecessors[20] or have critiqued mandated testing as federal overreach.[21]

## The Innovative Assessment Demonstration Authority (IADA)

A current federal mechanism for innovation — IADA — has produced disappointing results.[22] So far, of the seven state participants pursuing ideas such as performance assessment and through-year assessment, only North Carolina has transitioned to permanent statewide implementation in the timeline required by IADA, and the changes remain optional for K-12 schools.[23] In a November 2023 letter, the U.S. secretary of education announced changes to IADA meant to make it more attractive and feasible for states.[24]

## COVID-19 Pandemic

In the wake of the COVID-19 pandemic, education leaders at the federal, state, and local levels want to understand learning loss and support accelerated learning with accurate data. At the same time, disruptions in accountability and assessments due to students not testing during SY19-20[25] and incomplete testing in SY20-21[26] prompted many states to rethink their assessment systems[27] and consider different approaches to calculating growth.[28]

## Curricular Reform and Supplemental Learning Efforts by States

There are parallel efforts by states to support stronger educational outcomes via higher-quality, research-backed curricula and new approaches to maximize and supplement learning time (e.g., tutoring).[29] These trends fuel a desire for more timely and sophisticated measures of student learning to concretely support and inform instructional practice.

Despite the limitations of current summative assessments, the scores serve as an important foundation of accountability systems, state goal setting, and public transparency. Advocates and policymakers need to understand what they may gain and what they may lose, as well as the policy changes necessary when considering potential shifts in assessments.

This report — informed by literature reviews, previous work on assessments,[30] experience as state assessment staff, and conversations with assessment experts across the country — is designed to help advocates who support educational equity and policymakers understand the different models and their associated trade-offs, the potential impact on historically marginalized students (e.g., students of color, students with disabilities, English learners (ELs), and economically disadvantaged students), and necessary changes to law or U.S. Department of Education guidance (Sidebar 1). It concludes with a set of federal and state policy and advocacy recommendations to enable change so that summative assessments can better address the needs of policymakers, educators, families, and students.

Although conversations about changing summative assessments have been ongoing for years, now is an important time for policymakers and advocates to think through the goals and trade-offs associated with annual summative assessments.

SIDEBAR 1

**What Is Peer Review?**

The federal Every Student Succeeds Act (ESSA) requires that state assessment programs undergo a peer review where technical experts identified by the U.S. Department of Education review and give feedback on the state assessment system's technical quality, including alignment to state standards, overall design and validity, inclusion of all students, and reporting capabilities (Appendix A).

The criteria set out in the peer review guidance build upon ESSA requirements and are consistent with relevant sections of the professional standards for assessments, *Standards for Educational and Psychological Testing*.[31] In some cases, new assessment models would require changes in peer review guidance, but not necessarily law.

# Background and Assumptions

This report's comparison of alternative summative assessment models makes two assumptions about how the test scores will continue to be used:

1. **Produce valid, reliable, and comparable scores** to track state-, district-, and school-level performance and identify schools with low performance.

2. **Produce scores that can be disaggregated by student groups**, including by socioeconomic status, race and ethnicity, and special education and EL status.

These assumptions are made because these are the minimum requirements to identify low-performing schools and have transparency around the performance of historically marginalized student groups (e.g., students of color, students with disabilities, ELs, and economically disadvantaged students). ESSA's requirement that states disaggregate and publicly report data for these student groups[32] has been critical in highlighting equity gaps in student learning and for the school improvement process.[33]

The following two sections analyze two policy goals for changing state summative assessments along with assessment models that could address each goal (Table). Each model includes: 1) an overview of its key features, 2) why states and stakeholders may be interested, 3) what may be gained or lost in making a change, 4) what is unknown, and 5) what changes may be needed in ESSA or peer review guidance to facilitate the model's use.

**TABLE: POLICY GOALS AND POTENTIAL ASSESSMENT MODELS TO MEET THE GOALS**

| Policy Goal | Assessment Model Approaches |
|---|---|
| **1. Reduce Testing Footprint** | 1. Reducing Test Length<br>2. Matrix Sampling of Items<br>3. Sampling of Students<br>4. Grade-Band Testing |
| **2. Improve Instructional Relevance** | 1. Performance Assessment<br>2. Through-Year Assessment |

# *The State Summative Assessment Status Quo*

To assess potential changes, it is useful to understand what an "average" state summative assessment looks like today. This is particularly helpful as many states have substantially changed their tests in the last 10 years, in some cases multiple times.

### Grade Levels and Content Areas

ESSA requires that schools test all students in grades 3-8 and high school for ELA and mathematics as well as in elementary, middle school, and high school for science.[34]

### Administration Mode

All states except Tennessee administer the assessments on a computer, with paper and pencil allowed as an option in limited circumstances.[35]

### Length

Total testing times typically range from 4 to 8 hours for younger students, with slightly longer tests in later grades (Appendix B). Tests are often administered in two to four shorter sessions over a few days. Many are untimed, with a wide range of actual testing time for individual students.

### Time of Year

Each school year, testing typically takes place between March and May.[36]

### Reporting

ESSA requires that the assessments produce individual student scores that can be aggregated to the school, local educational agency, and state and disaggregated by subgroup.[37] Schools must provide families with their students' scores and provide them in an accessible manner (e.g., Braille or native language) when needed.[38]

How quickly families and schools receive the scores varies by state. Although some states have the technical capabilities to produce scores quickly, often states do not report scores until the summer or fall after a spring test. For instance, the Florida Assessment of Student Thinking system provides scores within 24 hours of testing in a parent portal,[39] whereas in Washington State, schools receive a student's test score within a few weeks of the student testing, but families may not receive a score report until September.[40]

### Cost to States

On average, $24-$25 per student for ELA and math.[41]

### How the Scores Are Used

In addition to using scores for federal and state school accountability and public reporting, states use summative scores in a variety of other ways: Some use scores as a high school graduation requirement;[42] some for promotion to the next grade;[43] some for teacher evaluation;[44] and some for charter school (re)authorization.[45]

Districts and schools also use the scores in various ways. For instance, districts may use the scores to help inform higher-level instructional decisions, like identifying gaps in a district's curriculum[46] or, in conjunction with other data, for purposes such as informing a student's course placement or need for remediation.[47]

# Policy Goal 1: Reduce Testing Footprint

Among some stakeholders (e.g., families, students, teachers, school leaders, and advocates), there is a concern that state tests take up a disproportionate amount of learning time for the value the information delivers. Although total testing time for summative assessments usually takes up only a day or two of the 180 days in a typical school year for an individual student, several factors can make testing seem burdensome and disruptive:

- Long windows for testing and associated administrative tasks, which might disrupt normal schedules and school activities.[48]
- Conflation of state tests with additional tests chosen by a district.[49]
- Longer tests in the past.[50]
- Perception of time spent on test prep, discussions about testing, or ways in which testing may shape the content taught throughout the year.[51]
- Stress on teachers, students, and school leaders associated with the tests and their uses.[52]

For some stakeholders, however, the goal is not just to cut down the time and effort surrounding testing but to reduce the emphasis on standardized testing in general and to limit the use of tests for decision-making around school accountability ratings, student promotion, and teacher evaluation, among others. Some of this sentiment comes from a concern that the systemwide emphasis on test scores has distorting or deleterious effects on student learning, as well as student and teacher well-being.[53] State tests with a reduced footprint that fulfill the assumptions of tracking school-level performance and producing data for student groups would mostly function as a check on the system, with less emphasis on individual student or teacher performance.

**For Policy Goal 1, four approaches can help reduce a given state's summative assessment footprint — each with its own opportunities, challenges, and risks.**

1. Reducing Test Length
2. Matrix Sampling of Items
3. Sampling of Students
4. Grade-Band Testing

APPROACH 1

## Reducing Test Length

Reducing the overall test length involves reducing the number of items on the test while maintaining the core goals of testing.

States and stakeholders are interested in this option, as it continues to serve as a check on the system and provides transparency but with less instructional time devoted to testing. However, most states already aim to have the shortest test possible that will produce the data they need to comply with federal and state law.

### *Gains*

**Testing Time**: This may moderately reduce testing time, potentially freeing up more time for instruction.[54] The amount of time reduced will depend on the type and quality of information the state wants to receive. For instance, in the early 2000s, the administration time for the Iowa Test of Basic Skills survey battery — a shortened version of the full assessment producing reading, language, and math scores without subscores

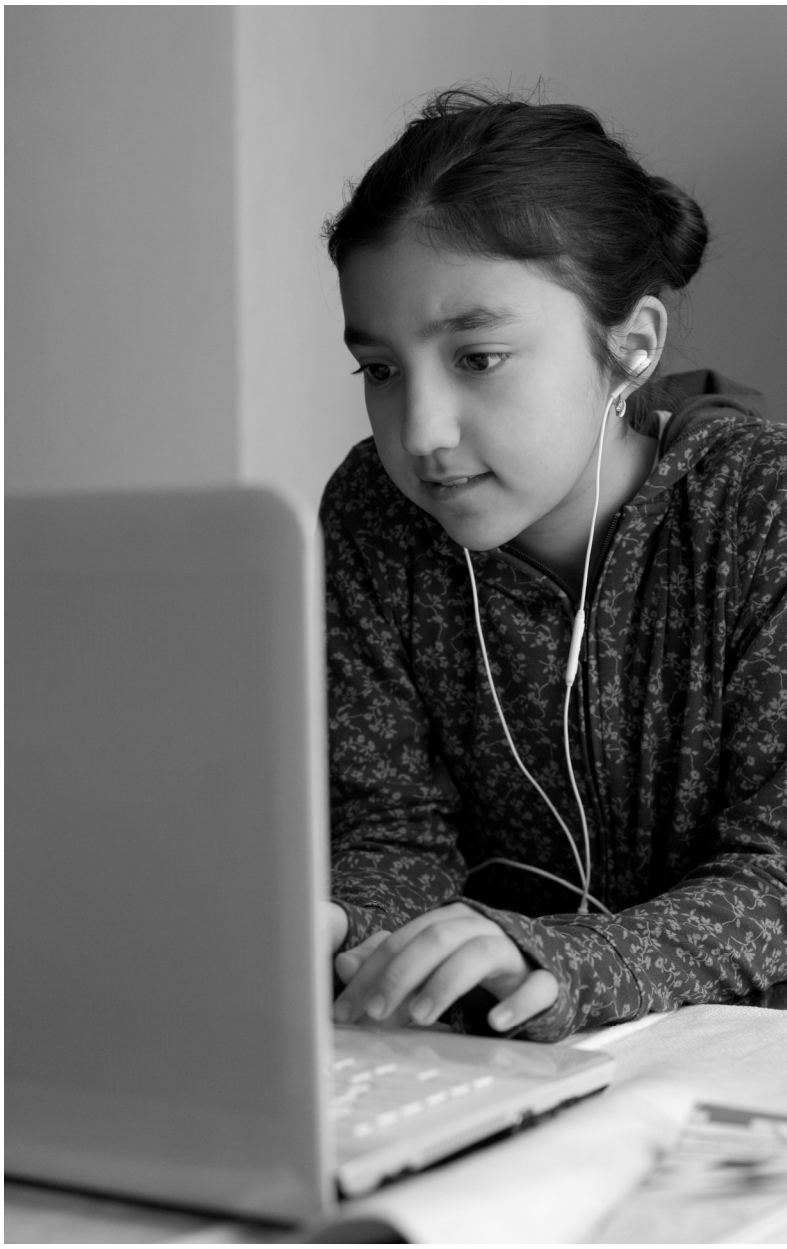
— was 90 minutes.[55] During the NCLB era, more than 85% of state-developed tests took less than 3 hours to complete, with nearly 30% taking less than an hour and a half.[56] These early 2000s tests may not have included any complex writing tasks and had time limits such that students would need an accommodation to receive extended time.[57] Under ESSA, in alignment with revised state standards, many states introduced more complex (and time-intensive) item types to assess skills at a deeper level and adopted untimed assessments with lengthier test windows.

**Administrative**: This may give districts more schedule flexibility.[58] For instance, if test session lengths were short enough to fit within a class period, it might be less disruptive to the school day.

### *Losses*

**Precision**: With fewer items, the scores will be less reliable.[59] This means all those who use the scores will be less sure whether the score accurately captures performance against state standards. Reduced reliability also limits the number of performance levels a state can support (e.g., Advanced, Proficient, or Below Basic), as there may be too much inconsistency to support performance levels beyond the three required by ESSA.[60] Less precision may also make it more difficult to accurately measure the knowledge and skills of high- and low-performing students because there are fewer opportunities to ask questions of greater or lesser difficulty. Computer adaptive testing could help improve some precision for high- and low-achieving students[61] but would not fully address the challenge given that the test would need to collect a certain amount of information about a student's performance before showing the student more challenging or easier items.

**Reporting**: A shortened test likely cannot support subscores, such as a writing subscore on an ELA test,[62] and, as noted earlier, may result in fewer performance levels.



**Equity**: Less information on specific strands of standards might mask inequities and achievement gaps.[63] For example, states that eliminate writing sections of ELA exams to reduce test time may not yield data to reveal a systemic challenge in writing instruction for a particular subgroup of students.

### *Risks and Unknowns*

**Administrative**: It may be that the greater time burdens of the summative assessment are the administrative tasks associated with testing. The demands of a secure and consistent test experience change the normal operations of a school day throughout the test period. This would not be fully addressed by slightly shortening the tests.

**Testing Time**: Most states already work to make tests as short as possible to cover the content required, and there simply may not be much more room to shave down testing time without substantial cuts in content or student performance data.

**Emphasis on Different Standards**: If the test is shortened by omitting particular standards, such as writing or critical thinking, it could encourage teachers to omit or deemphasize parts of the standards and curricula that are not tested. This is already a concern in current summative assessments, with speaking and listening standards typically not being assessed and potentially being underemphasized as a result.[64]

### *Changes Needed in ESSA or Peer Review Guidance*

Given that assessments are already brief in relation to the content covered, substantially reducing the test length may require multiple changes to ESSA as part of a reauthorized ESEA.

First, federal law might need to loosen the required alignment between the tests and state standards. ESSA does not require that a state test all the knowledge and skills covered within the standards each year, but over multiple years the assessment should cover all of the knowledge and skills represented in the standards.[65] ESSA also allows states to test ELA *or* reading, such that states could omit standards associated with ELA to reduce test length.[66] Even with those provisions, the next ESEA reauthorization may need to further relax the requirement to cover the "breadth and depth" of the state's content standards, as it would be difficult for a shorter assessment to fully cover a state's standards without narrowing the standards themselves.

Given that assessments are already brief in relation to the content covered, substantially reducing the test length may require multiple changes to ESSA as part of a reauthorized ESEA.

A second change is to the reporting structure. By reducing the number of required performance levels from three to two (proficient or not), a reauthorized ESEA could allow for a smaller number of items.

A third change is to ESSA's diagnostic requirement, which requires that scores allow "parents and educators to understand specific academic needs of students."[67] A shortened test would not provide sufficient content to understand specific student needs, so this requirement would need to be eliminated.[68]

Finally, the peer review guidance requires the assessment to be "adequately precise" across a full range of performance, including for "high- and low-achieving students."[69] Because the focus of a shortened test would be to identify whether students are proficient or not proficient according to the grade-level standards, there would not be room for much information on either extreme of the score scale. This requirement may, therefore, need to be eliminated.

APPROACH 2

## Matrix Sampling of Items

Matrix sampling of items is a model to "efficiently test a large amount of content and skills by distributing only a small subset of test questions to each student."[70] This means that individual students might not be tested on the entirety of the grade-level standards, but collective scores could roll up to the school level to cover them.

There are different ways to design a matrix-sampled test. In one design, all items are randomly assigned to students. For example, the test developer could divide a test into 10 parts and randomly assign students to one of the 10 parts. Partial matrix sampling is another design where every student is assessed on a minimum core of the standards, and deeper strands are randomized; for example, half of the items are the same for all students, and the other half of the items are randomly assigned (i.e., matrix sampled).

States and other stakeholders are interested in matrix sampling of items because it could allow for substantially shorter testing time for individual students while the schools still receive a similar level of information at the classroom or grade level to support higher-level instructional planning.[71] This model would be similar in some respects to the National Assessment of Educational Progress (NAEP), which uses a matrix-sampled design,[72] or might be perceived similarly to computer adaptive testing, which many states already use, where students receive different questions based on earlier responses to pinpoint performance.[73]

### *Gains*

**Testing Time**: Matrix sampling of items may substantially reduce testing time, as students would be tested on only a portion of the content.[74] For example, NAEP's total administration time, including the time for students to leave class and set up for testing, takes no longer than 90 minutes.[75] Testing times may be longer for partial matrix sampling to ensure there is enough content for students to receive a score.

**Reporting**: Compared with other alternative models to reduce the testing footprint, there is a potential for a robust measure of proficiency at the school level because the breadth of content is still covered across all of the students testing.

### *Losses*

**Individual Student Score**: If the state used a full matrix sampling approach, students, educators, and families would *not* receive a score for individual students. Instead, the assessment would produce only a school score that could be used for school accountability and high-level instructional decisions. This could be problematic for states that use the tests for promotion and graduation decisions.[76] And it would be especially problematic for families, and for student motivation, if they receive no individual information or feedback in return for the time and effort spent testing. We should note that, for some, the lack of a student score could be a benefit by reducing the pressure and emphasis on testing.

A partial matrix sampling *could* allow for individual student scores.

**Precision**: With a partial matrix sampling approach, the reliability of individual student scores and scores for smaller student groups or classrooms will be substantially lower.[77]

**Instructional Use**: A complaint about the current testing system is that scores are not sufficiently useful to inform school and educator planning. Switching to a matrix sample may contribute to less instructional usage than current tests if individual reports are less reliable and less detailed or absent altogether.

### *Risks and Unknowns*

**Reporting**: If a state uses a partial matrix design, the state may report only the common items to be able to have a comparison of student performance. The disadvantage of this approach is that individual students would not receive any feedback from a substantial portion of the test they took. It also would not support subscores at the student level.

### *Changes Needed in ESSA or Peer Review Guidance*

The changes needed in the next ESEA reauthorization will depend on the type of matrix sampling design.

If the state uses a full matrix sampling design, the law must be updated to remove any mention of individual student scores, including item analysis, diagnostic information, and individual score reporting to families. If the state uses the partial matrix sampling design with the student score based on the common items, it could be treated similarly to the reduced test length model, where a reauthorized ESEA may require changes in the areas of alignment, performance levels, diagnostic information, and precision across the full performance continuum.

If the state opted to use a partial matrix design with the score based on all items, there would likely be concerns about comparability because students would not have tested on the same standards.[78] Language related to comparability in a reauthorized ESEA and peer review would need to be adjusted to require comparable school scores but not student scores.

If the state uses a full matrix sampling design, the law must be updated to remove any mention of individual student scores, including item analysis, diagnostic information, and individual score reporting to families.

APPROACH 3

## Sampling of Students

Sampling of students is where students are randomly selected to take the assessment. The number of students selected will vary depending on the size of the grade, school, and student groups.[79] Typically, this model would be adopted in tandem with matrix sampling of items, but the impact of student sampling would be different in several ways.

States and stakeholders are interested in this model to reduce the number of students required to test while still having a measure of school performance. The model is familiar to the way that NAEP samples students.[80]

### *Gains*

**Testing Time**: Testing is reduced or eliminated for some students while continuing to generate an indicator of school performance.

### *Losses*

**Individual Student Score**: Students, educators, and families likely would not receive a score for individual students. Instead, like the fully matrix-sampled test, the assessment would produce only a school score that could be used for school accountability and high-level instructional decisions.

**Administrative**: Schools would likely continue to suspend instruction during test time so that the students selected for testing did not miss key content. Because of this, instructional time may be equally lost as if all students were testing.

**Precision**: The fewer students who test, the less reliable the school score will be.[81]

### *Risks and Unknowns*

**Equity**: To have a reliable measure, there may be a heavier testing burden on smaller schools and smaller student groups, requiring greater proportions of students in those groups to test. As part of the process of identifying schools for targeted support under ESSA, the state sets certain business rules, including the minimum number of students needed for a subscore to be reported or an achievement gap to be identified. These minimums range from 10 to 30 students.[82] To meet these minimums, small schools and smaller student groups may need to test nearly all students in each grade to indicate how each grade level is performing. The differential testing burden may result in pushback from families and teachers about who is selected to test. This model may also place additional pressure on the students selected for testing.

### *Changes Needed in ESSA or Peer Review Guidance*

Sampling of students is not currently allowed and would require multiple changes within a reauthorized ESEA. The most important change would be removing the requirement that all students test and replacing it with language allowing states to use a stratified random sample for testing,[83] if schools can still publicly report subgroup data.

The law would also need to be updated to remove any mention of individual student scores, including item analysis, diagnostic information, and individual score reporting to families.

APPROACH 4

## Grade-Band Testing

This model would reduce the grades in which tests are given. Some examples include the following:

- Testing only one subject per grade (e.g., ELA in grade 3 and math in grade 4).
- Testing every other year.
- Testing once in elementary, middle, and high school (e.g., ELA and math in grades 5, 8, and 10).

Interviewees noted that states may not be considering grade-band testing as an option because it is not allowed in ESSA. But it is currently in practice in various ways, including science and high school testing as part of ESSA,[84] state social studies assessments, the structure of NAEP,[85] and it was the norm in states before NCLB.[86]

### *Gains*

**Testing Time**: Students would spend less time testing during their K-12 experience.

### *Losses*

**Annual Progress Measures**: Schools would no longer have a yearly measure of performance in each grade and subject area. For example, schools would not have a yearly measure of grade 3 and grade 4 math.

**Student Growth**: Not having annual testing would make growth calculations more difficult, particularly where high student mobility results in missing test scores. Year-over-year growth is a key part of many states' accountability systems,[87] in addition to proficiency, and is less associated with individual student characteristics.[88]

**Equity**: As noted earlier, the state sets the minimum number of students needed for a subscore to be reported or an achievement gap to be identified. Fewer students testing increases the potential for missing data, especially for smaller groups of students, as there may not be enough students to publicly report results.[89] Testing every other year could potentially support determinations like the current system, but states would need to adjust other variables in their accountability systems.[90] Likewise, if a state is unable to include growth measures within the accountability system, it increases the likelihood that a school is designated as low-performing due to student background characteristics, which are more closely related to the test scores (i.e., students from higher-income families tend to have higher test scores), instead of the school's quality.[91]

### *Risks and Unknowns*

**Teacher Assigment**: Testing fewer grade levels could skew teacher assignments. The strongest teachers may be more likely to be assigned to tested classrooms.[92] These staffing dynamics are already problematic for non-tested grades and subjects such as elementary grades K-2.[93]

**Emphasis on Different Standards**: Standards-based reform and associated assessments can be advantageous in helping teachers focus on the standards. For example, in Colorado under NCLB, teachers reported a greater schoolwide emphasis on writing in response to the law.[94] However, the testing of grade-level standards might overly concentrate testing pressure in a few grades and harmfully deemphasize standards in other grades.[95]

### *Changes Needed in ESSA or Peer Review Guidance*

As mentioned, grade-band testing for ELA and math is not currently allowed in ESSA. For it to be allowable, the ESEA reauthorization would need to adjust the grade levels and content areas tested. States would also need to consider how to update proficiency goals and how (if at all) to measure and incorporate growth in how they set goals and identify schools in need of support.

# Policy Goal 2: Improve Instructional Relevance

Many stakeholders want to see greater instructional relevance from state summative assessments. Some educators simply do not find much utility in test data[96] and distrust that the tests accurately measure the state's standards.[97] The use case for individual teachers is limited: By the time the state summative test scores come back (often the following school year), it is too late for teachers to make changes at either the class or individual level. Having information earlier in the year could allow schools and educators to provide remediation.

States also want the earlier information to be high quality and aligned with state standards. Districts often use commercially developed interim assessments, but little public information is available to evaluate quality.[98] A state-developed tool provides additional confidence about the technical quality and interpretations. More importantly, the state can provide resources to support teachers in appropriately using the data.[99]

Another factor driving the interest in instructional relevance may be the emphasis on curricular quality in recent years (Sidebar 2). Historically, curricular decisions have been left to districts. States see the potential positive impact of a high-quality and more consistent curriculum on student outcomes and are trying to incentivize or force districts in that direction.[100] Aligning state tests more closely to vetted curricula could be a powerful lever to shift instructional practice, creating a clear loop of feedback between what is happening in the classroom and what is assessed by the state.

**For Policy Goal 2, two approaches can help reduce a given state's summative assessment footprint — each with its own set of opportunities, challenges, and risks.**

1. Performance Assessment
2. Through-Year Assessment

APPROACH 1

## Performance Assessment

Performance assessments are multistep activities where "students are asked to produce a product or carry out a performance (e.g., a musical performance) that is scored according to prespecified criteria, typically contained in a scoring guide or rubric."[101]

Performance tasks can be relatively brief. For example, the Smarter Balanced assessment includes a performance task for ELA and literacy involving targeted research and writing; for math, it allows students to demonstrate problem-solving, communicating reasoning, and modeling and data analysis.[102] Other states often include open-ended writing tasks.

The performance-based assessments discussed here are composed solely of performance tasks administered throughout the school year and aligned to clear learning targets.[103]

State policymakers and other stakeholders are interested in performance assessments for several reasons. One is because the assessment can signal a preferred type of instruction. As teachers may look to the state's assessment for an indicator of what instruction should look like, performance assessments may help to signal richer, more interactive classroom

tasks.[104] Another is because performance assessments can measure standards that are difficult to assess through traditional tests, such as speaking, research, or higher-order thinking skills.[105] They also embed teacher professional development within the design and scoring, which may increase general assessment literacy and use of the data.[106]

### *Gains*

**Higher-Order Tasks**: Performance assessments have the potential to measure more cognitively complex skills better than other assessments.

**Integration of Content Areas**: Performance assessments have the potential to measure performance in more than one subject area instead of the current model of separating content areas.[107] For example, math could be integrated into a science lab task.

**Educator Interest**: Teachers might see greater value in the assessment and be more likely to use the data.[108]

### *Losses*

**Administration**: If testing every student at every grade, then the logistical burden and cost is high, particularly given the time for teacher training to be able to administer the assessments.[109]

**Testing Time**: Performance assessments can lead to greater testing time, even if testing just one grade level.[110]

**Precision**: Performance assessments can lead to lower reliability for student scores because of the complexity of scoring unless additional multiple-choice items are added to create a hybrid test.[111]

**Reporting**: Performance assessments can lead to longer score reporting times due to the complexity of the test and scoring where student performance is measured against a rubric,[112] although various researchers and vendors are seeking to solve this challenge through technology-driven scoring methods.[113]

SIDEBAR 2

**Tension in Curriculum Alignment May Limit the Adoption of Instructionally Relevant Assessments**

In order for tests to be relevant to everyday instruction, there should be close alignment between a curriculum and an assessment so that students are tested on material they have recently been taught. This implies that instructionally relevant assessments depend on states having greater control and consistency in curriculum. For example, a state test might ask students questions about a book they all had read and discussed recently in class, rather than an excerpt of an unfamiliar text.

There are potential drawbacks to closely aligning state tests and classroom curricula. Tying assessments to specific curricula in this way may not adequately measure a student's ability to transfer learning in new situations (i.e., ability to comprehend a new text), such that states may need to include additional items based on unfamiliar texts, and might constrain the ability of schools to experiment with more innovative curricula or instructional models.

Additionally, if a state ties an assessment to specific curricula but otherwise does not require districts to use the curricula or provide the resources for purchase and implementation, it may disadvantage districts that do not have access to high-quality curricula, coaching, and implementation assistance (by choice or by lack of resources). Given districts' historically high level of control around curricular choices, state intervention into curriculum via testing may generate strong political pushback.

**Equity**: There are two equity concerns with performance assessment, both related to teacher involvement within the assessments. Given the considerable amount of training required to properly administer the assessments, there is a concern that schools with high proportions of economically disadvantaged students, which are more likely to have less-experienced teachers, will perform poorly because of the lack of teacher experience needed to administer and prepare students for the assessments.[114] Another concern is consistency in accommodations for students with disabilities.[115]

### *Risks and Unknowns*

**Validity**: If the assessments are not properly administered or scored,[116] there is a greater chance to affect the validity of the score.[117] This means there is less evidence to support the score as an accurate measure of the student's knowledge and skills.

**Local Control**: Performance assessments are heavily curriculum-based, which could affect implementation and scores if the timing of the assessment is out of step with when information is taught.[118]

### *Changes Needed in ESSA or Peer Review Guidance*

Performance assessments are allowable under ESSA but are difficult and expensive to implement while also meeting the technical requirements required for accountability, including reliability, validity, and alignment with the standards.

The changes needed in ESSA or peer review to facilitate the use of performance assessments would likely be extensive and compromise on the minimum assessment principles (i.e., a valid, reliable, and comparable score able to be disaggregated by student group).

APPROACH 2

## Through-Year Assessment

Through-year assessments replace one end-of-year assessment with multiple test events throughout the year. The most common design is three test events a school year, but at least one state (Montana) is exploring a design with as many as five testing windows. Each testing window includes multiple brief testing events designed to measure a smaller number of standards per test. Most through-year assessment designs aspire to test all standards at least once at points throughout the year, eliminating the need for a summative test covering all standards at the end of the year.[119]

Although the original concept of through-year assessment was to use the multiple test events to produce a single summative score, to date, most states are using only the final test in the series to produce a proficiency score.[120] Given the general shift in systems, these are often still considered a through-year assessment, as the scores from throughout the year *could* contribute to a single summative score.[121]

State policymakers and other stakeholders are interested in through-year assessments for a few reasons.[122] The first is the opportunity to provide more frequent feedback to students, families, and educators. Data earlier in the year may allow for earlier interventions. Second, state policymakers are interested in reducing the overall testing by supplanting district interim assessments. The state interims may offer closer alignment to state standards than commercially available or district-created interims.

Finally, state policymakers are interested in through-year assessments' ability to produce different kinds of growth indicators. Because of student mobility or the lack of an assessment in the prior grade, states may be unable to calculate a year-to-year growth score for all students. By having an assessment at the beginning of the year, state policymakers can create growth scores for more students within the state and better approximate what a student has learned during the school year.

Research on the implementation of through-year assessment is still in the early stages. Because the information is limited and there are a variety of potentially important design choices under the through-year banner, this model has more uncertainties.[123]

### *Gains*

**Opportunity for Student Remediation**: There are anecdotal reports that some teachers and schools are providing remediation based on earlier test scores, but more research is needed to understand how widespread the practice is as well as what interventions are being used and their effectiveness.[124]

**Opportunity for State Intervention**: A uniform measure of progress throughout the year may allow states to target resources and interventions to struggling schools or districts during the year instead of the following year.

**Within-Year Growth**: Some contend that the through-year model could produce useful new measures of growth by seeing how much students know at the beginning of the year compared with the end. Others contend that some through-year models measure a student's opportunity to learn the content because students have not always been taught the material before testing. This creates a challenge in interpreting the growth measures.[125]

**Curriculum Choices and Educator Planning**: Many through-year models require teachers to be thoughtful about when the standards are taught and tested throughout the school year. This could promote the adoption of higher-quality curricula and greater educator planning throughout the year to ensure all standards are covered.

### *Losses*

**Administration**: More testing times per year requires more administrative resources at both the state and the local level.

**Testing Time**: In a through-year assessment model, overall testing time would likely increase, particularly if the state allows for retesting of previously assessed standards or if districts continue to administer separate interim assessments.[126]

**Comparability**: To date, there is no agreed-upon standard for rolling up the test events into a single, comparable score.[127] Comparability is a challenge because of the timing of proficiency determinations. For instance, if a student reaches proficiency in the fall, would that score have the same meaning as that of a student who reaches proficiency in the spring? Similarly, if a student retests on the standards throughout the year, would the student's score have the same meaning as that of a student who does not retest?

**Equity**: Increased interruption to the school day with additional testing events can be disruptive, particularly to students with disabilities.[128] Another challenge is student mobility. Highly mobile students moving among districts or states would be unlikely to have the same test experience or opportunity for instruction as their peers.

### *Risks and Unknowns*

**Increased Emphasis on Testing**: Through-year assessments are designed to influence instruction more directly, creating a more frequent feedback loop between tests and day-to-day instruction. This is in tension with many stakeholders' desire to reduce the footprint of tests, and could create pushback from those who are skeptical of state tests in general.

**Instructional Use**: Through-year assessments might not provide detailed enough information to inform instruction in the ways that many schools currently use homemade or off-the-shelf interim assessments. Most state through-year assessments are very similar to traditional end-of-year assessments and have similar technical features. For example, cost issues related to test security requirements mean that teachers would have limited visibility into the content of the questions a student got right or wrong, which is important for diagnosing areas of missed learning.[129]

### *Changes Needed in ESSA or Peer Review Guidance*

ESSA explicitly permits states to use multiple assessments throughout the year.[130] However, certain elements, particularly in the peer review guidance, may make compliance more difficult for states.

As noted earlier, the biggest challenge is score comparability if the state tries to roll up the test events into a single score.[131] The peer review guidance would need to update the requirements about multiple test forms to allow for states to show evidence of comparability for *proficiency levels* as opposed to comparability for *particular student scores*.[132]

Test administration and security requirements within peer review would also need to be relaxed for the assessments to be maximally useful to educators. Teachers would likely want more specific information, such as test questions, to help guide instruction. Likewise, traditional test security practices, such as removing instructional content on walls, must be revisited for seamless integration into the classroom.

Research on the implementation of through-year assessment is still in the early stages. Because the information is limited and there are a variety of potentially important design choices under the through-year banner, this model has more uncertainties.

# Other Areas of Interest for States

There are a few other goals and ideas emerging in states that do not fit into the categories of reducing the footprint and increasing instructional relevance. These are less prevalent in current conversations, but they may be useful for advocates to consider.

### Multiple State Assessments

IADA allows states to experiment with multiple tests, but states must eventually scale the innovative pilot to replace the statewide assessment. However, there may be promising innovations that are unlikely to be scalable statewide due to administrative or political barriers. For instance, in Louisiana, the state's IADA through-year assessment is aligned with the Guidebooks 2.0 ELA curriculum, used in approximately 80% of Louisiana districts.[133] Without requiring the remaining 20% of the districts to adopt the curriculum, it would be challenging for the state to require the assessment statewide. One proposal is to allow states, like Louisiana, to offer districts choices among multiple assessment systems, provided that the state finds a way to link or compare the two or more tests. The main barrier to this approach is that it is resource-heavy for states to maintain two testing programs, particularly when the state must provide evidence of comparability between the assessments.

### Off-Grade-Level Items

ESSA requires that states test on-grade-level standards.[134] Some states are interested in including off-grade-level items on an assessment to better understand what an individual student knows and can do above or below grade level.[135] States can technically include off-grade-level items, but those items cannot contribute to the student's test score, making for a longer test. More off-grade-level items could diminish the usefulness of tests as a measure of student or school proficiency against grade level, which is of particular concern for advocates of students with disabilities,[136] but could more precisely pinpoint student performance for some students, thus increasing instructional relevance.

### Competency-Based Assessments for High School

State summative assessments are limited to measuring proficiency in certain academic areas. More than a dozen states have developed a Portrait of a Graduate, which includes other competencies graduates should exhibit, such as creativity, communication, and working collaboratively.[137] These competencies are important skills but are more difficult to measure than reading and math. States interested in creating competency-based assessments based on their Portrait of a Graduate likely would do so outside of ESSA accountability or in addition to the state summative assessment for reading and math.

### Workforce Readiness

States and federal policymakers are increasingly interested in work readiness skills. States are able to include measures of work readiness as an ESSA "other measure of school quality,"[138] but there may be interest in supplementing traditional high school tests or college-readiness tests with workforce-readiness indicators.

### Culturally Relevant Assessments

The widely implemented best practice in state summative assessments is to eliminate any content that "differentially favors individuals from some subgroups over others" through reviews for potential bias and other steps in the test design process.[139] But some researchers and experts argue that stripping out cultural specificity and responsiveness from test content advances a white-centric viewpoint to the detriment of students of color, ELs, and other test-takers.[140]

Research is underway to better understand how cultural relevance affects student experience and test scores, positively or negatively. For example, a recent study examined students' reactions to test items designed to highlight aspects of Black or Hispanic communities and cultures.[141] State interest in making definite steps toward culturally relevant assessment has bubbled up in limited ways so far. For instance, items for the Montana through-year assessment prioritize Montana and Indigenous authors, aligned with the state's "Indian Education for All" law.[142] However, continued research is needed to understand how best to address cultural relevance in state summative assessments.

Some researchers and experts argue that stripping out cultural specificity and responsiveness from test content advances a white-centric viewpoint to the detriment of students of color, ELs, and other test-takers.

# Barriers to Change

As states explore revising their assessment systems, it is important to consider the ecosystem in which they make decisions and the barriers they face when enacting changes.

The first is the **accountability context**, which complicates changes to the state summative assessments. There is a vast body of research literature about how perceived or real accountability consequences change educator leader, teacher, and student behavior around the tests. The changes range from positive behaviors, such as teaching the required standards[143] and identifying students for remediation, to negative behaviors, such as educator cheating[144] or deemphasis on non-testing subjects.[145]

Accountability systems can also influence how educators and the public emotionally respond to the assessments. Distrust, particularly from educators,[146] makes it harder to gather buy-in from the very people who need to use the results to achieve instructional usefulness. Given how accountability changes educator and student behavior, it may be impossible to have one test used for both accountability and day-to-day instructional purposes.

Another barrier is **state capacity** to implement change. State education agencies may not have sufficient resources to develop a new assessment system or guide districts and the public through a big change process, especially amid competing priorities and budgetary pressures. States are also constrained due to the test vendors that develop the assessments on the states' behalf, as the vendors lack the incentive to innovate without a push from state buyers. Finally, change carries risks that some state education agencies may be unwilling to take on. As one expert noted, state assessment departments work hard to build trust in the assessment and scores. If a solution does not deliver on its promises, states can lose traction with stakeholders.

**Politics** also plays a role. In addition to scores being highly politicized, in some cases politicians use assessments to accomplish other policy goals, such as signaling their support for teachers or emphasizing the value of college and career readiness.[147] Politics can also limit a state's ability to make certain changes. For instance, a state often has limited ability to control district curriculum or the timing of when certain standards are taught.

# Policy Recommendations

There is growing energy to make changes in assessments but less clarity on how to move forward. Some continued experimentation and variation among states could be beneficial, to pressure-test newer innovations and continue to allow states to emphasize different goals that match their educational and political context. But, not every potential option is worth the risks, trade-offs, or the implementation effort it would take to achieve.

## *Recommendations for State Policymakers*

For states committed to the goal of **reducing the footprint**, although all paths may be technically possible and achieve this report's baseline assumptions with changes in law, several approaches have more severe drawbacks that make them inadvisable.

| Approach | Recommendation | Rationale |
|---|---|---|
| Reducing Test Length | Not recommended | ***Loss of performance detail***<br><br>By reducing the test length, students and schools may save one to two hours of testing but lose additional information from the tests, such as subscores, which may help inform higher-level instructional decisions. |
| Full Matrix Sampling of Items | Not recommended | ***Loss of individual scores***<br><br>Student-level scores are a key indicator for families, allowing them to see how their student is performing against the state's standards and compared with other students in the state.<br><br>Additionally, sampling students would put a disproportionate and inequitable burden on students in smaller groups or those in smaller schools. |
| Sampling of Students | Not recommended | |
| Grade-Band Testing | Not recommended | ***Loss of information on student groups essential for equity***<br><br>Disaggregating scores by groups is a feature of current tests that civil rights advocates have held up as critically important for decades. During the last ESEA reauthorization, a strong coalition of civil rights groups and advocates for certain student groups (e.g., students with disabilities and Latino students) were adamant about annual public reporting "both overall and for all groups of students so families and taxpayers have honest, consistent information on how their schools are performing."[148] |

| Approach | Recommendation | Rationale |
|---|---|---|
| **Partial Matrix Sampling of Items** | **Recommended, with conditions** | With a partial matrix sample design that reduces the footprint of summative testing, **states could minimize test time while continuing to allow for growth scores, individual student score reports, and subgroup performance transparency**. Attention to sample design and test design would still be essential to preserve equity and ensure usable results combined with significantly less time spent testing. And changes in law would also be necessary.<br><br>A state using a partial matrix sample of items might reduce test time and simultaneously answer the call for instructional relevance and more detailed information outside the summative assessment. For instance, states can build flexible, open item banks where educators can select items aligned to what is happening in the classroom to help inform day-to-day instruction. This path may enable states to provide comparable student scores and information to guide high-level accountability decisions while informing day-to-day instruction through other means. |

States primarily interested in **instructional relevance** face greater implementation challenges. Despite interest in pursuing this goal, some experts hold that any test tied to accountability will have limited appeal and usability for educators on a day-to-day basis.

| Approach | Recommendation | Rationale |
|---|---|---|
| | | ***Overly burdensome when tied to accountability*** |
| Performance Assessment | Not recommended | Despite educator excitement over performance-based assessment, states and schools should anticipate an expensive and arduous process to yield valid, comparable scores. |
| **Through-Year Assessment** | **Recommended, with conditions** | Many through-year assessment designs so far do not represent a substantial change from the assessment status quo given the need to produce comparable scores that can be used for accountability and allow for variation in curriculum. A better path forward for instructional relevance while meeting this report's baseline assumptions could be **a flexible through-year model where students are tested on recently taught standards**. This approach lets districts maintain local control over curriculum and provides information throughout the year. **It is unclear, however, if this approach would provide the comparable student data** that advocates believe to be critical for identifying schools for additional resources and intervention, and it would not answer the call to reduce the footprint of testing. |

# *Recommendations for Federal Policymakers*

To enable state policymakers to pursue these paths with the best potential impact, federal policy changes will be necessary.

**Change ESEA's assessment requirements to support different types of assessments.**

To enable states to *reduce the summative assessment footprint* via matrix sampling of items, the next reauthorization of ESEA could focus on requirements that summative assessments provide data to inform accountability decisions and high-level district and school instructional decisions. Federal policymakers should continue to require an individual student score but eliminate requirements focused on informing classroom instruction and relax some standards of depth and diagnostic precision. For example, the next ESEA reauthorization should remove language requiring the assessment to provide diagnostic information for students.

To support the path of *increased instructional relevance*, ESEA could loosen score comparability requirements, focusing on proficiency-level comparability instead. By focusing comparability on the proficiency levels, states can allow for greater flexibility regarding when tests are administered throughout the year.

**Provide additional support for assessment-related change.**

Support can come in multiple ways. One is tying the Competitive Grants for State Assessments (CGSA), which supports assessment innovation beyond the state assessments,[149] to the Innovative Assessment Demonstration Authority.

Another is requiring states to develop a system of assessment professional development for educators. Doing so could help educators understand how to appropriately use scores from different types of assessments and create better classroom-based assessments to inform day-to-day instruction.

# *Recommendations for Advocates*

In either path forward, advocates will be essential, to both inform policy change and be watchdogs for state implementation. For advocates interested in educational equity, we recommend the following.

**Figure out what you are not willing to compromise.**

All assessment design changes include trade-offs. Advocates must identify what they are not willing to compromise on (e.g., annual reporting of comparable student group data at the school, district, or state level), and then hold firm with federal and state policymakers for these critical elements. This clarity could also help build stronger, more effective coalitions.

**Ask questions to understand the state's policy goal and how a proposed change to assessments helps to advance that goal.**

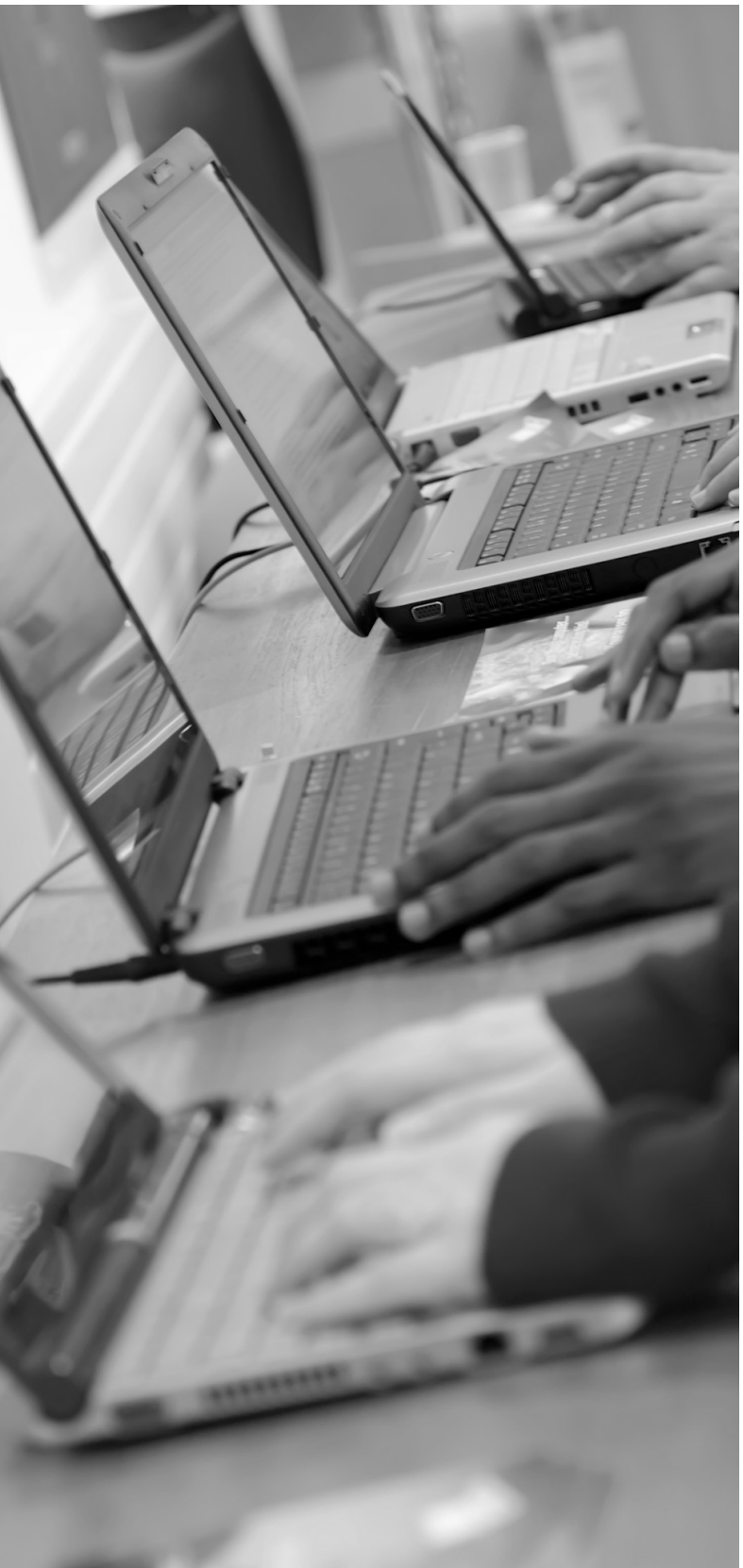
Advocates should ask questions of their state leaders to help clarify the main policy goals and to identify where, if at all, those goals conflict with one another. It may be useful to map out what policymakers, educators, and families, particularly families of historically marginalized students, want in assessments and create a system of assessments that can provide that type of information.

**Take steps to ensure the state supports policy implementation.**

As states implement new assessments, they must provide districts with the resources for implementation. These resources include information and technical assistance to the educators who are administering the assessment and using the data, and information for families. Advocates can help support this process by providing input into the design of resources so that they are accessible and communicate the information community members are most interested in.

Implementation also requires a system for continuous improvement. Advocates should encourage the state to develop a formal monitoring system that includes

monitoring student experiences. States are required to monitor the testing process for test security reasons, but advocates should encourage states to extend the monitoring to include how students experience the assessment, particularly to monitor for unintended consequences.



# Conclusion

State summative assessments are an important tool to provide stakeholders with credible, comparable information about student learning. In an era marked by large declines in student learning due to the pandemic, reliable barometers of student progress are more important than ever to track improvement and shine a light on inequities.

As policymakers and advocates at the state and federal levels think about updates and innovations to assessments, it is essential to understand that even promising changes will come at a cost, which could include data usability, test reliability, implementation effort, or political pushback. Policymakers and advocates should prioritize policy goals, determine which goals are most appropriate for the summative assessment, and design the summative assessment to meet those needs. In making these decisions, state departments of education should involve a broad coalition of stakeholders early in the design process.

Although these discussions are complex and time-intensive, they can result in assessments that are better positioned to meet students' needs in the decades to come. For the goals that the summative assessment is unable to address, the state may develop or otherwise support districts (through funding and/or professional development) to create other tools, assessments, or resources to meet those goals.

To do otherwise in this tenuous political moment risks losing state assessments, and the valuable data they produce, entirely. ✦

# Appendix A

**ESSA AND PEER REVIEW GUIDANCE REQUIREMENTS**

*What does ESSA and assessment peer review guidance require?*

### Which Students Test and in Which Subjects

- The same assessment for all students with limited exceptions.[150]
- All students[151] in grade levels 3-8 and high school in ELA or reading and math; science in certain grade spans.[152]

### Alignment to Standards

- Must be aligned with the state's content standards.[153]
- Must be aligned "to the depth and breadth" of the standards.[154]

### Technical Quality

- Reliable, valid, and consistent with professional standards.[155]
- Adequate technical quality for the purposes of ESSA.[156]
- Technical quality is further defined in peer review to include elements such as the following:[157]
    - Validity of the scores, including that "the scoring and reporting structures of its assessments are consistent with the sub-domain structures of the State's academic content standards."
    - Full performance continuum (i.e., the assessment is "adequately precise," including for "high- and low-achieving students").
    - Multiple forms and versions (e.g., paper and pencil and computer-based).
    - Comparable within and across school years.

### Results

- Coherent and timely information about student attainment of the standards and whether the student is performing at grade level.[158]
- Three achievement levels.[159]
- Individual student interpretative, descriptive, and diagnostic report "that allows parents and educators to understand specific academic needs of students and that are provided to parents … as soon as practicable and in an understandable and uniform format."[160]
- Disaggregated by student groups.[161]
- Itemized score analysis to see specific needs of students.[162]

### Ability to Innovate

- Language in ESSA permits states, in addition to the IADA to innovate through alternatives to selected-response, which may include growth and be partially delivered in the form of portfolios, projects, or extended performance tasks.[163]

***Note**: Any change to ESSA assessment requirements may also affect secondary ESSA provisions, such as setting goals.*

# Appendix B

**SAMPLE OF STATEWIDE STUDENT ASSESSMENT TESTING TIMES**

The below testing times are gathered from administration manuals for SY22-23 and SY23-24. Where states' tests were untimed, the suggested time block for schedulers is listed as a proxy for estimating total test length.

| States | Grade 3 | | | Grade 8 | | |
|---|---|---|---|---|---|---|
| | *ELA* | *Math* | *Total* | *ELA* | *Math* | *Total* |
| Georgia (2023)[164] | 2:05-4:10 | 1:00-2:10 | 3:05-6:20 | 2:05-4:10 | 1:00-2:10 | 3:05-6:20 |
| New York (2023)[165] | *Untimed 2:10-2:30 | *Untimed 1:55-2:15 | 4:05-4:45 | *Untimed 2:50-3:10 | *Untimed 2:35-2:55 | 5:25-6:05 |
| California (2024)[166] | *Untimed 2:45 | *Untimed 1:45 | 4:30 | *Untimed 2:45 | *Untimed 2:00 | 4:45 |
| Arizona (2024)[167] | *Untimed 3:00-4:30, including 30 min Oral Fluency | *Untimed 2:00-2:50 | 5:00-7:20 | *Untimed 2:30-4:00 | *Untimed 2:00-2:50 | 4:30-6:50 |
| Pennsylvania (2023)[168] | 2:55-3:40 | 2:50-3:20 | 5:40-7:00 | 4:10-4:55 | 2:50-3:20 | 7:00-8:15 |
| South Dakota (2024)[169] | *Untimed 3:30 | *Untimed 2:30 | 6:00 | *Untimed 3:30 | *Untimed 3:00 | 6:30 |
| Colorado (2023)[170] | 4:30 | 3:15 | 7:45 | 5:30 | 3:15 | 8:45 |

# Endnotes

1 Data Quality Campaign, *Growth Data: It Matters, and It's Complicated*, January 2019, 1, https://files.eric.ed.gov/fulltext/ED593508.pdf.

2 American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, eds., *Standards for Educational and Psychological Testing* (Lanham, MD: American Educational Research Association, 2014), 220, https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf.

3 AERA et al., *Standards for Educational and Psychological Testing*, 221.

4 AERA et al., *Standards for Educational and Psychological Testing*, 222–223.

5 *APA Dictionary of Psychology*, s.v. "sampling," updated April 19, 2018, https://dictionary.apa.org/sampling.

6 Michelle Croft, Hailly T. N. Korman, and Titilayo Tinubu Ali, *Demystifying Statewide Standardized Assessments: Developing High-Quality Assessments and Items*, Bellwether, April 2023, 5, https://bellwether.org/wp-content/uploads/2023/03/DemystifyingStandardizedAssessments_Brief-3_Bellwether_April2023.pdf.

7 AERA et al., *Standards for Educational and Psychological Testing*, 11.

8 Croft et al., *Demystifying Statewide Standardized Assessments*.

9 No Child Left Behind Act of 2001, 20 U.S.C. § 1111 (b)(3)(A) (2002), https://www.congress.gov/107/plaws/publ110/PLAW-107publ110.pdf; Every Student Succeeds Act, 20 U.S.C. § 1111 (b)(2)(A) (2015), https://www.congress.gov/114/plaws/publ95/PLAW-114publ95.pdf.

10 Croft et al., *Demystifying Statewide Standardized Assessments*, 2.

11 Gallup & Learning Heroes, *B-flation: How Good Grades Can Sideline Parents*, November 2023, 3, https://bealearninghero.org/wp-content/uploads/2023/11/B-flation_Gallup_Learning-Heroes_Report-FINAL.pdf.

12 Chad Aldeman, "Schools, Teachers & Parents Need Rapid State Test Results. Why Are They So Slow?," *The 74*, September 18, 2023, https://www.the74million.org/article/schools-teachers-parents-need-rapid-state-test-results-why-are-they-so-slow/.

13 Michelle Croft, Titilayo Tinubu Ali, and Bonnie O'Keefe, *Testing the Waters: Insights Into Parent Perspectives on Through-Year Assessment Implementation*, 2023, 18, https://bellwether.org/publications/testing-the-waters/.

14 Jonathan Supovitz, "Is High-Stakes Testing Working?," University of Pennsylvania Graduate School of Education, https://www.gse.upenn.edu/review/feature/supovitz; Thomas S. Dee and Brian A. Jacob, "The Impact of No Child Left Behind on Students, Teachers, and Schools," *Brookings Papers on Economic Activity*, no. 2 (Fall 2010): 149, https://www.brookings.edu/wp-content/uploads/2010/09/2010b_bpea_dee.pdf; Thomas S. Dee and Brian A. Jacob, "Evaluating NCLB," *Education Next* 10, no. 3 (2010), https://www.educationnext.org/evaluating-nclb/.

15 The Every Student Succeeds Act (ESSA), the latest reauthorization of the ESEA, was signed into law in December 2015 and technically was due for reauthorization in the 2020-21 school year. Catherine A. Paul, "Elementary and Secondary Education Act of 1965," VCU Libraries Social Welfare History Project, https://socialwelfare.library.vcu.edu/programs/education/elementary-and-secondary-education-act-of-1965/.

16 NCLB was slated for reauthorization in 2007 but wasn't reauthorized until 2015. U.S. Department of Education, *Every Student Succeeds Act* (ESSA), https://www.ed.gov/essa?src=rn.

17 Lynn Olson and Craig Jerald, *The Big Test: The Future of Statewide Standardized Assessments*, FutureEd, April 2020, 7, https://www.future-ed.org/wp-content/uploads/2020/04/TheBigTest_Final-1.pdf.

18 Olson and Jerald, *Big Test*, 1.

19 John Rosales and Tim Walker, "The Racist Beginnings of Standardized Testing," *National Education Association News*, March 20, 2021, https://www.nea.org/nea-today/all-news-articles/racist-beginnings-standardized-testing.

20 Katelyn Cordero and Juan Perez Jr., "'A Bizarre Coalition:' Red and Blue States Weigh Big Changes to Testing Requirements," *Politico*, December 6, 2023, https://www.politico.com/news/2023/12/06/standardized-testing-changes-backlash-00129688.

21 Dana Goldstein, "After 10 Years of Hopes and Setbacks, What Happened to the Common Core?" *New York Times*, updated November 4, 2021, https://www.nytimes.com/2019/12/06/us/common-core.html; Ashley Jochim and Patrick McGuinn, "The Politics of the Common Core Assessments," *Education Next* 16, no. 4 (2016): 44–52, https://www.educationnext.org/the-politics-of-common-core-assessments-parcc-smarter-balanced/.

22 U.S. Secretary of Education Miguel A. Cardona to the Chief State School Officers, November 20, 2023, https://oese.ed.gov/files/2023/11/23-0431-DCL-IADA-os-approved-11.17.2023.pdf; Alyson Klein, "The Feds Gave States the Chance to Create Better Standardized Tests. There Were Few Takers," *Education Week*, April 19, 2023, https://www.edweek.org/teaching-learning/the-feds-gave-states-the-chance-to-create-better-standardized-tests-there-were-few-takers/2023/04.

23 Office of Elementary and Secondary Education, *Innovative Assessment Demonstration Authority (IADA)*, https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/iada/; "NC Check-Ins 2.0," State Tests, North Carolina Department of Public Instruction, https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests/nc-check-ins-20.

24 U.S. Secretary of Education Miguel A. Cardona to the Chief State School Officers, November 20, 2023, Key Policy Letters Signed by the Education Secretary or Deputy Secretary, U.S. Department of Education, https://www2.ed.gov/policy/gen/guid/secletter/231120.html.

25 Catherine Gewertz, "See Which States Have Cancelled Spring Tests Because of Coronavirus," *Education Week*, March 17, 2020, https://www.edweek.org/teaching-learning/see-which-states-have-cancelled-spring-tests-because-of-coronavirus/2020/03.

26 Catherine Gewertz, "State Test Results Are In. Are They Useless?" *Education Week*, October 21, 2021, https://www.edweek.org/teaching-learning/state-test-results-are-in-are-they-useless/2021/10.

27 Cordero and Perez, "Bizarre Coalition." For instance, as an emergency measure, Maine used interim assessments to evaluate student learning and has since developed a new assessment system modeled around the use of interim assessments. Robbie Feinberg, "Maine Students Will Take a New Standardized Test This Spring," *Maine Public*, January 20, 2023, https://www.mainepublic.org/maine/2023-01-20/maine-students-will-take-a-new-standardized-test-this-spring; Naaz Modan, "Maine Rebuts Ed Department Threat to Withhold Some Federal Funds," *K-12 Dive*, updated March 15, 2023, https://www.k12dive.com/news/Maine-Title-I-funds-jeopardy-over-assessment-data/643757/. Similarly, Smarter Balanced introduced a shorter form that has continued beyond the pandemic: Smarter Balanced Assessment Consortium, *2020-21 Summative Technical Report*, October 14, 2022, chap. 6, https://technicalreports.smarterbalanced.org/2020-21_summative-report/_book/test-admin.html#test-duration.

28 Office of Governor Ron DeSantis, "Governor DeSantis Announces End of the High-Stakes FSA Testing to Become the First State in the Nation to Fully Transition to Progress Monitoring," September 14, 2021, https://www.flgov.com/2021/09/14/governor-desantis-announces-end-of-the-high-stakes-fsa-testing-to-become-the-first-state-in-the-nation-to-fully-transition-to-progress-monitoring/; Data Quality Campaign, Alliance for Excellent Education, and Collaborative for Student Success, *Measuring Growth in 2021: What State Leaders Need to Know*, August 2020, 3, https://dataqualitycampaign.org/wp-content/uploads/2020/07/Measuring-Growth-in-2021_What-State-Leaders-Need-to-Know.pdf.

29 Sy Doan et al., *How States Are Creating Conditions for Use of High-Quality Instructional Materials in K–12 Classrooms: Findings From the 2021 American Instructional Resources Survey* (Santa Monica, CA: RAND Corporation, 2022), https://www.rand.org/pubs/research_reports/RRA134-13.html; Jocelyn Pickford and Kate Poteet, "States Take Many Paths to Advance High-Quality Curriculum and Align Professional Learning," *Journal of the National Association of State Boards of Education* 24, no. 1 (2024), https://www.nasbe.org/states-take-many-paths-to-advance-high-quality-curriculum-and-align-professional-learning/; National Student Support Accelerator, *A Snapshot of State Tutoring Policies*, November 2023, https://studentsupportaccelerator.org/sites/default/files/Snapshot%20of%20State%20Tutoring%20Policies.pdf; Autumn Rivera, "Beyond the Bell: Out-of-School Learning Programs Can Help Kids Catch Up and Excel," National Conference of State Legislatures, February 12, 2024, https://www.ncsl.org/state-legislatures-news/details/beyond-the-bell-out-of-school-learning-programs-can-help-kids-catch-up-and-excel.

30 Croft et al., *Testing the Waters*; Croft et al., *Demystifying Statewide Standardized Assessments*; Bonnie O'Keefe and Brandon Lewis, *The State of Assessment: A Look Forward on Innovation in State Testing Systems*, Bellwether, July 2019, https://files.eric.ed.gov/fulltext/ED596503.pdf.

31 ESSA, 20 U.S.C. § 1111(a) and 1111(b)(2)(B)(iii)-(iv), https://www2.ed.gov/documents/essa-act-of-1965.pdf; 34 C.F.R. § 200.2(b)(4) and (5) and (d), https://www.law.cornell.edu/cfr/text/34/200.2; U.S. Department of Education, *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process*, September 2018, https://oese.ed.gov/files/2020/07/assessmentpeerreview.pdf; AERA et al., *Standards for Educational and Psychological Testing*.

32 ESSA, 20 U.S.C. § 1111(b)(2)(B)(xi), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

33 Mary Moran, "Standards and Assessments: The New Measure of Adequacy in School Finance Litigation," *Journal of Education Finance* 25, no. 1 (1999): 37, http://www.jstor.org/stable/40704086.

34 ESSA, 20 U.S.C. § 1111(b)(2)(B)(v), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

35 Tennessee continues to administer paper-and-pencil tests but does so only at the lower elementary grades (i.e., grades 3-5), with older students testing online. Some states allow a paper-and-pencil administration for reasons including a lack of school resources or for a student who needs paper and pencil as an accommodation for a disability. National Center for Education Statistics (NCES), "State Education Practices (SEP)," Table 2.22, https://nces.ed.gov/programs/statereform/tab2_22.asp. Since the NCES data collection, Kentucky now requires computer-based testing except for students with accommodations ("Kentucky Summative Assessment," Kentucky Department of Education, March 22, 2024, https://www.education.ky.gov/AA/Assessments/Pages/KentuckySummativeAssessment-.aspx). North Carolina's 3-8 end-of-grade tests are required to be administered online unless there is a technological hardship or for students with accommodations ("State Tests," Testing and School Accountability, North Carolina Department of Public Instruction, https://www.dpi.nc.gov/districts-schools/testing-and-school-accountability/state-tests#End-of-GradeEOGTests-1576). Tennessee grades 3-5 are using paper-based assessments, but grades 6-8 are administered online (Tennessee Department of Education, *Tennessee Comprehensive Assessment Program (TCAP) Calendar 2023-24 School Year*, July 2022, https://www.tn.gov/content/dam/tn/education/events/2023-24_TCAP_Testing_Calendar_06152023.pdf). Oklahoma has moved administration online (Oklahoma Department of Education, *2022-2023 Assessment Calendar*, https://sde.ok.gov/sites/default/files/2023-2024%20Assessment%20Calendar%20With%20New%20Change.pdf). Wyoming has moved administration online ("WY-TOPP," Wyoming Department of Education, https://edu.wyoming.gov/for-district-leadership/state-assessment/wy-topp/).

36 Catherine Gewertz, "Coronavirus Throws Spring Testing Into Disarray," *Education Week*, March 16, 2020, https://www.edweek.org/leadership/coronavirus-throws-spring-testing-into-disarray/2020/03.

37 ESSA, 20 U.S.C. § 1111(b)(2)(B)(xi), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

38 ESSA, 20 U.S.C. § 1111(b)(2)(B)(x), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

39 Florida Department of Education, *2023-24 FAST Fact Sheet Grades 3-10 ELA Reading and Grades 3-8 Mathematics*, December 2023, 2, https://www.fldoe.org/core/fileparse.php/20102/urlt/2324FASTGrd310FS.pdf.

40 "State Testing Frequently Asked Questions," State Testing, Washington Office of Superintendent of Public Instruction, https://ospi.k12.wa.us/student-success/testing/state-testing/state-testing-frequently-asked-questions.

41 Assessment Solutions Group, personal communication, April 2, 2024. The cost varies by state, with large states paying less and small states paying more.

42 Catherine Gewertz, "What Tests Does Each State Require?," *Education Week*, updated March 5, 2019, https://www.edweek.org/teaching-learning/what-tests-does-each-state-require/2017/02.

43 Marcus A. Winters, *The Costs and Benefits of Test-Based Promotion*, July 30, 2018, available at SSRN, http://dx.doi.org/10.2139/ssrn.3222671; Council of Chief State School Officers, *Third Grade Reading Laws: Implementation and Impact*, July 2019, 4, https://files.eric.ed.gov/fulltext/ED603144.pdf.

44 National Council on Teacher Quality, "Measures of Student Growth National Results" (data set, State Teacher Policy Database, 2022), https://www.nctq.org/yearbook/national/Measures-of-Student-Growth-95.

45 National Association of Charter School Authorizers, *State of Authorizing Report*, https://qualitycharters.org/authorizing-evolving-definitions-of-excellence/.

46 Croft et al., *Demystifying Statewide Standardized Assessments*, 2.

47 For example, in California, 71% of districts used the state's SBAC test as part of math placement decisions. Niu Gao and Sara Adan, *Math Placement in California's Public Schools*, Public Policy Institute of California, November 2016, 6, https://www.ppic.org/wp-content/uploads/content/pubs/report/R_1116NGR.pdf.

48 Pennsylvania breaks out the amount of time spent testing and the administrative tasks related to testing (e.g., having students log in to the computer and reading instructions) in the test administration manual: Pennsylvania Department of Education, *Pennsylvania System of School Assessment Handbook for Assessment Coordinators*, 2023, 39–40, https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/PSSA%20Handbook%20for%20Assessment%20Coordinators.pdf.

49 Ray Hart et al., *Student Testing in America's Great City Schools: An Inventory and Preliminary Analysis, Council of the Great City Schools*, October 2015, https://www.cgcs.org/cms/lib/dc00001581/centricity/domain/87/testing%20report.pdf.

50 During the early days of the assessment consortia, the PARCC assessment was approximately 10 to 11 hours long, administered during two testing windows. The first window was 12 weeks for a performance task, and the second was an 8-week window for the end-of-year component. Catherine Gewertz, "PARCC Shortens Its Common-Core Test," *Education Week*, May 21, 2015, https://www.edweek.org/teaching-learning/parcc-shortens-its-common-core-test/2015/05.

51 Laura S. Hamilton et al., *Standards-Based Accountability Under No Child Left Behind*, RAND, April 29, 2007, https://www.rand.org/pubs/monographs/MG589.html. Some test prep can be beneficial by giving students an opportunity to practice in an environment similar to how they will be tested, so that scores better represent what a student knows and can do instead of familiarity with the testing process. However, too much test preparation can interfere with instructional time.

52 Libby Stanford, "Educators Feel Growing Pressure for Students to Perform Well on Standardized Tests," *Education Week*, September 1, 2023, https://www.edweek.org/teaching-learning/educators-feel-growing-pressure-for-students-to-perform-well-on-standardized-tests/2023/09; Sarah D. Sparks, "Teaching Students to De-Stress Over Testing," *Education Week*, May 24, 2017, https://www.edweek.org/teaching-learning/teaching-students-to-de-stress-over-testing/2017/05.

53 See, for example, Matt Barnum, "Why One Harvard Professor Calls American Schools' Focus on Testing a 'Charade,'" *Chalkbeat*, January 19, 2018, https://www.chalkbeat.org/2018/1/19/21104211/why-one-harvard-professor-calls-american-schools-focus-on-testing-a-charade/; Richard J. Shavelson et al., "Problems With the Use of Student Test Scores to Evaluate Teachers," Briefing Paper no. 278, Economic Policy Institute, August 27, 2010, https://www.epi.org/publication/bp278/.

54 John Fensterwalk, "California Will Give a Short Version of Its Standardized Math and English Tests Next Spring," *EdSource*, September 10, 2021, https://edsource.org/2021/california-will-give-a-short-version-of-its-standardized-math-and-english-tests-next-spring/661039.

55 H. D. Hoover et al., *The Iowa Tests: Interpretative Guide for Teachers and Counselors Forms A and B Levels 9-14* (Itasca, IL: Riverside Publishing, 2003).

56 Hart et al., *Student Testing*, 32.

57 An accommodation is an adjustment in testing (e.g., testing presentation, environment, or format) that does not change what is being measured. Some examples include testing in a small group or large print. AERA et al., *Standards for Educational and Psychological Testing*, 215.

58 Fensterwalk, "California."

59 Robert L. Ebel, "Why Is a Longer Test Usually a More Reliable Test?," *Educational and Psychological Measurement* 32, no. 2 (1972): 249–253, https://journals.sagepub.com/doi/abs/10.1177/001316447203200202.

60 Fensterwalk, "California"; Smarter Balanced Assessment Consortium, *2020-21 Summative Technical Report*.

61 Tim Davey, *Practical Considerations in Computer-Based Testing*, ETS, January 2011, 3, https://www.ets.org/Media/Research/pdf/CBT-2011.pdf.

62 Fensterwalk, "California."

63 Ibid.

64 Center on Standards & Assessment Implementation, *Integrating Speaking and Listening Standards Into Instruction—A Review of Resources*, WestEd and CRESST, January 2017, 3, https://csaa.wested.org/wp-content/uploads/2020/02/CSAI-Report_SpeakingandListeningResources.pdf.

65 U.S. Department of Education, *State's Guide*, 24–25.

66 ESSA, 20 U.S.C. § 1111(b)(2)(A), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

67 ESSA, 20 U.S.C. § 1111(b)(2)(B)(x), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

68 Scott Marion and Derek Briggs, "Just Give Us a Little," *Centerline Blog*, Center for Assessment, July 13, 2022, https://www.nciea.org/blog/just-give-us-a-little/.

69 U.S. Department of Education, *State's Guide*, 37.

70 Scott Marion and Will Lorié, "Can We Reduce Testing in K-12 Schools?," *Centerline Blog*, Center for Assessment, October 18, 2023, https://www.nciea.org/blog/can-we-reduce-testing-in-k-12-schools/.

71 Ruth A. Childs and Andrew P. Jaciw, "Matrix Sampling of Items in Large-Scale Assessments," *Practical Assessment, Research, and Evaluation* 8, Article 16 (January 2019): 6, https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1125&context=pare.

72 National Center for Education Statistics, "National Assessment of Educational Progress (NAEP): 4. Survey Design," in *NCES Handbook of Survey Methods*, updated May 2017, https://nces.ed.gov/statprog/handbook/naep_surveydesign.asp.

73 Smarter Balanced Assessment Consortium, *2020-21 Summative Technical Report*.

74 Edward Roeber, *What Does It Mean to Use Matrix Sampling in Student Assessment?*, Think Point document (Mason, MI: Michigan Assessment Consortium), 1, https://michiganassessmentconsortium.org/wp-content/uploads/ThinkPoint_MatrixSampling3.pdf.pdf.

75 Institute of Education Sciences, *FAQs: Students*, National Assessment of Educational Progress, 2, https://nces.ed.gov/nationsreportcard/subject/students/pdf/students_10_4.pdf.

76 Gewertz, "What Tests?"; Winters, *Costs and Benefits*; Council of Chief State School Officers, *Third Grade Reading Laws*, 4. Center on Enhancing Early Learning Outcomes and Council of Chief State School Officers, *Third Grade Reading Laws: Implementation and Impact*, 4.

77 Childs and Jaciw, "Matrix Sampling," 6.

78 Ibid., 1.

79 For example, see how NAEP selects schools. National Center for Education Statistics, *NAEP Assessment Sample Design*, updated October 25, 2023, https://nces.ed.gov/nationsreportcard/tdw/sample_design/.

80 National Center for Education Statistics, "National Assessment of Educational Progress."

81 Childs and Jaciw, "Matrix Sampling," 6.

82 Alliance for Excellent Education, *N-Size in ESSA State Plans*, November 2018, 1, https://all4ed.org/wp-content/uploads/2018/11/N-Size-in-ESSA-State-Plans.pdf.

83 Johnnie Daniel, "Choosing the Type of Probability Sampling," chap. 5 in *Sampling Essentials: Practical Guidelines for Making Sampling Choices* (Thousand Oaks, CA: SAGE, 2012), 131, https://www.sagepub.com/sites/default/files/upm-binaries/40803_5.pdf; National Center for Education Statistics, "National Assessment of Educational Progress."

84 ESSA, 20 U.S.C. § 1111(b)(2)(B)(v), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

85 National Center for Education Statistics, "National Assessment of Educational Progress."

86 Richard J. Coley and Margaret E. Goertz, *Educational Standards in the 50 States: 1990*, ETS Research Report Series, no. 1 (Princeton, NJ: Educational Testing Service, June 1990): 23, https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.1990.tb01347.x.

87 Juan D'Brot, *Considerations for Including Growth in ESSA State Accountability Systems*, Council of Chief State School Officers, January 2017, 6, https://ccsso.org/sites/default/files/2017-10/CCSSOGrowthInESSAAccountabilitySystems1242017.pdf.

88 Andy Hegedus, *Evaluating the Relationships Between Poverty and School Performance*, NWEA Research, October 2018, 7, https://files.eric.ed.gov/fulltext/ED593828.pdf.

89 Marion and Lorié, "Can We Reduce Testing?"

90 Ibid.

91 Hegedus, *Evaluating the Relationships*, 7.

92 Jason A. Grissom, Demetra Kalogrides, and Susanna Loeb, "Strategic Staffing? How Performance Pressures Affect the Distribution of Teachers Within Schools and Resulting Student Achievement," *American Educational Research Journal 54*, no. 6 (2017): 1079–1116, https://files.eric.ed.gov/fulltext/EJ1162344.pdf. Similarly, teachers in tested grades who had low value-added scores are more likely to move to non-tested positions within their current school or exit teaching. Matthew M. Chingos and Martin R. West, "Promotion and Reassignment in Public School Districts: How Do Schools Respond to Differences in Teacher Effectiveness?," *Economics of Education Review* 30, no. 3 (2011): 419–433, http://nrs.harvard.edu/urn-3:HUL.InstRepos:4889479.

93 Grissom et al., "Strategic Staffing?"

94 Grace Taylor et al., *A Survey of Teachers' Perspectives on High-Stakes Testing in Colorado: What Gets Taught, What Gets Lost*, National Center for Research on Evaluation, Standards, and Student Testing and Center for Research on Evaluation, Diversity and Excellence, September 2001, 27, https://nepc.colorado.edu/sites/default/files/Cosurvey.pdf.

95 For instance, see Dee and Jacob, "Impact of No Child Left Behind," 151, who found a deemphasis on science instruction post–No Child Left Behind.

96 Nicholas Munyan-Penney and Sarah Mehrotra, *Future of Assessments: Centering Equity and the Lived Experiences of Students, Families, and Educators*, Education Trust, April 2023, 12, https://edtrust.org/wp-content/uploads/2014/09/Future_Assessments_Embargoed.pdf.

97 Educators for Excellence, *Voices From the Classroom: A Survey of America's Educators*, 2023, 21, https://e4e.org/sites/default/files/voices_from_the_classroom_2023.pdf; Julia H. Kaufman et al., *U.S. Teachers' Support of Their State Standards and Assessments: Findings From the American Teacher Panel*, RAND, October 4, 2017, https://www.rand.org/pubs/research_reports/RR2136.html.

98 EdReports, *The Pursuit for Transparency in Commercial Interim Assessment Products*, May 24, 2023, https://www.edreports.org/resources/article/the-pursuit-for-transparency-in-commercial-interim-assessment-products; Emma Kate Fittes, "Are Interim Assessments Living Up to Their Billing? New Review Aims to Find Out," *EdWeek Market Brief*, July 12, 2022, https://marketbrief.edweek.org/marketplace-k-12/interim-assessments-living-billing-new-review-aims-find/.

99 Croft et al., *Testing the Waters*, 29.

100 Doan et al., *How States Are Creating*; Pickford and Poteet, "States Take Many Paths."

101 Scott Marion and Katie Buckley, "Design and Implementation Considerations of Performance-Based and Authentic Assessments for Use in Accountability Systems," in *Meeting the Challenges to Measurement in an Era of Accountability*, ed. Henry Braun (New York: Routledge, 2016), 51, http://dx.doi.org/10.4324/9780203781302-4.

102 Smarter Balanced Assessment Consortium, *2020-21 Summative Technical Report*, Section 4.5: Performance Task Design.

103 Marion and Buckley, "Design and Implementation Considerations"; see New Hampshire's Innovative Assessment Demonstration Authority application for an example: State of New Hampshire Department of Education, *Application for the New Authorities Under the Innovative Assessment Demonstration Authority*, 2018, https://www.education.nh.gov/sites/g/files/ehbemt326/files/inline-documents/iada-application.pdf.

104 Marion and Buckley, "Design and Implementation Considerations," 54.

105 For certain standards, such as speaking or longer research tasks, performance assessments are needed because the skill could not be measured (or cannot be measured as well) through selected-response items. We should note that although performance assessments, per se, do not always measure higher-order thinking skills, there are greater opportunities to develop performance tasks that measure higher-order thinking and may help facilitate deeper understanding of knowledge so that students can better transfer and retain knowledge. Ronald K. Hambleton et al., *Psychometric Review of the Maryland School Performance Assessment Program (MSPAP)*, December 2000, xi, https://www.researchgate.net/publication/237501323_Psychometric_Review_of_the_Maryland_School_Performance_Assessment_Program_MSPAP; Scott Marion and Paul Leather, "Assessment and Accountability to Support Meaningful Learning," *Education Policy Analysis Archives* 23, no. 9 (February 2, 2015), https://epaa.asu.edu/index.php/epaa/article/view/1984/1489; D. E. (Sunny) Becker et al., *Formative Evaluation of New Hampshire's Performance Assessment of Competency Education (PACE): Final Report* (Alexandria, VA: HumRRO, March 13, 2017), https://www2.ed.gov/policy/elsec/guid/stateletters/nhpaceformativevalrpt2017.pdf.

106 For instance, in New Hampshire, teachers participated in extensive professional development so that they could develop the tasks, create rubrics for evaluating performance, administer the assessment, and evaluate student performance. Becker et al., *Formative Evaluation*, xviii.

107 Hambleton et al., *Psychometric Review*, iv.

108 Becker et al., *Formative Evaluation*, xxii.

109 Hambleton et al., *Psychometric Review*, x; see also letter from the New Hampshire Department of Education to Dr. Donald Peasley, March 9, 2022, https://oese.ed.gov/files/2022/04/NHIADAWithdrawal3.9.2022.pdf; Becker et al., *Formative Evaluation*, xxviii.

110 Hambleton et al., *Psychometric Review*, 18.

111 Hambleton et al., *Psychometric Review*; Marion and Buckley, "Design and Implementation Considerations," 64.

112 Hambleton et al., *Psychometric Review*, 18.

113 See, e.g., Tomoya Okubo et al., "AI Scoring for International Large-Scale Assessments Using a Deep Learning Model and Multilingual Data," OECD Education Working Papers, no. 287 (Paris: OECD Publishing, 2023), https://dx.doi.org/10.1787/9918e1fb-en.

114 Hambleton et al., *Psychometric Review*, x.

115 Becker et al., *Formative Evaluation*, xxxi.

116 For instance, New Hampshire tried to increase the trustworthiness of the ratings by having two teachers from a different district score student work. Becker et al., *Formative Evaluation*, 14.

117 Marion and Buckley, "Design and Implementation Considerations"; Hambleton et al., *Psychometric Review*, 18; see also letter from the New Hampshire Department of Education to Dr. Donald Peasley, March 9, 2022, https://oese.ed.gov/files/2022/04/NHIADAWithdrawal3.9.2022.pdf.

118 Hambleton et al., *Psychometric Review*, xi.

119 "Montana Aligned to Standards Through-Year Pilot," Montana Office of Public Instruction, https://opi.mt.gov/Leadership/Assessment-Accountability/Montana-Alternative-Student-Testing-Pilot-Program.

120 Dave Powell et al., *What Are Through-Year Assessments? Exploring Multiple Approaches to Through-Year Design*, Education First, 6, https://www.education-first.com/download/12842/?tmstv=1702665297.

121 Ibid.

122 Croft et al., *Testing the Waters*.

123 Nathan Dadey, Carla M. Evans, and Will Lorié, *Through-Year Assessment: Ten Key Considerations*, Center for Assessment, March 2023, https://www.nciea.org/wp-content/uploads/2023/03/Ten-Key-Considerations-Through-Year-Assessment-Report-March2023-F.pdf.

124 Croft et al., *Testing the Waters*, 30.

125 Expert interview.

126 Croft et al., *Testing the Waters*, 26.

127 Nathan Dadey and Brian Gong, *Using Interim Assessments in Place of Summative Assessments? Consideration of an ESSA Option* (Washington, DC: Council of Chief State School Officers, 2017), https://ccsso.org/sites/default/files/2017-12/ASR_ESSA_Interim_Considerations-April.pdf; Brian Gong, "Why Has It Been So Difficult to Develop a Viable Through-Year Assessment System?," *CenterLine Blog*, Center for Assessment, February 23, 2021, https://www.nciea.org/blog/why-has-it-been-so-difficult-to-develop-a-viable-through-year-assessment/.

128 Croft et al., *Testing the Waters*, 21.

129 If a state is going to reuse items in future years, the state must keep the item secure and not release it. States could release items to parents and teachers annually, but doing so would greatly increase the cost of item development.

130 ESSA, 20 U.S.C. § 1111(b)(2)(B)(vi), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

131 Dadey and Gong, *Using Interim Assessments?*; Gong, "Why Has It Been So Difficult?"

132 U.S. Department of Education, *State's Guide*, 57.

133 Andrew Ujifusa, "Overhauling Student Assessments: A View from the Pilot's Seat," *Education Week*, April 2, 2019, https://www.edweek.org/policy-politics/overhauling-student-assessments-a-view-from-the-pilots-seat/2019/04.

134 ESSA, 20 U.S.C. § 1111(b)(2)(B)(ii), https://www2.ed.gov/documents/essa-act-of-1965.pdf; U.S. Department of Education, *State's Guide*, 25.

135 Croft et al., *Testing the Waters*, 24.

136 Lindsay Kubatzky and Meg Benner, *Inclusive, Innovative Assessments for Students With Learning Disabilities*, National Center for Learning Disabilities, 2023, 17, https://ncld.org/wp-content/uploads/2023/08/Inclusive_Innovative_Assessments_for_Students_With_Learning_Disabilities.NCLD_.Final_.pdf.

137 Libby Stanford, "More States Are Creating a 'Portrait of a Graduate.' Here's Why," *Education Week*, December 11, 2023, https://www.edweek.org/policy-politics/more-states-are-creating-a-portrait-of-a-graduate-heres-why/2023/12.

138 ESSA, 20 U.S.C. § 1111(c)(4)(B)(v)(I), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

139 AERA et al., *Standards for Educational and Psychological Testing*, 54.

140 Jennifer Randall et al., "Our Validity Looks Like Justice. Does Yours?" *Language Testing* 41, no. 1 (2024): 203–219, https://journals.sagepub.com/doi/10.1177/02655322231202947; Jennifer Randall, "'Color-Neutral Is Not a Thing: Redefining Construct Definition and Representation Through a Justice-Oriented Critical Antiracist Lens," *Educational Measurement Issues and Practice* 40, no. 4 (2021): 82–90, https://onlinelibrary.wiley.com/doi/abs/10.1111/emip.12429; Center for Measurement Justice, https://measurementjustice.org/.

141 Molly Faulkner-Bond and Priya Kannan, "Student Reactions to Culturally Responsive Mathematics Assessment Items," WestEd, October 5, 2023, https://www.wested.org/wested-bulletin/equity-in-focus/student-reactions-to-culturally-responsive-mathematics-assessment-items/.

142 Tammy Elser, *The Framework: A Practical Guide for Montana Teachers and Administrators Implementing Indian Education for All*, Montana Office of Public Instruction, 2020, https://opi.mt.gov/Portals/182/Page%20Files/Indian%20Education/Indian%20Education%20101/Framework.pdf.

143 Dee and Jacob, "Impact of No Child Left Behind."

144 Corey Mitchell, "Atlanta Educators Convicted in Test-Cheating Trial," *Education Week*, April 2, 2015, https://www.edweek.org/teaching-learning/atlanta-educators-convicted-in-test-cheating-trial/2015/04.

145 Hamilton et al., *Standards-Based Accountability*.

146 Educators for Excellence, "Voices From the Classroom," 21; Kaufman et al., "U.S. Teachers' Support."

147 Will Sentell, "Governor's Bid for Major Changes in Teacher Evaluations Sparking Pushback," *The Advocate*, April 23, 2017, https://www.theadvocate.com/baton_rouge/news/politics/legislature/governors-bid-for-major-changes-in-teacher-evaluations-sparking-pushback/article_02532f12-2484-11e7-90f2-1b78f1652dea.html; Massachusetts Department of Elementary and Secondary Education, *An Introduction to PARCC – Frequently Asked Questions*, January 2014, https://www.mass.gov/doc/an-introduction-to-parcc-frequently-asked-questions-0/download.

148 From ESEA Core Coalition to Senators on the Health, Education, Labor and Pensions Committee and Representatives on the House Education and the Workforce Committee, January 2015, https://edtrust.org/wp-content/uploads/2013/10/ESEACoreCoalitionLetter.pdf.

149 U.S. Department of Education, "Competitive Grants for State Assessments," https://oese.ed.gov/offices/office-of-formula-grants/school-support-and-accountability/competitive-grants-for-state-assessments/.

150 ESSA, 20 U.S.C. § 1111(b)(2)(B)(i)(I), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

151 ESSA, 20 U.S.C. § 1111(b)(2)(B)(i)(II), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

152 ESSA, 20 U.S.C. § 1111(b)(2)(B)(v), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

153 ESSA, 20 U.S.C. § 1111(b)(2)(B)(ii), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

154 ESSA, 20 U.S.C. § 1111(b)(2)(B)(iii), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

155 ESSA, 20 U.S.C. § 1111(b)(2)(B)(iii), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

156 ESSA, 20 U.S.C. § 1111(b)(2)(B)(iv), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

157 U.S. Department of Education, *State's Guide*.

158 ESSA, 20 U.S.C. § 1111(b)(2)(B)(ii), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

159 ESSA, 20 U.S.C. § 1111(b)(1)(A), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

160 ESSA, 20 U.S.C. § 1111(b)(2)(B)(x), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

161 ESSA, 20 U.S.C. § 1111(b)(2)(B)(xi), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

162 ESSA, 20 U.S.C. § 1111(b)(2)(B)(xii), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

163 ESSA, 20 U.S.C. § 1111(b)(2)(B)(vi), https://www2.ed.gov/documents/essa-act-of-1965.pdf.

164 Georgia Department of Education, *Assessment Guide: Introduction and Overview*, Georgia Milestones Assessment System, 2023, 10, https://lor2.gadoe.org/gadoe/file/d0d5b8f7-63d6-4a7f-8e5f-68df45b837d1/1/GM_AssessmentGuide_Intro.pdf.

165 Values represent the average testing time. New York State Testing Program, *English Language Arts and Mathematics Tests: School Administrator's Manual*, 2023, 4–5, https://www.nysed.gov/sites/default/files/programs/state-assessment/38-sam-2023.pdf.

166 California Assessment of Student Performance and Progress, "General Test Administration Information," in *CAASP Online Test Administration Manual*, https://ca-toms-help.ets.org/caaspp-otam/prep-and-planning/general-test-admin-info/#testing-time-and-order-of-administration.

167 Arizona Department of Education, *Test Coordinator's Manual: Grades 3-8 ELA and Math*, Arizona's Academic Standards Assessment, Spring 2024, 9, https://www.azed.gov/sites/default/files/2023/12/AASA_TCM_Spr24.pdf.

168 Pennsylvania Department of Education, *Pennsylvania System of School Assessment Handbook for Assessment Coordinators*, 2023, 39–40, https://www.education.pa.gov/Documents/K-12/Assessment%20and%20Accountability/PSSA/PSSA%20Handbook%20for%20Assessment%20Coordinators.pdf.

169 South Dakota Department of Education, *South Dakota's Online, Summative, Test Administration Manual*, 2024, 34–35, https://sd.portal.cambiumast.com/-/media/project/client-portals/south-dakota/pdf/sd_2023-24-online-summative-tam_final4524.pdf.

170 Colorado Department of Education, *Colorado Statewide Assessments: 2022–23 Testing At-a-Glance*, February 2023, https://www.cde.state.co.us/communications/2023assessmentataglance.

# About the Authors

### MICHELLE CROFT

Michelle Croft is an associate partner at Bellwether in the Policy and Evaluation practice area. She can be reached at **michelle.croft@bellwether.org**.

### BONNIE O'KEEFE

Bonnie O'Keefe is a senior associate partner at Bellwether in the Policy and Evaluation practice area. She can be reached at **bonnie.okeefe@bellwether.org**.

### MARISA MISSION

Marisa Mission is a senior analyst at Bellwether in the Policy and Evaluation practice area. She can be reached at **marisa.mission@bellwether.org**.

### JULIET SQUIRE

Juliet Squire is a senior partner at Bellwether in the Policy and Evaluation practice area. She can be reached at **juliet.squire@bellwether.org**.

# About Bellwether

Bellwether is a national nonprofit that exists to transform education to ensure systemically marginalized young people achieve outcomes that lead to fulfilling lives and flourishing communities. Founded in 2010, we work hand in hand with education leaders and organizations to accelerate their impact, inform and influence policy and program design, and share what we learn along the way. For more, visit **bellwether.org**.


### ACKNOWLEDGMENTS

We would like to thank the many individuals who gave their time and shared their knowledge with us to inform our work, including Chad Aldeman, Ashley Eden, Gretchen Guffy, Christy Hovanetz, Khaled Ismail, Patricia Levesque, Scott Marion, and Laura Slover. Thank you also to the Walton Family Foundation for its financial support of this project.

We would also like to thank our Bellwether colleague Alexis Richardson for her support. Thank you to Amy Ribock, Kate Neifeld, Andy Jacob, Zoe Campbell, Julie Nguyen, and Amber Walker for shepherding and disseminating this work, and to Super Copy Editors.

The contributions of these individuals and entities significantly enhanced our work; however, any errors in fact or analysis remain the responsibility of the authors.