Executive Summary

Approximately one-third of American freshmen at two-year and four-year colleges require remedial coursework and over 40 percent of employers rate new hires with a high school diploma as "deficient" in their overall preparation for entry-level jobs.[9, 10] Yet, over the past decade, as these students marched through America's public education system, officials repeatedly told them, and their parents, that they were on track for success. They passed their courses, got good grades, and aced state annual tests. To put it plainly, it was all a lie. Imagine being told year after year that you're doing just fine—only to find out when you apply for college or a job that you're simply not as prepared as you need to be.

Thankfully, states have taken courageous steps to address this preparedness gap. Over the past five years, every state has upgraded its K–12 academic standards to align with the demands of college and career readiness (CCR), either by adopting the Common Core State Standards (CCSS) or working with their own higher education and career training providers to strengthen or develop standards. New assessments intended to align to these more-rigorous standards made their debut in the past year or two, and, as was widely expected (and, indeed, inevitable), student proficiency rates are lower than on previous tests—often significantly lower. State and local officials must decide whether to forge ahead with the new tests and higher expectations or back down in order to cast more schools and students in a positive (if, perhaps, illusory) light.

Of course, test scores that more accurately predict students' readiness for entry-level coursework or training are not enough. The content of state assessments, too, is an important predictor of the impact of those tests on what is taught and learned. For instance, low-quality assessments poorly aligned with the standards will undermine the content messages of the standards; given the tests' role in accountability under the newly reauthorized Elementary and Secondary Education Act, it is only logical that such tests might contribute to poor-quality instruction.

In short, good tests matter. Of critical importance to this conversation, therefore, is whether the new tests are indeed good and worth fighting for. That's the central question this study seeks to answer.

The Tests

In the pages that follow, we evaluate the quality of four standardized assessments—three new, multi-state assessments and a well-regarded existing state assessment—to determine whether they meet new criteria developed by the Council of Chief State School Officers (CCSSO) for test quality. These new criteria, as explained in the following pages, ask that evaluators take a deep look at whether the assessments target and reliably measure the essential skills and knowledge needed at each grade level to achieve college and career readiness by the end of high school.

9. National Center for Education Statistics (NCES), Digest of Education Statistics, *Percentage of First-Year Undergraduate Students Who Took Remedial Education Courses, by Selected Characteristics: 2003–04 and 2007–08,* Table 241 (Washington, D.C.: NCES, 2010), https://nces.ed.gov/programs/digest/d10/tables/dt10_241.asp.

10. Conference Board et al., "Are They Really Ready To Work? Employers' Perspectives on the Basic Knowledge and Applied Skills of New Entrants to the 21st Century U.S. Workforce" (New York, NY: Conference Board, 2006), http://www.p21.org/storage/documents/FINAL_REPORT_PDF09-29-06.pdf.

We evaluate English language arts/literacy and mathematics assessments for grades 5 and 8 for this quartet of testing programs:

- ACT Aspire
- The Partnership for Assessment of Readiness for College and Careers (PARCC)
- The Smarter Balanced Assessment Consortium (Smarter Balanced)
- The Massachusetts Comprehensive Assessment System (MCAS, 2014)

The Study Design

The analysis that follows was designed to answer three questions:

1. Do the assessments place strong emphasis on the most important content for college and career readiness (CCR), as called for by the Common Core State Standards and other CCR standards? (**Content**)
2. Do they require all students to demonstrate the range of thinking skills, including higher-order skills, called for by those standards? (**Depth**)
3. What are the overall strengths and weaknesses of each assessment relative to the examined criteria for ELA/Literacy and mathematics? (**Overall Strengths and Weaknesses**)

To answer these questions, we use a new methodology based on the CCSSO's 2014 "Criteria for Procuring and Evaluating High-Quality Assessments."[11] Developed by experts at the National Center for the Improvement of Educational Assessment (NCIEA), this methodology evaluates the degree to which test items and supporting program documentation (e.g., test blueprints and documents describing the item creation process) measure the critical competencies reflected in college and career readiness standards, thereby sending clear signals about the instructional priorities for each grade.[12]

The evaluation was conducted by review panels composed of practitioners, content experts, and specialists in assessment. Following reviewer training and a calibration exercise, the panels evaluated test items across various dimensions, with three to four experts reviewing each test form. Results were aggregated for each test form, discussed among the panel members, combined with results from a review of program documentation, and turned into group ratings and summary statements about each program.

The quality and credibility of an evaluation of this type rests largely on the expertise and judgment of the individuals serving on the review panels. To recruit highly qualified yet impartial reviewers, the study team requested recommendations from each of the four testing programs; from other respected content, assessment, and alignment experts; and from several national and state organizations. Reviewers were carefully vetted for their familiarity with the CCSS, their experience with developing or evaluating assessment items, and potential conflicts of interest. Individuals currently or previously employed by participating testing organizations and writers of the CCSS were not considered. (For more information, see Section I, *Selection of Review Panels*.) To ensure fairness and a balance of reviewer familiarity with each assessment, each of the panels included at least one reviewer recommended by each testing program.

Two university-affiliated content leads facilitated and reviewed the work of the ELA/Literacy and math review panels. Dr. Charles Perfetti, Distinguished University Professor of Psychology at University of Pittsburgh, served as the ELA/Literacy content lead, and Dr. Roger Howe, Professor of Mathematics at Yale University, served as the mathematics content lead. The names and biographical summaries of all panelists appear in Appendix E.

11. Council of Chief State School Officers (CCSSO), "Criteria for Procuring and Evaluating High-Quality Assessments" (Washington, D.C.: CCSSO, 2014).

12. The National Center for the Improvement of Educational Assessment, Inc. (NCIEA), "Guide to Evaluating Assessments Using the CCSSO Criteria for High Quality Assessments: Focus on Test Content" (Dover, NH: NCIEA, February 2016): http://www.nciea.org/publication_PDFs/Guide%20to%20Evaluating%20CCSSO%20Criteria%20Test%20Content%20020316.pdf.

This study evaluates English language arts and math assessments at grades 5 and 8, while a parallel study led by the Human Resources Research organization (HumRRO) evaluates the high school assessments from the same four testing programs (see Table ES-1). Because both organizations used the same methodology, it made sense to conduct two portions of the review jointly and across all grades: the documentation review and the accessibility review. Documentation results specific to grades 5 and 8 are addressed in this report. Please see HumRRO's report for the results from their evaluation of the high school assessments, as well as results from the joint accessibility review (all grades).[13]

TABLE ES-1

Overview of the Parallel Fordham and HumRRO Studies

	ELA/Literacy Review	Math Review	Documentation Review	Accessibility Review
Fordham Study	Grades 5 and 8	Grades 5 and 8	Joint Panel	Joint Panel
HumRRO Study	High School	High School	(grades 5 and 8 findings presented in this report; high school findings presented in HumRRO report)	(presented in HumRRO report)

The CCSSO Criteria for High-Quality Assessments

To evaluate assessments intended to measure student mastery of the Common Core State Standards, we needed a new methodology that would capture their key dimensions. Traditional alignment methodologies offer the advantage of having been studied extensively, but treat each of the grade-level standards with equal importance, creating an inadvertent incentive for tests—and instruction—to be "a mile wide and an inch deep."

The CCSSO's "Criteria for Procuring and Evaluating High-Quality Assessments" was the basis of the new methodology. Specifically designed to address tests of college and career readiness, these criteria focus the evaluation on the highest priority skills and knowledge at each grade in the CCSS, addressing foundational as well as complex skills. By using the CCSSO Criteria as the basis of the methodology, the evaluation rewards those tests that focus on the essential skills and give clear signals about the instructional priorities for each grade.

The CCSSO Criteria address six domains, but only two pertain to the research questions addressed in this study: those for the assessment of ELA/Literacy standards and the assessment of mathematics standards (see Table ES-2).

In addition, CCSSO defined ratings for test content and depth, each of which is based on a subset of ratings. The Content rating reflects the degree to which each test assesses the material most needed for college and career readiness, and the Depth rating reflects the degree to which each test assesses the depth and complexity of the college and career readiness standards.

13. This study also originally included an evaluation of test program transparency, or the extent to which programs provide sufficient information to the public regarding assessment design and expectations (CCSSO criterion A.6). Due to several challenges associated with this review, however, we ultimately decided to drop this criterion from our study. Review panelists were not able to review all relevant documentation for each program, due to the vast volume of materials provided and publicly available. In addition, many test programs continued to release additional information (such as sample items) since our review occurred, rendering this panel's findings somewhat outdated.

TABLE ES-2

CCSSO Criteria Evaluated in This Study

Assessment of ELA/Literacy Standards

Test Content Criteria

- B.3 Requiring students to read closely and use evidence from texts
- B.5 Assessing writing from sources
- B.6 Emphasizing vocabulary and language skills
- B.7 Assessing research and inquiry
- B.8 Assessing speaking and listening

Test Depth Criteria

- B.1 Using a balance of high-quality literary and informational texts
- B.2 Focusing on the increasing complexity of texts across grades
- B.4 Requiring a range of cognitive demand
- B.9 Ensuring high-quality items and a variety of item types

Assessment of Mathematics Standards

Test Content Criteria

- C.1 Focusing strongly on the content most needed for success in later mathematics (i.e., the major work of the grade)
- C.2 Assessing a balance of concepts, procedures, and applications

Test Depth Criteria

- C.3 Connecting mathematics practices to mathematical content
- C.4 Requiring a range of cognitive demand
- C.5 Ensuring high-quality items and a variety of item types

Findings

Results are organized around the key research questions above.

✓ RESULTS FOR QUESTIONS #1 AND #2

Do the assessments place strong emphasis on the most important content for college and career readiness (CCR) as called for by the Common Core State Standards and other CCR standards? **(Content)**

Do they require all students to demonstrate the range of thinking skills, including higher-order skills, called for by those standards? **(Depth)**

The panels assigned one of four ratings to each ELA/Literacy and math criterion: Excellent Match, Good Match, Limited/Uneven Match, or Weak Match. To generate these, each panel reviewed the ratings from the grade 5 and grade 8 test forms, considered the results of the documentation review, and came to consensus on the criterion rating.

Table ES-3 shows the ratings for test content and depth in ELA/Literacy and mathematics across the four programs.

The PARCC and Smarter Balanced assessments earned an Excellent or Good Match to the CCSSO Criteria for both ELA/Literacy and mathematics. While ACT Aspire and MCAS did well regarding the quality of items (see Section I, *Results*) and the Depth of Knowledge assessed (Depth), the panelists found that these two programs do not adequately assess—or may not assess at all—some of the priority content in both ELA/Literacy and mathematics at one or both grades in the study (Content).

TABLE ES-3

Overall Content and Depth Ratings for ELA/Literacy and Mathematics

	ACT Aspire	MCAS	PARCC	Smarter Balanced
ELA/Literacy CONTENT	L	L	E	E
ELA/Literacy DEPTH	G	G	E	G
Mathematics CONTENT	L	L	G	G
Mathematics DEPTH	G	E	G	G

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

Criterion Level Results for ELA/Literacy and Mathematics

The Content and Depth ratings are based on the results of subsets of the CCSSO Criteria, as described above. NCIEA also recommended that certain criteria be "emphasized," meaning awarded greater weight in the final determinations (though precise weightings were not specified). The panels, however, sometimes chose not to adhere to the weighting based on their level of confidence in reviewing each criterion (see Section I, *Methodology Modifications*).

Tables ES-4A and 4B show the distribution of the ELA/Literacy and math criteria ratings. Immediately striking in ELA is that the two consortia assessments (PARCC and Smarter Balanced, which received development grants from the U.S. Department of Education) earned twice as many ratings of Good and Excellent Match as the other two programs, earning eight high ratings to the four of ACT Aspire and MCAS. PARCC earned the most Excellent Match ratings (six), while Smarter Balanced was the only assessment with no ratings of Weak Match (partly because it was also the only program to test listening on the summative assessment).

TABLE ES-4A

ELA/Literacy Ratings Tally by Program

ACT Aspire	E	G	G	G	L	L	L	W	W
MCAS	E	G	G	G	L	L	W	W	W
PARCC	E	E	E	E	E	E	G	G	W
Smarter Balanced	E	E	E	E	G	G	G	G	L

TABLE ES-4B

Mathematics Ratings Tally by Program[14]

ACT Aspire	E	E	L	W					
MCAS	E	E	E	L					
PARCC	E	G	G	G					
Smarter Balanced	E	G	G	L					

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

14. Although all four programs require the assessment of conceptual understanding, procedural skill/fluency, and applications (criterion C.2), final ratings could not be determined with confidence due to variations in how reviewers understood and implemented this criterion.

The ratings for mathematics (Table ES-4B) were more similar between programs, with PARCC earning four Excellent or Good Match ratings, Smarter Balanced and MCAS three each, and ACT Aspire two. MCAS scored particularly well on the three Depth criteria in mathematics, while PARCC is the only assessment that earned all Good Match or better scores.

Tables ES-5A and ES-5B on the following pages provide the final criterion ratings for each program, organized by Content and Depth. They also provide the specific attributes required to fully meet each criterion as indicated in the methodology.[15] Those criteria followed by an asterisk were awarded greater emphasis during development of the Content and Depth ratings.

TABLE ES-5A

Criterion Ratings for ELA/Literacy

CONTENT		ACT Aspire	MCAS	PARCC	Smarter Balanced
B.3*	**Reading:** Items require close reading and use of direct textual evidence, and focus on central ideas and important particulars. **To Meet the Criterion:** 1) Nearly all reading items require close reading and analysis of text, rather than skimming, recall, or simple recognition of paraphrased text. 2) More than half of the reading score points are based on items that require direct use of textual evidence. 3) Nearly all items are aligned to the specifics of the standards. 4) More than half of the reading score points are based on items that require direct use of textual evidence.	L	G	E	E
B.5*	**Writing:** Test programs assess a variety of types and formats of writing and the use of writing prompts that require students to confront and use evidence from texts or other stimuli directly. **To Meet the Criterion**: 1) All three writing types (expository, narrative, and persuasive/argument) are approximately equally represented across all forms in the grade band (K–5 and 6–12), allowing blended types (those that combine types) to contribute to the distribution. 2) All writing prompts require writing to sources (are text-based).	L	W	E	E
B.6	**Vocabulary and Language Skills:** Test forms place adequate emphasis on language and vocabulary items on the assessment, assess vocabulary that reflect requirements for college and career readiness, and focus on common student errors in language questions. **To Meet the Criterion:** 1) The large majority of vocabulary items (i.e., three-quarters or more) focuses on Tier 2 words and requires use of context, and more than half assess words important to central ideas. 2) A large majority (i.e., three-quarters or more) of the items in the language skills component and/or scored with a writing rubric mirror real-world activities, focus on common errors, and emphasize the conventions most important for readiness. 3) Vocabulary is reported as a sub-score or at least 13 percent of score points are devoted to assessing vocabulary/language. 4) Same as #3 for language skills.	G	L	E	G

* Criterion awarded greater weight in determination of Content and Depth rating.

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

15. Note: As first implementers of the methodology, the reviewers made a number of modifications they deemed important for improvement. See Section I, *Methodology Modifications*.

CONTENT		ACT Aspire	MCAS	PARCC	Smarter Balanced
B.7	**Research and Inquiry:** Test forms include research items/tasks requiring students to analyze, synthesize, organize, and use information from multiple sources. **To Meet the Criterion:** The large majority (i.e., three-quarters or more) of the research items require analysis, synthesis, and/or organization of information.	L	W	E	E
B.8	**Speaking and Listening:** (*Not yet required by the criteria, so not included in the Content rating. Listening requirements are listed here because one program assesses listening.)* Items assess students' listening skills using passages with adequate complexity and assess students' speaking skills through oral performance tasks. **To Meet the Criterion:** 1) The large majority (i.e., at least three-quarters) of listening items meet the requirements outlined in B.1 and B.2 and evaluate active listening skills.	W	W	W	L
DEPTH		**ACT Aspire**	**MCAS**	**PARCC**	**Smarter Balanced**
B.1*	**Text Quality and Types:** Test forms include a variety of text types (narrative and informational) that are of high quality, with an increasing focus on diverse informational texts across grades. **To Meet the Criterion:** 1) Approximately half of the texts at grades 3–8 and two-thirds at high school are informational, and the remainder literary. 2) Nearly all passages are high quality (previously published or of publishable quality). 3) Nearly all informational passages are expository in structure. 4) For grades 6–12, the informational texts are split nearly evenly between literary nonfiction, history/social science, and science/technical texts.	G	G	G	E
B.2	**Complexity of Texts:** *(based on documentation review only)* Assessments include texts that have appropriate levels of text complexity for the grade or grade band (grade bands identified in the CCSS are K–5 and 6–12). **To Meet the Criterion:** 1) The documentation clearly explains how quantitative data are used to determine grade band placement. 2) Texts are then placed at the grade level recommended by qualitative review. 3) Text complexity rating process results in nearly all passages being placed at a grade band and grade level justified by complexity data.	G	G	G	G
B.4	**Matching the Complexity of the Standards:** Each test form contains an appropriate range of cognitive demand that adequately represents the cognitive demand of the standards. **To Meet the Criterion:** 1) The distribution of cognitive demand of the assessment matches the distribution of cognitive demand of the standards as a whole and matches the higher cognitive demand (DOK 3+) of the standards. *(Note: This is not a rating of test difficulty. Assessments that do not match the DOK distribution of the standards, even if there are too many high DOK items, may receive a rating less than Excellent Match. See Appendix A for more information.)*	W	L	E	G
B.9	**High-Quality Items and Variety of Item Types:** Test items are of high quality, lacking technical or editorial flaws and each test form contains multiple item types including at least one type in which students construct, rather than select, a response. **To Meet the Criterion:** 1) All or nearly all operational items reviewed reflect technical quality and editorial accuracy. 2) At least two item formats are used, including one that requires students to generate, rather than select, a response.	E	E	E	G

* Criterion awarded greater weight in determination of Content or Depth rating.

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

Cells for which the ratings are not used in determining Content and Depth ratings (See Section I, *Weighting of Criteria for Content and Depth Ratings*.)

TABLE ES-5B

Criterion Ratings for Mathematics

CONTENT		ACT Aspire	MCAS	PARCC	Smarter Balanced
C.1*	**Focus:** Each test form contains a strong focus on the content most crucial for success in later mathematics (i.e., the major work of the grade). **To Meet the Criterion:** The vast majority (i.e., at least three-quarters in elementary grades, at least two-thirds in middle school grades, and at least half in high school) of score points in each assessment focus on the content that is most important for students to master in that grade band in order to reach college and career readiness (the major work of the grade).	W	L	G	G
C.2	**Concepts, Procedures, and Applications:** Each test form contains items that assess conceptual understanding, procedural skill/fluency, and application in approximately equal proportions. **To Meet the Criterion:** The distribution of score points reflects a balance of mathematical concepts, procedures/fluency, and applications.	*Due to variations in how reviewers understood and implemented this criterion, final ratings could not be determined with confidence.*			
DEPTH		**ACT Aspire**	**MCAS**	**PARCC**	**Smarter Balanced**
C.3	**Connecting Practice to Content:** Assessments test students' use of mathematical practices through test items that connect these practices with grade-level content standards. **To Meet the Criterion:** All or nearly all items that assess mathematical practices also align to one or more content standards.	E	E	E	E
C.4*	**Matching the Complexity of the Standards:** Each test form contains an appropriate range of cognitive demand that adequately represents the cognitive demand of the standards. **To Meet the Criterion:** 1) The distribution of cognitive demand of the assessment matches the distribution of cognitive demand of the standards as a whole and matches the higher cognitive demand (DOK 3+) of the standards. *(Note: This is not a rating of test difficulty. Assessments that do not match the DOK distribution of the standards, even if there are too many high DOK items, may receive a rating less than Excellent Match. See Appendix A for more information.)*	L	E	G	G
C.5*	**High-Quality Items and Variety of Item Types:** Test items are of high quality, lacking technical or editorial flaws, and each test form contains multiple item types, including at least one type in which students construct, rather than select, a response. **To Meet the Criterion:** 1) All or nearly all operational items reviewed reflect technical quality and editorial accuracy. 2) At least two item formats are used, including one that requires students to generate, rather than select, a response.	E	E	G	L

* Criterion awarded greater weight in determination of Content or Depth rating.

LEGEND E Excellent Match G Good Match L Limited/Uneven Match W Weak Match

In the ELA/Literacy assessments, all four programs receive high ratings for the quality of items and variety of item types. In addition, all pay close attention to the use of high-quality informational and literary texts and increasing the complexity of tests across grades, which are significant advances over many previous state ELA assessments. Significant differences exist across the testing programs, however, in the degree to which their writing tasks require students to use evidence from sources and the extent to which research skills are assessed. In these areas, PARCC and Smarter Balanced perform well, receiving higher ratings than either ACT Aspire, which receives a rating of Limited/Uneven Match on these criteria, or MCAS, which receives a rating of Weak Match. PARCC and Smarter Balanced assessments also contain a distribution of cognitive demand that better reflects that of the standards, when compared to ACT Aspire and MCAS.

In mathematics, PARCC and Smarter Balanced receive a rating of Good Match for the degree to which their tests focus on the most important content of the grade. ACT Aspire test forms receive a rating of Weak Match on this prioritized criterion, due to their test design choice, in which off-grade standards are assessed in order to monitor mastery across grades. MCAS receives a rating of Limited/Uneven because its grade 5 forms do not contain

Supplemental Analysis: Assessment of Higher-Order Thinking Skills

CCSSO criteria B.4 and C.4 capture the degree to which the range of cognitive demand on the test forms match that of the CCSS. We used Webb's Depth of Knowledge (DOK) taxonomy to assess cognitive demand, as it is by far the most widely used approach to categorizing cognitive demand Webb's DOK is composed of four levels. Level 1 is the lowest level (recall), Level 2 requires use of a skill or concept, and Levels 3 and 4 are higher-order thinking skills. We compared the DOK of the assessments to those of the Common Core State Standards, which were coded by content experts (see Section I, *Selection of Review Panels and Assignment to Forms*). We also compared the tests' DOK distributions to those of fourteen highly regarded previous state assessments, as well as the distribution reflected in several national and international assessments—including Advanced Placement (AP), the National Assessment of Education Progress (NAEP), and the Program for International Student Assessment (PISA).[16, 17]

We found that the CCSS call for greater emphasis on higher-order skills than fourteen highly regarded previous state assessments in ELA/Literacy at both grades 5 and 8 as well as in math at grade 8 (they are similar at grade 5). In addition, the grade 8 CCSS in both ELA/Literacy and math call for greater emphasis on higher-order thinking skills than either NAEP or PISA, both of which are considered to be high-quality, challenging assessments.

Overwhelmingly, the assessments included in our study were found to be more challenging—placing greater emphasis on higher-order skills—than prior state assessments, especially in mathematics (where prior assessments rarely included items at DOK 3 or 4 at all). In some cases, the increase was dramatic: PARCC's DOK in grade 8 exceeds even that of AP and PISA in both subjects. See Appendix A for more.

However, the panels found significant variability in the degree to which the four assessments match the distribution of DOK in the CCSS. In some cases, the panels found significant variability between the grade 5 and grade 8 assessments for a given program. PARCC tests generally have the highest DOK in ELA/Literacy, while ACT Aspire had the highest in mathematics. See Section I, Tables 14 and 22 for the DOK distribution of each program.

16. L. Yuan and V. Le, *Estimating the Percentage of Students who were Tested on Cognitively Demanding Items through the State Achievement Tests* (Santa Monica, CA: RAND Corporation, 2012).

17. *Ibid*.

sufficient focus on the critical content for the grade. With respect to item quality, ACT Aspire and MCAS receive the highest rating of Excellent Match, whereas PARCC receives a rating of Good Match and Smarter Balanced a rating of Limited/Uneven Match.[18]

✓ RESULTS FOR QUESTION #3

What are the overall strengths and weaknesses of each assessment relative to the examined criteria for ELA/Literacy and mathematics? **(Overall Strengths and Weaknesses)**

Each of the review panels developed summary statements for each assessment program, detailing their strengths and areas of improvement in ELA/Literacy and mathematics. In addition, they created summary statements for each test's Content and Depth ratings based on the prioritization of criteria recommended in the study methodology (see Appendix F). They also generated final statements summarizing the observed strengths and areas of improvement for each program.

ACT Aspire

English Language Arts:

In ELA/Literacy, ACT Aspire receives a Limited/Uneven to Good Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The combined set of ELA/Literacy tests (reading, writing, and English) requires close reading and adequately evaluates language skills. More emphasis on assessment of writing to sources, vocabulary, and research and inquiry, as well as increasing the cognitive demands of test items, will move the assessment closer to fully meeting the criteria. Over time, the program would also benefit by developing the capacity to assess speaking and listening skills.

Content: ACT Aspire receives a Limited/Uneven match to the CCSSO Criteria for Content in ELA/Literacy. The assessment program includes an emphasis on close reading and language skills. However, the reading items fall short on requiring students to cite specific textual information in support of a conclusion, generalization, or inference and in requiring analysis of what has been read. In order to meet the criteria, assessing writing to sources, vocabulary, as well as research and inquiry need to be strengthened.

Depth: ACT Aspire receives a rating of Good Match for Depth in ELA/Literacy. The program's assessments are built on high-quality test items and texts that are suitably complex. To fully meet the CCSSO Criteria, more cognitively demanding test items are needed at both grade levels, as is additional literary narrative text—as opposed to literary informational texts.[19]

Mathematics:

In mathematics, ACT Aspire receives a Limited/Uneven to Good Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. Some of the mismatch with the criteria is likely due to intentional program design, which requires that items be included from previous and later grades.

18. The nature and timing of this review required Smarter Balanced to make the test items and forms available to reviewers through an alternate test interface that was more limited than the actual student interface used for the summative assessments, particularly with regard to how items appeared on the screen and how erroneous responses were handled. Though reviewers were not able to determine the extent to which these interface limitations impacted their findings, the study team worked with Smarter Balanced to ascertain which item issues were caused by interface differences and which were not. All item-relevant statements in the report reflect data not prone to interface differences.

19. ACT Aspire does not classify literary nonfiction texts that are primarily narrative in structure as "informational." See Appendix G for more information about ACT Aspire's interpretation of CCSSO criterion B.1.

The items are generally high quality and test forms at grades 5 and 8 have a range of cognitive demand, but in each case the distribution contains significantly greater emphasis at DOK 3 than reflected in the standards. Thus, students who score well on the assessments will have demonstrated a strong understanding of the standards' more complex skills. However, the grade 8 test may not fully assess standards at the lowest level of cognitive demand. The tests would better meet the CCSSO Criteria with an increase in the number of items focused on the major work of the grade and the addition of more items at grade 8 that assess standards at DOK 1.

Content: ACT Aspire receives a Limited/Uneven Match to the CCSSO Criteria for Content in Mathematics. The program does not focus exclusively on the major work of the grade, but rather, by design, assesses material from previous and later grades. This results in a weaker match to the criteria. The tests could better meet the criteria at both grades 5 and 8 by increasing the number of items that assess the major work of the grade.

Depth: ACT Aspire receives a good match to the CCSSO Criteria for Depth in Mathematics. The items are well crafted and clear, with only rare instances of minor editorial issues. The ACT Aspire tests include proportionately more items at high levels of cognitive demand (DOK 3) than the standards reflect and proportionately fewer at the lowest level (DOK 1). This finding is both a strength, in terms of promoting strong skills, and a weakness, in terms of ensuring adequate assessment of the full range of cognitive demand within the standards. While technically meeting the criterion for use of multiple item types, the range is nonetheless limited, with the majority comprising multiple-choice items. The program would better meet the criteria for Depth by including a wider variety of item types and relying less on traditional multiple-choice items.

MCAS

English Language Arts:

In ELA/Literacy, MCAS receives a Limited/Uneven to Good Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The test requires students to closely read high-quality texts and a variety of high-quality item types. However, MCAS does not adequately assess several critical skills—including reading informational texts, writing to sources, language skills, and research and inquiry; further, too few items assess higher-order skills. Addressing these limitations would enhance the ability of the test to signal whether students are demonstrating the skills called for in the standards. Over time, the program would also benefit by developing the capacity to assess speaking and listening skills.

Content: MCAS receives a Limited/Uneven Match to the CCSSO Criteria for Content in ELA/Literacy. The assessment requires students to read closely well-chosen texts and presents test questions of high technical quality. However, the program would be strengthened by assessing writing annually, assessing the three types of writing called for across each grade band, requiring writing to sources, and placing greater emphasis on assessing research and language skills.

Depth: MCAS receives a rating of Good Match for Depth in ELA/Literacy. The assessments do an excellent job in presenting a range of complex reading texts. To fully meet the demands of the CCSSO Criteria, however, the test needs more items at higher levels of cognitive demand, a greater variety of items to test writing to sources and research, and more informational texts—particularly those of an expository nature.

Mathematics:

In mathematics, MCAS receives a Limited/Uneven Match to the CCSSO Criteria for Content and an Excellent Match for Depth relative to assessing whether students are on track to meet college and career readiness standards. The MCAS mathematics test items are of high technical and editorial quality. Additionally, the content is distributed well across the breadth of the grade level standards, and test forms closely reflect the range of cognitive demand of the standards. Yet the grade 5 tests have an insufficient degree of focus on the major work of the grade.

While mathematical practices are required to solve items, MCAS does not specify the assessed practices(s) within each item or their connections to content standards. The tests would better meet the criteria through increased focus on major work at grade 5 and identification of the mathematical practices that are assessed—and their connections to content.

Content: MCAS receives a Limited/Uneven Match to the CCSSO Criteria for Content in Mathematics. While the grade 8 assessment focuses strongly on the major work of the grade, the grade 5 assessment does not, as it samples more broadly from the full range of standards for the grade. The tests could better meet the Criteria through increased focus on the major work of the grade on the grade 5 test.

Depth: MCAS receives an Excellent Match to the CCSSO Criteria for Depth in Mathematics. The assessment uses high-quality items and a variety of item types. The range of cognitive demand reflects that of the standards of the grade. While the program does not code test items to math practices, mathematical practices are nonetheless incorporated within items. The program might consider coding items to the mathematical practices and making explicit the connections between specific practices and content standards.

PARCC

English Language Arts:

In ELA/Literacy, PARCC receives an Excellent Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The tests include suitably complex texts, require a range of cognitive demand, and demonstrate variety in item types. The assessments require close reading, assess writing to sources, research, and inquiry, and emphasize vocabulary and language skills. The program would benefit from the use of more research tasks requiring students to use multiple sources and, over time, developing the capacity to assess speaking and listening skills.

Content: PARCC receives an Excellent Match to the CCSSO Criteria for Content in ELA/Literacy. The program demonstrates excellence in the assessment of close reading, vocabulary, writing to sources, and language, providing a high-quality measure of ELA/Literacy content as reflected in college and career readiness standards. The tests could be strengthened by the addition of research tasks that require students to use two or more sources and, as technologies allow, a listening and speaking component.

Depth: PARCC receives a rating of Excellent Match for Depth in ELA/Literacy. The PARCC assessments meet or exceed the depth and complexity required by the Criteria through a variety of item types that are generally of high quality. A better balance between literary and informational texts would strengthen the assessments in addressing the Criteria.

Mathematics:

In mathematics, PARCC receives a Good Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The assessment is reasonably well aligned to the major work of each grade. At grade 5, the test includes a distribution of cognitive demand that is similar to that of the standards. At grade 8, the test has greater percentages of higher-demand items (DOK 3 and 4) than reflected by the standards, such that a student who scores well on the grade 8 PARCC assessment will have demonstrated strong understanding of the standards' more complex skills. However, the grade 8 test may not fully assess standards at the lowest level (DOK 1) of cognitive demand.

The test would better meet the CCSSO Criteria through additional focus on the major work of the grade, the addition of more items at grade 8 that assess standards at DOK 1, and increased attention to accuracy of the items—primarily editorial, but in some instances mathematical.

Content: PARCC receives a Good Match to the CCSSO Criteria for Content in Mathematics. The test could better meet the criteria by increasing the focus on the major work at grade 5.

Depth: PARCC receives a Good Match to the CCSSO Criteria for Depth in Mathematics. The tests include items with a range of cognitive demand, but at grade 8, that distribution contains a higher percentage of items at the higher levels (DOK 2 and 3) and significantly fewer items at the lowest level (DOK 1). This finding is both a strength, in terms of promoting strong skills, and a weakness, in terms of ensuring adequate assessment of the full range of cognitive demand within the standards. The tests include a variety of item types that are largely of high quality. However, a range of problems (from minor to severe) surfaced relative to editorial accuracy and, to a lesser degree, technical quality. The program could better meet the Depth criteria by ensuring that all items meet high editorial and technical standards and by ensuring that the distribution of cognitive demand on the assessments receives sufficient information across the range.

Smarter Balanced

English Language Arts:

In ELA/Literacy, Smarter Balanced receives a Good to Excellent Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The tests assess the most important ELA/Literacy skills of the CCSS, using technology in ways that both mirror real-world uses and provide quality measurement of targeted skills. The program is most successful in its assessment of writing and research and inquiry. It also assesses listening with high-quality items that require active listening, which is unique among the four programs. The program would benefit by improving its vocabulary items, increasing the cognitive demand in grade 5 items, and, over time, developing the capacity to assess speaking skills.

Content: Smarter Balanced receives an Excellent Match to the CCSSO Criteria for Content in ELA/Literacy. The program demonstrates excellence in the areas of close reading, writing to sources, research, and language. The listening component represents an important step toward adequately measuring speaking and listening skills—a goal specifically reflected in the standards. Overall, Smarter Balanced is a high-quality measure of the content required in ELA/Literacy, as reflected in college and career readiness standards. A greater emphasis on Tier 2 vocabulary would further strengthen these assessments relative to the criteria.

Depth: Smarter Balanced receives a rating of Good Match for Depth in ELA/Literacy. The assessments use a variety of item types to assess student reading and writing to source. The program could better meet the depth criteria by increasing the cognitive demands of the grade 5 assessment and ensuring that all items meet high editorial and technical quality standards.

Mathematics:

In mathematics, Smarter Balanced has a Good Match to the CCSSO Criteria relative to assessing whether students are on track to meet college and career readiness standards. The test provides adequate focus on the major work of the grade, although it could be strengthened at grade 5.

The tests would better meet the CCSSO Criteria through increased focus on the major work at grade 5 and an increase in the number of items on the grade 8 tests that assess standards at the lowest level of cognitive demand. In addition, removal of serious mathematical and/or editorial flaws, found in approximately one item per form, should be a priority.[20]

Content: Smarter Balanced receives a Good Match to the CCSSO Criteria for Content in Mathematics. The tests could better meet the criteria by increasing the focus on the major work for grade 5.

Depth: Smarter Balanced receives a Good Match to the CCSSO Criteria for Depth in Mathematics. The exam includes a range of cognitive demand that fairly represents the standards at each grade level. The tests have a strong variety of item types including those that make effective use of technology. However, a range of problems (from minor to severe) surfaced relative to editorial accuracy and, to a lesser degree, technical

20. See footnote 18 for more on Smarter Balanced test interface.

quality. A wide variety of item types appear on each form, and important skills are assessed with multiple items, as is sound practice. The program could better meet the Depth criteria by ensuring that all items meet high editorial and technical standards and that a given student is not presented with two or more virtually identical problems.

******

For too many years, state assessments have generally focused on low-level skills and have given parents and the public false signals about students' readiness for postsecondary education and the workforce. They often weren't very helpful to educators or policymakers either. States' adoption of college and career readiness standards has been a bold step in the right direction. Using high-quality assessments of these standards will require courage: these tests are tougher, sometimes cost more, and require more testing time than the previous generation of state tests. Will states be willing to make the tradeoffs?