THOMAS B. FORDHAM INSTITUTE
ADVANCING EDUCATIONAL EXCELLENCE

MEMO

To: Dr. Mitchell Chester, Commissioner of Elementary and Secondary Education, Massachusetts Department of Elementary and Secondary Education

CC: Carrie Conaway, Associate Commissioner of Planning, Research, and Delivery Systems, Massachusetts Department of Elementary and Secondary Education

James Peyser, Massachusetts Secretary of Education

Laura Slover, Chief Executive Officer of PARCC Inc/Partnership for the Assessment of Readiness for College and Careers

Francine Markowitz, Program Associate: Policy, Research and Design of PARCC Inc/Partnership for the Assessment of Readiness for College and Careers

From: Michael J. Petrilli and Amber Northern, the Thomas B. Fordham Institute

Date: October 30, 2015

Re: Evaluation of the Content and Quality of the 2014 MCAS and PARCC relative to the CCSSO Criteria for High Quality Assessments

We are pleased to provide this special report as requested by the Massachusetts Department of Elementary and Secondary Education. We hope that it will provide useful information for planning purposes in Massachusetts relative to future assessments of English language arts/literacy and mathematics for students in grades 3–8.

The data and analyses provided herein are a **subset** of the final results of an evaluation that examined the content and quality of four assessments relative to the CCSSO Criteria for High Quality Assessments in ELA/literacy and mathematics. The Thomas B. Fordham Institute convened expert panels in summer 2015 to evaluate operational grade five and eight assessments (end of elementary and middle school), while the Human Resources Research Organization (HumRRO) conducted a similar evaluation at the high school level (reported separately).

This special report includes high-level takeaways from the Fordham study for two of the four assessment programs: the 2014 MCAS and the Partnership for Assessment of Readiness for College and Careers (PARCC). The full report, to be released in January 2016, will add results from Smarter Balanced and ACT Aspire, as well as detailed supplementary analyses (that do not impact the final ratings contained herein). The report was co-authored by Nancy Doorey and Morgan Polikoff, whose biographies appear at the end.

Should you have any questions about the results that follow, do not hesitate to contact us.

Kindly,

Michael J. Petrilli	Amber M. Northern
President	Senior Vice President for Research
mpetrilli@edexcellence.net	anorthern@edexcellence.net

Special Report:

Evaluation of the Massachusetts Comprehensive Assessment System (MCAS) and the Partnership for the Assessment of Readiness for College and Careers (PARCC)

By Nancy Doorey and Morgan Polikoff

Overview

The Thomas. B. Fordham Institute is engaged in a study intended to provide much-needed information to researchers, policymakers, and practitioners about the quality of new assessments and their potential to support effective implementation of college and career readiness standards. The study takes an unprecedented, in-depth look at three new multi-state assessments and an existing single-state assessment. All four purport to measure college and career readiness standards, and the Common Core State Standards (CCSS) specifically.

This special report provides a **subset** of the final results of the larger study (above) that will be released in January 2016. We release these partial results now per a request of the Massachusetts Department of Elementary and Secondary Education. The state of Massachusetts is in the process of deciding which assessment it will use to assess students' mastery of English language arts and mathematics in the future, and these results are of interest. Both the larger "four-test" study and this "two-test" component of it examine the content and quality of assessment programs at grades five and eight, relative to the CCSSO Criteria for High Quality Assessments in ELA/literacy and mathematics.[1]

Three of the evaluated assessments are currently in use across multiple states: the Partnership for the Assessment of Readiness for College and Careers (PARCC), the Smarter Balanced Assessment System (Smarter Balanced), and ACT Aspire. The fourth assessment, the 2014 version of the Massachusetts Comprehensive Assessment System (MCAS), is a highly regarded state test. Widely known as the "best in class" of the previous generation of state assessments, it serves as a comparison point or best-case example.

This special report provides results for MCAS and PARCC only. Results for ACT Aspire and Smarter Balanced will be added to the final January report.

The study addresses the following questions:

1) Do the assessments place strong emphasis on the most important content of college and career readiness standards for the pertinent grade level, including the Common Core State Standards (CCSS)? Do they require all students to demonstrate the higher-order thinking skills reflected in those standards? (**Content** and **Depth**)
2) Are they accessible to all students, including students with disabilities and English language learners? (**Accessibility**)
3) What are the **overall strengths and weaknesses** of each assessment relative to the criteria?

[1] According to the CCSSO document, it provides "criteria for states to consider as they develop procurements and evaluate options for high-quality state summative assessments aligned to college and career readiness standards." It also provides the type of "evidence [that] states could ask vendors to provide to demonstrate that the criteria have been or will be met."

Approach

The Thomas B. Fordham Institute convened expert panels in summer 2015 to evaluate the test items and program documentation for the grade five and eight assessments (end of elementary and middle school).[2] The study provides an unprecedented, in-depth examination of the content and quality of these assessments. Panels of expert reviewers were granted access to operational test forms from each program.

The reviewers used a new methodology developed by the National Center for the Improvement of Educational Assessment (NCIEA), under the leadership of Brian Gong and Scott Marion, to determine how well these tests measure the knowledge and critical skills required at each grade level to attain college and career readiness.[3] The methodology is based on the content-specific portions of the Council of Chief State School Officers (CCSSO)'s "Criteria for Procuring and Evaluating High Quality Assessments" (sections B and C, alignment to standards for both math and English language arts (ELA), as well as section A5, which covers accessibility of the assessment for special populations (see Appendix A for full criteria, including those not examined in this study).[4] While the CCSSO criteria were developed for evaluating assessments aligned to college and career readiness standards generally, they (and the methodology) include elements that are specific to the CCSS.

This methodology differs from earlier generations of test alignment studies in two noteworthy ways. First, most prior studies of this kind assumed a one-to-one match between standards and test items. The CCSS and other standards of college and career readiness, however, describe numerous complex competencies that cannot be assessed with individual test items—such as the ability to craft clearly written arguments supported by evidence from multiple sources, or applying multiple mathematical skills to solve a complex problem. Such competencies are essential for postsecondary education, career training, and citizenship, and states need to know whether their assessments are measuring them. This new approach to alignment and quality asks evaluators to determine the degree to which the tests measure these more complex competencies.

The second key difference of our approach is that it focuses the evaluation on the highest-priority skills and knowledge at each grade. Prior approaches treated each of the grade-level standards with equal importance, creating an inadvertent incentive for tests—and instruction—that are "a mile wide and an inch deep." The CCSSO criteria (hereafter "the Criteria") acknowledge the prioritized skills and competencies defined for each grade level within college and career readiness standards, including the CCSS.[5]

[2] Time and budgetary considerations prohibited examination of all grade levels. These two were selected because they are the "capstone" grades at the elementary and middle school levels.

[3] The Center's published methodology document is available upon request and will also be published this fall.

[4]Also available here: http://www.ccsso.org/Documents/2014/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

[5] This study, and the parallel high school evaluation conducted by HumRRO, were the first to use this new methodology developed by the NCIEA. There are noteworthy strengths to these methods, particularly in how they provide a deep look at critical aspects of tests designed to measure college and career readiness. That said, they are not perfect, and choices were made during the development and implementation of the methodology that will inevitably be refined over time. This memo attempts to be clear about how the criteria were operationalized so that readers can overlay their own priorities. The final report will contain recommendations for future enhancements to the methodology, as well as responses from the assessment programs to the findings.

The evaluations of the mathematics and ELA/literacy assessments of each test program were conducted by panels of experts who reviewed operational test forms (more below). A subset of these experts also reviewed program documentation to evaluate the generalizability of the findings. A short description of the study components and processes follow, then the results. We close with a summary of each program's strengths and weaknesses relative to the Criteria.

Study Components

Panel Recruitment and Training

Great care was taken in recruiting, vetting, and selecting panelists for the mathematics and ELA/literacy panels. Recommendations for neutral, high-quality reviewers were requested from all four participating test programs and from other national assessment and content experts and organizations. We sought individual reviewers from three categories: classroom educators, higher education math and ELA content experts and consultants, and assessment experts. Former or current employees of participating testing organizations and writers of the CCSS were not eligible for consideration (though individuals may have served on a feedback or technical advisory group at some point prior to our evaluation). Recruitment priorities included deep knowledge of the Common Core and prior experience with alignment studies.

Given that most content and assessment experts have become experts by working on prior alignment or assessment development studies—and that it is impossible to find individuals with zero conflicts who are also experts—we prioritized balance and fairness. We recruited at least one reviewer recommended by each testing program to serve on each panel. This strategy helped to ensure fairness by balancing reviewers' familiarity with the various assessments on each panel. (Reviewer names and biographies will be included in the final January report once the larger study has been completed.)

Two university-affiliated "content leads" facilitated the work of the content area panels. Dr. Charles Perfetti, distinguished university professor of psychology at the University of Pittsburgh, served as the ELA/literacy Content Lead, and Dr. Roger Howe, professor of mathematics at Yale University, served as mathematics Content Lead. In total, sixteen reviewers conducted the item review for MCAS (eight for math and eight for ELA), and twenty-three reviewers assessed PARCC items and forms (twelve for math and eleven for ELA). [6] Within each content area, reviewers were randomly assigned to test forms while maintaining the desired balance of areas of expertise.

We conducted extensive training and calibration preparation prior to the reviews to build common understanding and consistent judgments across reviewers. Through a combination of online and in-person sessions, training was provided on the design of the four programs (delivered by program representatives) as well as the Criteria and the scoring guidelines provided to reviewers. Portions of the training were delivered by Student Achievement Partners, HumRRO, and several consultants, including alignment and assessment experts.

Next, we briefly describe the two phases of the assessment review process itself: the item review and the documentation review.

[6] The smaller number of reviewers assigned to MCAS was due solely to the fact that only one test form is available. See additional explanation under "Item Review".

Item Review

The study methodology calls for review of two different test forms per grade level and content area.[7] We reviewed the PARCC assessments using these specifications. However, as the MCAS assessment is composed of one operational form per grade level and content area, our reviewers evaluated only the one existing form. Reviewers accessed online test form items for PARRC and evaluated documents (PDFs) of MCAS's paper forms (the format in which each test is administered in Massachusetts).[8]

For both mathematics and ELA/literacy, reviewers evaluated the technical quality of test items, the degree to which test forms place sufficient emphasis (per the Criteria) on the most important skills and knowledge at the pertinent grade level, and the degree to which each test embodies the depth and complexity of the standards.

Documentation Review

The NCIEA methodology also requires panels to evaluate whether the results from the review of one or two test forms per grade are generalizable. In other words, would item-review ratings likely remain the same, improve, or decline if all test forms built from the same blueprints and other test specifications had been reviewed?

Fordham and HumRRO convened a joint review panel composed of eight math and ELA/literacy reviewers from across the two studies. This group reviewed the documentation provided by each program and developed ratings for each criterion regarding the generalizability of the findings. For example, criterion B.1 calls for text passages used within the assessments to be balanced across literary and informational text types, with more informational than literary texts used at the higher grade bands. To determine whether the results from the item review are generalizable, this joint panel reviewed the program documentation to determine whether it required the prescribed balance across the grade bands for all test forms.

Development of Program Ratings: ELA/Literacy and Mathematics

In accord with the methodology, steps were taken to aggregate the ratings of individual reviewers across forms, grade levels, and programs.

First, individual reviewers independently completed their evaluations of individual test forms. They rated the forms against each individual CCSSO criterion according to a 0, 1, 2 scale, which indicated "Does Not Meet," "Partially Meets," and "Meets" the criterion, respectively. Next, scores from all individual reviews were gathered, and reviewers discussed each individual score and the associated reviewers' comments. The group then determined a "group match score" for *each form*, which was simply their agreed-upon score after talking through rater differences and settling on a score that represented the collective judgment of the panel.[9] (If they were unable to reach consensus, minority opinions were reflected in the final summary statements.)

[7] The review only evaluates the summative assessments offered by the testing programs, not the formative assessments.

[8] PARCC also offered paper-based test administration to schools and districts in Massachusetts (and, for that matter, nationally) as an alternative to its online version.

[9] Note that panelists were encouraged to use their professional judgment during all phases of the review. Some panels exercised this right freely while others chose to adhere more closely to the scoring guidance.

Next, the panels reviewed the ratings from grade five and grade eight and began to develop *program*-level ratings. At this point, they considered the rating that had been generated for the Criteria based on the program documentation (which helped to inform the generalizability of the findings). If the program documentation supported their item review findings, suggesting that the results would hold regardless of the number of forms reviewed,[10] the final criterion rating was based solely on the Group Rating, as noted above and explained in the methodology. If, however, the documentation *failed to provide* positive evidence that examination of additional forms would lead to the same rating, the panel determined whether to adjust the final criterion rating and, if so, stated the rationale. These ratings, which "roll up" the prior form- and grade-level ratings, were based on four levels of "match" to the Criteria: Excellent, Good, Limited/Uneven, and Weak Match. Lastly, the panels also developed their final "match" ratings relative to the Content and Depth of the assessments, which was the largest "grain size" upon which the tests were evaluated. In determining these Content and Depth ratings, reviewers were instructed by the methodology to weight certain criteria more strongly, though in the end the Content and Depth ratings were determined by professional judgment.[11] The review panels accompanied these final ratings with summary statements for each program regarding the observed strengths and aspects needing improvement, based on the Criteria.

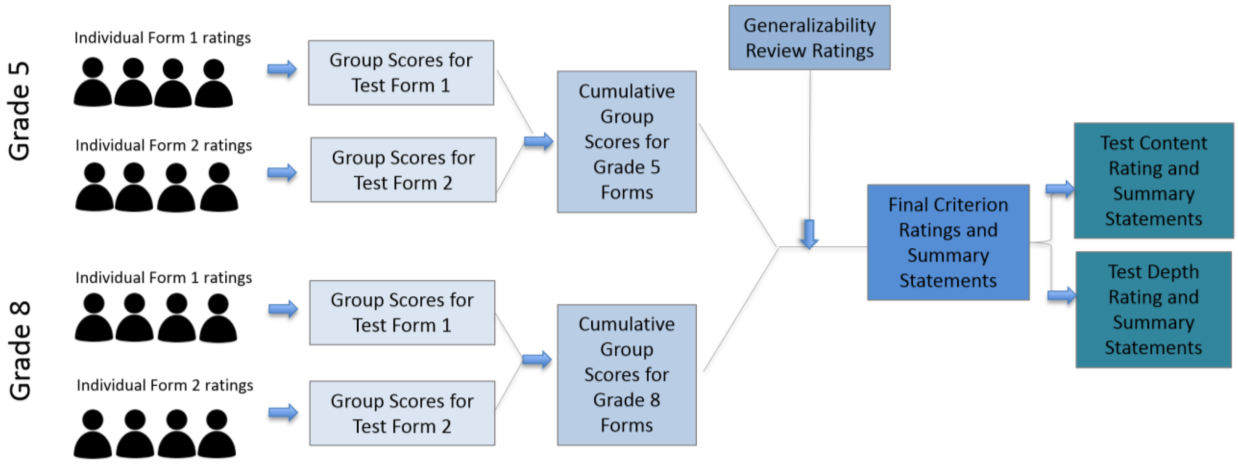
In sum, each consecutive step of the evaluation built upon the prior step, which fostered shared understanding within the panel and strengthened the internal consistency of the results. Below is a diagram showing how the "layers" of the evaluation ultimately helped to generate the final Content and Depth scores, as well as the summary statements.[12]

[10] The reference to more forms applies to forms developed in the same or different years that are based on the same test blueprints and specifications.

[11] Per the methodology, the prioritized criteria are B.3 and B.5 for ELA Content; B.1 and B.2 for ELA Depth; C.1 for Mathematics Content; and C.3 for Mathematics Depth.

[12] Unless otherwise noted in the summary statements, the ratings for the grade five and eight forms were largely consistent across grades and were ultimately combined at the "form" level.

The Process for Generating Final Program Ratings for ELA/Literacy and Mathematics


Grade 5
Individual Form 1 ratings
Group Scores for Test Form 1
Individual Form 2 ratings
Group Scores for Test Form 2
Cumulative Group Scores for Grade 5 Forms
Grade 8
Individual Form 1 ratings
Group Scores for Test Form 1
Individual Form 2 ratings
Group Scores for Test Form 2
Cumulative Group Scores for Grade 8 Forms
Generalizability Review Ratings
Final Criterion Ratings and Summary Statements
Test Content Rating and Summary Statements
Test Depth Rating and Summary Statements

Results

ELA/Literacy

Question 1a: Do the ELA/Literacy assessments place strong emphasis on the most important content of college and career readiness standards for the pertinent grade level, including the Common Core State Standards (CCSS)? (CCSSO Criteria B.3, B.5, B.6, B.7, and B.8)

This question is addressed through a review of five areas: reading, writing, language/vocabulary, research, and speaking/listening. **Criterion B.3** asks whether the reading items require close reading and direct textual evidence.[13] PARCC earned a score of "Excellent Match" on this Criterion, and MCAS earned a score of "Good Match." Reviewers noted that large proportions of PARCC reading items required close reading, focused on central ideas, and required direct textual evidence. The large majority of MCAS items also required close reading and focused on central ideas, but relatively few items required direct textual evidence. For example, at grade five, reviewers found that 92 percent of PARCC reading comprehension items required direct textual evidence, as compared to just 21 percent for MCAS. Reviewers also noted that MCAS items were less likely to be aligned to the specifics of the reading standards than PARCC items.

[13] These are defined in the CCSS Anchor Standards for Reading this way: "Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text."

Criterion B.5 asks whether the assessments emphasize writing tasks that require students to write to sources (draw on textual evidence after reading a text or passage) and assess narrative, expository, and persuasive/argument writing across each grade band (K–5, 6–12). Because the state of Massachusetts does not assess writing at grade five or eight—it only does so once within a grade band—the MCAS does not offer sufficient opportunity to assess the several types of writing called for in this criterion. It therefore earned a score of "Weak Match" on this criterion.[14] In contrast, the PARCC assessment writing prompts do require writing to sources, and documentation indicates that the assessment includes a variety of writing types within the grade band. It receives a score of "Excellent Match."

Criterion B.6 evaluates the degree to which the tests sufficiently emphasize the assessment of language conventions and vocabulary. On this criterion, PARCC earns a score of "Excellent Match," and MCAS earns a score of "Limited/Uneven Match." PARCC receives the highest match score because it includes adequate numbers of language and vocabulary items, and these items focus on common student errors and assess Tier 2 words (words commonly used in written texts, which the CCSS refer to as "general academic words"). MCAS fell short on this criterion in several areas. First, at grade five, the proportion of vocabulary items testing Tier 2 words and phrases was 67.5 percent, lower than the suggested goal of at least three-quarters in the scoring guidance. (The comparable PARCC score was 88 percent.) Second, MCAS language items were not evaluated within writing assessments that mirror real-world activities (including editing and revision).

Criterion B.7 asks whether students must demonstrate research and inquiry skills through tasks that require them to analyze, synthesize, and/or organize information from two or more sources. The MCAS earns a score of "Weak Match" on this criterion because it does not assess research tasks. The PARCC assessment earns a score of "Excellent Match," with reviewers noting that its research tasks require analysis, synthesis, and/or organization; each task meets the requisite thresholds for this criterion. Still, PARCC's research tasks could have been strengthened by requiring students to always use two or more sources.

Finally, **criterion B.8** addresses speaking and listening skills. Both assessments earned a score of "Weak Match" on this criterion because neither assesses speaking or listening at this time. The panels urged the programs to add assessments of these skills over time as technologies allow. Note that the methodology states that this criterion is to be met "over time, as assessment advances allow." As such, this rating is not included in the overall rating for Content.

Considering the five content criteria, the **overall ELA/Literacy Content rating** for PARCC is "Excellent Match," while the overall Content rating for MCAS is "Limited/Uneven Match." Reviewers found that PARCC assessed reading, writing, language/vocabulary, and research with excellence, falling short only in its lack of attention to speaking and listening. In contrast, while MCAS received a score of "Good Match" on the reading criterion, it earned a "Weak Match" on the other four criteria.

[14] Given that MCAS does evaluate writing at grades four and seven, we asked our panelists to review those writing items and evaluate them based on the associated CCSS grade-level standards. This exercise was conducted only to give the program information about its treatment of writing; it was not a part of the final scoring. Ultimately, both the writing prompts on the fourth- and seventh-grade assessments did not require writing to sources, which likely would have also resulted in a "Weak Match" score had those prompts been a part of the actual review.

Question 1b: Do the ELA/Literacy assessments require all students to demonstrate the higher-order thinking skills reflected in the standards? (CCSSO Criteria B.1, B.2, B.4, and B.9)

The Depth rating for the ELA/literacy assessments is based on four criteria: text type, text complexity, depth of knowledge, and item type/quality.[15]

Criterion B.1, which focuses on text type, calls for a balance of high-quality informational and literary texts. Both MCAS and PARCC earn a score of "Good Match." Reviewers noted that both assessments had a variety of high-quality texts. However, among the informational texts used, reviewers found that PARCC had insufficient focus on narrative informational texts (writing about an event or series of events using literary style), whereas the MCAS had an insufficient focus on expository texts (writing that explains or informs about a specific topic).

In the item review, we were unable to apply **criterion B.2**, which calls for increasing levels of text complexity across the grades, as our evaluation did not include consecutive grade levels. We had planned to evaluate program-supplied data relative to text complexity, but those data (both quantitative and qualitative) were inconsistent and extremely difficult to present uniformly to reviewers. Thus, our ratings for B.2 are based only on blueprints and other documentation. Both PARCC and MCAS earned a score of "Good Match" on this criterion because their respective documentation clearly and explicitly requires texts to increase in complexity grade-by-grade and to be placed in grade bands and grade levels based on appropriate quantitative and qualitative data.

Criterion B.4 focuses on the depth of knowledge (DOK) of the assessments. It stipulates that the distribution of cognitive demand on a test form should be comparable to, and at least as rigorous as, the distribution of the state's standards.[16] MCAS earned a score of "Limited/Uneven Match," with reviewers finding inadequate coverage of the higher levels of cognitive demand (DOK 3 and 4) set forth in the CCSS.[17] In contrast, PARCC earned a score of "Excellent Match," with reviewers noting that the DOK of the test met (or, for eighth grade, exceeded) the DOK of the standards. A supplementary DOK analysis is presented in Appendix B.

Finally, **criterion B.9** focuses on item type and quality. The methodology asks whether the program uses at least two item types, including at least one in which students *generate* rather

[15] Item quality is included within the Depth rating because reviewers must determine whether an item requires students to provide evidence of the targeted skill. Particularly in the case of higher-order skills, items that appear to address them may in fact not yield such evidence, but be answerable through use of lower-level skills. For example, criterion B.3 addresses "close reading" of texts, for which the criterion (and CCSS) require analysis of the information in the text(s) to support a conclusion, generalization, or inference—not simply recall what was read. Reviewers had to determine whether items aligned to B.3 required the complexity called for by the criterion.

[16] Fordham used Webb's Depth of Knowledge (DOK) taxonomy to classify cognitive demand and commissioned an analysis of the DOK of the standards at grades five and eight in advance of the study. The DOK analysts coded each standard as falling into one or more of the DOK levels. All standards were equally weighted, and the ratings for each standard were averaged to arrive at the final DOK distribution for the standards. For the assessments, reviewers evaluated the DOK of the items on each form using the same process. Thus, reviewers were allowed to assign items to multiple DOK levels. However, because reviewers did not have access to answer keys or scoring rubrics, they could not take into account scoring guidance, such as the allocation of partial credit points, in determining the DOK of items. Appendix B for more.

[17] DOK 3 items typically require generalization and inference across an entire passage or the identification of abstract themes and concepts. DOK 4 items typically require deep conceptual understanding and generalization across multiple passages or sources. See Appendix B for more.

than *select* a response, and whether the items are of high editorial and technical quality. Both assessments earn a score of "Excellent Match" on this criterion. Both utilize multiple item types, though the paper-and-pencil nature of the MCAS does not allow for technology-enhanced items. Both assessments were also rated highly on item quality, with no major concerns noted.

Considering the four depth criteria, the **overall ELA/Literacy Depth rating** for PARCC is "Excellent Match," while the overall Depth rating for MCAS is "Good Match." The primary difference between PARCC and MCAS was for B.4 (depth of knowledge), for which PARCC earned "Excellent Match" and MCAS earned "Limited/Uneven Match." On the other three depth criteria, both assessments earned "Good Match."

Mathematics

Question 1a: Do the mathematics assessments place strong emphasis on the most important content of the Common Core State Standards (CCSS) for the grade level? (CCSSO Criteria C.1 and C.2)

To evaluate the degree to which the assessments place strong emphasis on the most important content of the CCSS math standards for the applicable grade level, the panel examined the degree of focus on the "major work of the grade"—those grade-level standards identified as most essential to keeping students on track for college and career readiness (**criterion C.1).** For example, the critical areas in grade five mathematics, according to the standards, include understanding the place value system and using equivalent fractions as a strategy to add and subtract fractions. Reviewers determined whether or not individual items were assessing the grade's major work topics, and "focus" was calculated based on the percentage of score points derived from these items. On this criterion, PARCC earned a score of "Good Match," and MCAS earned a score of "Limited/Uneven Match." At both grades, a greater proportion of PARCC's score points were allocated for items primarily assessing the major work of the grade. At fifth grade, only 52 percent of MCAS score points addressed these critical areas, which is significantly below the goal of at least three-quarters per the scoring guidance. In comparison, 72 percent of PARCC score points addressed the major work. At eighth grade, both assessments met the criterion, with 78 percent of MCAS score points and 81 percent of PARCC score points addressing the major work of the grade.

Reviewers also considered the distribution of items across three categories: procedural fluency, conceptual understanding, and application, with a goal of approximately equal distribution **(criterion C.2).** However, due to variations in how reviewers understood and implemented this criterion, we only report the panel's qualitative impressions.[18] In general, PARCC was found to have a good balance across these three item types. But reviewers sometimes found that the application items on PARCC had superficial or irrelevant contexts, meaning they were not necessary or important to the problem (and potentially distracting). As for MCAS, reviewers noted an imbalance at grade five; the test was found to focus more on

[18] For instance, the training was not sufficiently clear that items invoking a "trivial" context ought not be coded as Application. Further, the distinction between trivial and meaningful context is often subtle, and the methodology did not allow reviewers to choose two item types. This resulted in many items that also assessed Conceptual Understanding and/or Procedural Skill/Fluency being categorized as only Application, even if the context was presumably trivial.

application and procedural skills, with a deficit of problems addressing conceptual understanding.

Because we did not include C.2 in our results, the overall Content rating was the same as the C.1 score. Thus, for Content, PARCC earned a score of “Good Match” to the CCSSO criteria because of its greater emphasis on the major work of the grade, while MCAS earned a score of “Limited/Uneven Match” because it had inadequate focus on the major work at grade five (rather, it samples content across the full set of standards for the grade).

Question 1b: Do the assessments require all students to demonstrate the higher-order thinking skills called for in the standards? (CCSSO Criteria C.3, C.4, and C.5)

The mathematics rating for Depth is based on three criteria: connecting mathematical practices to content;[19] requiring a range of cognitive demand; and item type/quality. **Criterion C.3** connects mathematical practices to content, focusing on the extent to which test items measuring the standards for mathematical practice also connect to content standards. Reviewers rated both assessments “Excellent Match” on this criterion. While MCAS does not code test items to the mathematics practice standards—meaning that items are not linked to the practice standards in the MCAS program documentation or in their underlying item metadata—reviewers nonetheless believed that many items did in fact assess them and that all items that did so also assessed content standards. PARCC did code items to the standards for mathematical practices, and all of these items also assessed content standards.

With regard to the cognitive demand/depth of knowledge (**criterion C.4**), the same process was used as described above for ELA/literacy. MCAS earned a score of “Excellent Match” and PARCC earned a score of “Good Match.” The MCAS analysis found that the assessment at both grades matched the distribution of DOK of the math standards quite well. For PARCC, the fifth-grade assessment also matched the standards’ DOK well. At eighth grade, however, PARCC focused more on higher levels of DOK and less on lower levels relative to the standards; thus, reviewers were concerned that lower-level CCSS standards might not be adequately assessed.

Astute readers will observe that the ELA and math panels arrived at different conclusions regarding how to treat PARCC, which exceeded the DOK of the standards at grade eight in both subjects. The guidance recommended a DOK index of .80 to receive a score of “Excellent Match” (i.e., 80 percent agreement on the cognitive demand emphasis), which was not met for that grade in either subject. The math panelists found that DOK 1 was under-assessed (see above), which was not an issue for ELA. Further, the panelists were encouraged to use professional judgment in arriving at their final ratings. The mathematics panel tended to follow the letter of the guidance more directly. This propensity, coupled with the DOK-1 finding, resulted in PARCC receiving a lower score for criterion C.4. In contrast, the ELA reviewers judged the DOK distribution to be appropriate, and even desirable, so they adjusted the score to “Excellent Match.”

[19] Mathematical practices refer to the different varieties of expertise that mathematics educators seek to develop in their students (for example, abstract reasoning, constructing and critiquing arguments, modeling); these cut across grade levels. In contrast, content standards refer to the specific knowledge and skills that students are supposed to master at each grade level.

Finally, with regard to item type and quality (**criterion C.5**), the methodology asks whether the program uses at least two item types, including at least one in which students generate rather than select a response, and whether the items are of high editorial and technical quality. MCAS earned a score of "Excellent Match" and PARCC a score of "Good Match." Both assessments included a variety of item types, including both selected response and constructed response items. Reviewers were very satisfied with the editorial and mathematical quality of items on MCAS, noting only very minor problems. For PARCC, reviewers noticed somewhat more problems with item quality. Most of these were editorial, but some mathematical issues were also noted.

Considering all three depth criteria, MCAS earned a score of "Excellent Match" for depth, while PARCC earned a score of "Good Match." Reviewers noted only minor issues relating to depth for MCAS, particularly around coding items to math practices and making explicit the connections between math practices and content. For PARCC, reviewers were most concerned with item quality, noting some major and some minor concerns.

See Table 1 for the full ratings and summary statements for PARCC and MCAS, grades five and eight.

Accessibility

CCSSO criterion A.5 calls for assessment programs to "provid[e] accessibility to all students, including English learners and students with disabilities." Note that HumRRO and Fordham jointly recruited a panel to evaluate accessibility, so the results below apply for grade five, grade eight, and high school, as well as across the ELA and mathematics subject areas.

Panel Recruitment and Training

The evaluation of test accessibility for English learners (ELs) and students with disabilities (SWDs) requires specialized expertise. HumRRO and Fordham jointly recruited a separate panel of educators and researchers with expertise in ELs, SWDs and Universal Design for Learning (UDL) and who are also familiar with the Common Core State Standards (CCSS) in either English language arts/literacy (ELA) or mathematics. As was the case with all reviewers, the prospective accessibility panelists completed a questionnaire detailing their expertise, knowledge of the CCSS and the assessments to be reviewed, and possible conflicts of interest. Applications were reviewed, individuals with conflicts were removed from consideration, and the final panel was composed to ensure representation across the different areas of expertise and perspectives. Nine qualified reviewers were distributed across the various areas of expertise required for the study (including math and English language arts, ELs, SWDs, and UDL).

All panelists participated in online and in-person training and a calibration activity, provided by HumRRO, prior to conducting reviews of the online assessments. Participants were trained on the factors to consider when assigning their ratings as well as the procedures for doing so. Reviewers were monitored as they completed their assigned activities, and additional training and guidance was provided by HumRRO staff as needed.

Review of Documentation and Exemplars

Reviewers evaluated each of the scoring components under criterion A.5 and used the scoring rubrics and guidance to determine the tentative score to assign for ELs and then for SWDs. The rubrics included four possible categories/scores: Meets/2; Partially Meets/1; Does Not Meet/0, and Insufficient Evidence/0. If reviewers disagreed with a tentative score, they could override it, provided they gave a rationale.

In addition to program documentation, reviewers evaluated exemplar items and item meta-data provided by the vendors. The item meta-data included the content standard, depth of knowledge level, and other data that reviewers could use to understand what the item was measuring. For MCAS, the item exemplars were physical artifacts and included an American Sign Language compact disc, text-to-speech DVD, mathematics Braille manipulatives (block and ruler), a Spanish test version, and a large print test version. A sample of the items provided for the ELA and math exemplars was reviewed for construct integrity (whether the identified standard was consistent across different accommodated conditions) and ease of use for ELs and for SWDs separately (if appropriate). The PARCC exemplars provided reviewers the opportunity to see operational sets of items under certain accommodated conditions, such as use of a screen-reader (which converts text into computerized speech) and text-to-speech (which describes tables, figures, and illustrations). Reviewers determined whether the exemplars provided suitable evidence that what had been described in the program documentation would be implemented with consistency.

Results

Question 2: Are the assessments accessible to all students, including students with disabilities and English language learners? (Accessibility) (CCSSO criterion A.5)

The panelists found that a great deal of thought has been give across both programs to accessibility features and allowable accommodations that enable English learners and students with disabilities to be assessed using the same instrument as their peers. PARCC, as the newer, computer-based assessment, tended to use the most recent knowledge available for developing accessible items, embedding accessibility tools and features, and implementing student accommodations—as opposed to retrofitting items and tools developed for the general population. When MCAS was first developed, it was considered a state-of-the-art assessment. It remains a rigorous state assessment; however, because it is paper-and-pencil and was not developed using evidence-centered design, it received a “Weak Match” on the Criteria. PARCC’s accessibility was judged as a “Good Match.” The online nature of the assessment facilitates the inclusion of features for accessibility and universal design. See Table 2 for more.

Program Strengths and Weaknesses

Question 4: What are the overall strengths and weaknesses of each assessment relative to the CCSSO criteria?

The PARCC assessments received final ratings of “Good Match” or “Excellent Match” across all but one of the research questions regarding Content and Depth of the ELA/literacy and mathematics assessments, as well as accessibility for English learners and students with disabilities. Overall, then, our study finds the PARCC tests to be high-quality assessments of college and career readiness as judged against the CCSSO Criteria evaluated in this study.

PARCC tests pay close attention to the prioritized skills called for in the CCSS, including close reading, writing to sources, and the critical mathematics topics of each grade. One area that is a strength but also a possible weakness for PARCC is the heavy emphasis on higher-order skills: the degree of emphasis is significantly greater than found in the Common Core standards themselves and in some well-respected international assessments, particularly in ELA/literacy and grade-eight mathematics (see Appendix B). This attention to higher-demand items allows students to demonstrate (or not) strong understanding of the standard’s more complex skills. At the same time, attention to these more complex skills could result in inadequate measurement of lower-level foundational skills.

The 2014 MCAS, while a highly respected and long-standing assessment program with items of high editorial and technical quality, fell short of several of the criteria for high-quality assessments of college and career readiness. Overall, half of the final scores were “Good Match” or “Excellent Match,” and half were “Limited/Uneven Match” or “Weak Match.”

In ELA/literacy, this paper-and-pencil assessment infrequently assesses writing and does not assess writing to sources or research skills. Also, its emphasis on higher-order skills, such as analysis and synthesis, is inadequate, although this is likely due in part to the very limited assessment of writing on MCAS. The quality of the texts and assessment of close reading, however, are strengths.

In mathematics, the MCAS tests accurately reflect the distribution of cognitive demand in the CCSS, but the grade-five tests show inadequate focus on the major work of the grade, and too few items address conceptual understanding. (There were multiple items on non-critical areas like the use of parentheses, brackets, or braces in numerical expressions and converting different standard measurement units in solving multi-step, real world problems.)

For students with disabilities and English language learners, the accessibility of the MCAS assessments is rated a “Weak Match,” largely due to the limitations of a paper-based assessment, whereas the PARCC assessment has numerous accessibility tools and features embedded within the delivery system.

In the following pages, the expert review panels provide their ratings and their summary statements. The statements should be read carefully, as they provide additional guidance to those considering use of either assessment.

KEY: Degree of Match to CCSSO Criteria

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

Table 1. Final English Language Arts/Literacy Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

<table>
<tr><th colspan="5">ENGLISH LANGUAGE ARTS/LITERARY</th></tr>
<tr><th colspan="5">I. Assesses the content most needed for College and Career Readiness.</th></tr>
<tr><td colspan="2">MCAS receives a <b>Limited/Uneven Match</b> 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.</td><td>L</td><td>PARCC receives an <b>Excellent Match</b> 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 content as reflected in college- and career-ready 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.</td><td>E</td></tr>
<tr><th>Content Sub-Criteria</th><th>MCAS Rating</th><th>MCAS Summary Statement</th><th>PARCC Rating</th><th>PARCC Summary Statement</th></tr>
<tr><td><b>B.3 Reading:</b> Tests require students to read closely and use specific evidence from texts to obtain and defend correct responses.</td><td>G</td><td>On “requiring students to read closely and use evidence from texts,” the rating is <b>Good Match.</b> Most reading items require close reading and focus on central ideas and important particulars. Some questions, however, do not require the students to provide direct textual evidence to support their responses. In addition, too many items do not align closely to the specifics of the standards.</td><td>E</td><td>On “requiring students to read closely and use evidence from texts,” the rating is <b>Excellent Match.</b> Nearly all reading items require close reading, the understanding of central ideas, and the use of direct textual evidence.</td></tr>
<tr><td><b>B.5 Writing:</b> Tasks require students to engage in close reading and analysis of texts. Across each grade band, tests include a balance of expository, persuasive/argument, and narrative writing.</td><td>W</td><td>On “assessing writing,” the rating is <b>Weak Match</b>. Writing is assessed at only one grade level per band, and there is insufficient opportunity to assess writing of multiple types. In addition, the writing assessments do not require students to use sources. As a result, the program inadequately assesses the types of writing required by college and career readiness standards.</td><td>E</td><td>On “assessing writing,” the rating is <b>Excellent Match</b>. The assessment meets the writing criterion, which requires writing to sources. Program documentation shows that a balance of all three writing types is required across each grade band.</td></tr>
</table>

Table 1. Final English Language Arts/Literacy Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

Content Sub-Criteria	MCAS Rating	MCAS Summary Statement	PARCC Rating	PARCC Summary Statement
B.6 Vocabulary and language skills: Tests place sufficient emphasis on academic vocabulary and language conventions as used in real-world activities.	L	On "emphasizing vocabulary and language skills," the rating is **Limited/Uneven Match**. Vocabulary items are sufficient and generally aligned to the criterion; however, the grade five items need more words at the Tier 2 level. Furthermore, a lack of program documentation means that the quality of vocabulary assessments cannot be substantiated across forms. MCAS does not meet the criterion for assessing language skills, which call for them to be assessed within writing assessments that mirror real-world activities including editing and revision.	E	On "emphasizing vocabulary and language skills," the rating is **Excellent Match**. The test contains an adequate number of high-quality items for both language use and Tier 2 vocabulary and awards sufficient score points, according to the program's documentation, to both of these areas.
B.7 Research and inquiry: Assessments require students to demonstrate the ability to find, process, synthesize and organize information from multiple sources.	W	On "assessing research and inquiry", the rating is **Weak Match**. The assessment has no test questions devoted to research.	E	On "assessing research and inquiry," the rating is **Excellent Match**. The research items require analysis, synthesis, and/or organization and the use of multiple sources, therefore meeting the criterion for Excellent.
B.8 Speaking and listening: Over time, and as assessment advances allow, the assessments measure speaking and listening communication skills.	W	On assessing "speaking and listening," the rating is **Weak Match**. The program does not assess speaking or listening at this time. Because this criterion is to be met "over time, as assessment advances allow," **this rating is not included in the overall rating for Content.**	W	On assessing "speaking and listening," the rating is **Weak Match.** The program does not assess speaking or listening at this time. Because this criterion is to be met "over time, as assessment advances allow," **this rating is not included in the overall rating for Content.**

Table 1. Final English Language Arts/Literacy Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

II. Assesses the depth that reflects the demands of College and Career Readiness.			
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.	G	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.	E

Depth Sub-Criteria	MCAS Rating	MCAS Summary Statement	PARCC Rating	PARCC Summary Statement
B.1 Text quality and types: Tests include an aligned balance of high-quality literary and informational texts.	G	On "the balance of high-quality literary and informational texts," the rating is **Good Match**. The quality of the texts is very high. Regarding the balance of text types, some forms had too few informational texts.	G	On "the balance of high-quality literary and informational texts," the rating is **Good Match**. Although the passages are consistently of high quality, the tests would have better reflected the criterion with additional literary nonfiction passages.
B.2 Complexity of texts: Test passages are at appropriate levels of text complexity, increasing through the grades, and multiple forms of authentic, high-quality texts are used.	G	On "use of appropriate levels of text complexity," the rating is **Good Match**. It is based solely on the review of program documentation, which is determined to have met the criterion. The test blueprints and other documents clearly and explicitly require texts to increase in complexity grade by grade and to be placed in grade bands and grade levels based on appropriate quantitative and qualitative data.	G	On "use of appropriate levels of text complexity," the rating is **Good Match.** It is based solely on the review of program documentation, which is determined to have met the criterion. The test blueprints and other documents clearly and explicitly require texts to increase in complexity grade by grade and to be placed in grade bands and grade levels based on appropriate quantitative and qualitative data.
B.4 Cognitive demand: The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.	L	On "requiring a range of cognitive demand," the rating is **Limited/Uneven Match**. More items that measure the higher levels of cognitive demand are needed to sufficiently assess the depth and complexity of the standards.	E	On "requiring a range of cognitive demand," the rating is **Excellent Match**. The test is challenging overall; indeed, the cognitive demand of the grade eight test exceeds that of the CCSS.

Table 1. Final English Language Arts/Literacy Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

<table>
<tr><th>Depth Sub-Criteria</th><th>MCAS Rating</th><th>MCAS Summary Statement</th><th>PARCC Rating</th><th>PARCC Summary Statement</th></tr>
<tr><td><b>B.9 High-quality items and variety of item types:</b> Items are of high technical and editorial quality and each test form includes at least two items types including at least one that requires students to generate rather than select a response.</td><td>E</td><td>On “ensuring high-quality items and a variety of item types,” the rating is <b>Excellent Match</b>. Multiple item formats are used, including student-generated response items. The items exhibit high technical quality and editorial accuracy. The paper-and-pencil format precludes the use of technology-enhanced items, but the criterion for multiple item types is met.</td><td>E</td><td>On “ensuring high quality items and a variety of item types,” the rating is <b>Excellent Match</b>. The tests use multiple item formats, including student-constructed responses.</td></tr>
<tr><th colspan="3">MCAS ELA/Literacy Overall Summary</th><th colspan="2">PARCC ELA/Literacy Overall Summary</th></tr>
<tr><td colspan="3">In ELA/literacy, MCAS receives a limited-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.</td><td colspan="2">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.</td></tr>
</table>

Table 2. Final Mathematics Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

MATHEMATICS

I. Assesses the content most needed for College and Career Readiness.

MCAS provides a **Limited/Uneven Match** to the CCSSO Criteria for Content in Mathematics. While the grade eight assessment focuses strongly on the major work of the grade, the grade five 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 five test.	L	PARCC provides 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 five.	G

Content Sub-Criteria	MCAS Rating	MCAS Summary Statement	PARCC Rating	PARCC Summary Statement
C.1 Focus: Tests focus strongly on the content most needed in each grade or course for success in later mathematics (i.e. Major Work).	L	On "focusing strongly on the content most needed for success in later mathematics," the rating is **Limited/Uneven Match.** The grade eight assessment is focused on the major work of the grade. The grade five assessment is significantly less focused on the major work of the grade than called for by the criterion, as it samples content across the full set of standards for the grade.	G	On "focusing strongly on the content most needed for success in later mathematics," the rating is **Good Match.** While the grade eight tests focus strongly on the major work of the grade, the grade five tests fall barely short of the threshold required for the top rating.
C.2: Concepts, procedures, and applications: Assessments place balanced emphasis on the measurement of conceptual understanding, fluency and procedural skill, and the application of mathematics.*	--	(Qualitative summary only) The test forms contain items that assess conceptual understanding, procedural fluency, and application. Particularly in fifth grade, however, test forms are overly focused on procedural skill/fluency and application relative to conceptual understanding.	--	(Qualitative summary only) The test forms contain items that assess conceptual understanding, procedural fluency, and application. Some of the application problems, however, have shallow contexts that are not necessary or important to the problem.

**The program requires, in their program documentation, the assessment of conceptual understanding, procedural skill/ fluency, and application, although most do not clearly distinguish between procedural skill/fluency and conceptual understanding. Also, specific balance across these three types is not required. Due to variation across reviewers in how this criterion was understood and implemented, final ratings could not be determined with confidence. Therefore, only qualitative observations are provided.*

Table 2. Final Mathematics Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

II. Assesses the depth that reflects the demands of College and Career Readiness.			
MCAS provides 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 specific content standards.	E	PARCC provides a **Good Match** to the CCSSO Criteria for Depth in Mathematics. The tests include items with a range of cognitive demand, but at grade eight, 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 provides sufficient information across the range.	G

Depth Sub-Criteria	MCAS Rating	MCAS Summary Statement	PARCC Rating	PARCC Summary Statement
C.3 Connecting practice to content: Test questions meaningfully connect mathematical practices and processes with mathematical content.	E	On "connecting practice to content," the rating is **Excellent Match.** Although no items are coded to mathematical practices, the practices were assessed within items that also assessed content.	E	On "connecting practice to content" the rating is **Excellent Match.** All items that are coded to mathematics practices are also coded to one or more content standard.
C.4 Cognitive demand: The distribution of cognitive demand for each grade level is sufficient to assess the depth and complexity of the standards.	E	On "requiring a range of cognitive demand" the rating is **Excellent Match.** At each grade level, the distribution of cognitive demand closely reflects that of the standards.	G	On "requiring a range of cognitive demand" the rating is **Good Match.** The distribution of cognitive demand of items reflects that of the standards very well at grade five, while the grade eight test includes proportionately more items at the higher levels of cognitive demand (DOK 2 and 3). As a result, grade eight standards that call for the lowest level of cognitive demand may be under-assessed.

Table 2. Final Mathematics Ratings and Summary Statements, Grades Five and Eight (MCAS and PARCC)

Depth Sub-Criteria	MCAS Rating	MCAS Summary Statement	PARCC Rating	PARCC Summary Statement
C.5 High-quality items and variety of item types: Items are of high technical and editorial quality and are aligned to the standards, and each test form includes at least two item types, including at least one that requires students to generate rather than select a response.	E	On "ensuring high-quality items and a variety of item types," the rating is **Excellent Match.** Both grade five and grade eight forms include multiple item types, including constructed-response. The items are of high technical and editorial quality, with very minor issues of editing, language, and accuracy at grade eight.	G	On "ensuring high-quality items and a variety of item types" the rating is **Good Match.** The program includes a wide variety of item types, including several that require student-constructed responses. However, there are a number of items with quality issues, mostly minor editorial but sometimes mathematical.

MCAS Mathematics Overall Summary	PARCC Mathematics Overall Summary
In mathematics, MCAS receives a limited 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 five 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 five and identification of the mathematical practices that are assessed—and their connections to content.	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 five, the test includes a distribution of cognitive demand that is similar to that of the standards. At grade eight, 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 eight PARCC assessment will have demonstrated strong understanding of the standard's more complex skills. However, the grade eight 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 eight that assess standards at DOK 1, and increased attention to accuracy of the items—primarily editorial, but in some instances mathematical.

Table 3. Final Ratings and Summary Statements for Accessibility in ELA/Mathematics (Grades Five, Eight, and high school)

ACCESSIBILITY			
I. Providing accessibility to all students, including English learners and students with disabilities.			
MCAS Summary Statement		**PARCC Summary Statement**	
For students with disabilities, the MCAS documentation indicates an attempt to provide accessibility for this paper-based test. However, some of the procedures around implementation, communication, and quality of exemplars are lacking. For example, there is little documentation detailing how accommodations should be assigned or the potential impact of using multiple accommodations concurrently. Also, additional documentation is needed to show how data and feedback would be used to improve accessibility and future test items. The accessibility features/accommodations provided to English Learners taking this paper-based test are much narrower than the range of research-based supports available at this time. Attempts to provide accessibility to English learners, if referenced, are inconsistently applied across the testing program.	W	The assessment succeeds at pushing the framework for traditionally identified supports for English Learners and Students with Disabilities. For Students with Disabilities, PARCC was rated highly for sensitivity to item design that reflects the individual needs of students. However, little attention is paid to disability categories, how multiple features can be administered at once, or the implications of how multiple accessibility features impact students' performance. For English Learners, PARCC is research based and is inclusive of existing research to the extent possible. The PARCC documentation indicates that additional research will be conducted. PARCC proposes a multi-tiered approach to accessibility with improved guidelines, which is viewed as positive.	G

Appendix A: The CCSSO Criteria for Procuring and Evaluating High-Quality Assessments, 2014 (Only bolded criteria are included in this evaluation.)

A. Meet Overall Assessment Goals and Ensure Technical Quality

A.1 Indicating progress toward college and career readiness

A.2 Ensuring that assessments are valid for required and intended purposes

A.3 Ensuring that assessments are reliable

A.4 Ensuring that assessments are designed and implemented to yield valid and consistent test score interpretations within and across years

A.5 Providing accessibility to all students, including English learners and students with disabilities

A.6 Ensuring transparency of test design and expectations

A.7 Meeting all requirements for data privacy and ownership

B. Align to Standards – English Language Arts/Literacy

B.1 Assessing student reading and writing achievement in both ELA and literacy

B.2 Focusing on complexity of texts

B.3 Requiring students to read closely and use evidence from texts

B.4 Requiring a range of cognitive demand

B.5 Assessing writing

B.6 Emphasizing vocabulary and language skills

B.7 Assessing research and inquiry

B.8 Assessing speaking and listening

B.9 Ensuring high-quality items and a variety of item types

C. Align to Standards – Mathematics

C.1 Focusing strongly on the content most needed for success in later mathematics

C.2 Assessing a balance of concepts, procedures, and applications

C.3 Connecting practice to content

C.4 Requiring a range of cognitive demand

C.5 Ensuring high-quality items and a variety of item types

D. Yield Valuable Reports on Student Progress and Performance

D.1 Focusing on student achievement and progress to readiness

D.2 Providing timely data that inform instruction

E. Adhere to Best Practices in Test Administration

E.1 Maintaining necessary standardization and ensuring test security

F. State Specific Criteria (as desired)

Sample criteria might include

• Requiring involvement of the state’s K-12 educators and institutions of higher education

• Procuring a system of aligned assessments, including diagnostic and interim assessments

• Ensuring interoperability of computer-administered items

Appendix B: Depth of Knowledge (DOK) of MCAS and PARCC as Compared to Common Core Standards and Other Policy-Relevant Assessments

We also conducted a supplementary analysis of the depth of knowledge (DOK) of the MCAS and PARCC assessments against the Common Core standards. HumRRO had previously conducted a DOK analysis of the CCSS, so we started with those results. Next, we contracted with one subject-area expert in each subject to independently code the standards in terms of DOK. Finally, we contracted with another content-area expert in each subject to adjudicate in any instance where our reviewers disagreed with the HumRRO reviewers. In each instance, reviewers were allowed to place each standard into one or more DOK levels, and the standards were equally weighted to arrive at the final DOK distribution for the standards, as is common in alignment studies (e.g., Porter, 2002).[20] To help contextualize the findings, we compared the DOK of MCAS, PARCC, and the Common Core against an analysis of national and international assessments previously conducted by the RAND Corporation in 2014.[21] The results of this analysis are shown below.

In broad terms, depth of knowledge refers to the cognitive demands required by a task to produce an acceptable response. In each subject, there are four levels of DOK. Generally, level 1 refers to rote or reproductive skills (e.g., identifying an obvious detail in a text, conducting a straightforward one-step operation in mathematics). Level 2 refers to skills and concepts such as multi-step operations in mathematics and comprehension across one or more sentences. Level 3 refers to strategic thinking, such as solving a mathematics problem that has multiple possible approaches or identifying complex themes across an entire passage. Level 4 refers to extended thinking, such as extended mathematical investigations or synthesis and analysis across multiple texts.

The results in mathematics, shown in Table B.1, show that the PARCC assessment has somewhat higher levels of cognitive demand than the MCAS assessment. In fifth grade, for instance, the PARCC assessment has 11 percent of score points at DOK 3, whereas MCAS has 2 percent (and the Common Core has 7 percent of its content at DOK 3). The difference is starker in eighth grade, where the PARCC assessment has about 10 percent more DOK 3 score points and about 27 percent fewer DOK 1 score points. The DOK of the PARCC eighth-grade math test is among the highest of the assessments studied by Yuan and Le for RAND, far exceeding AP and PISA, for instance. However, the eighth-grade MCAS test's DOK distribution is more in line with the DOK distribution reflected in the Common Core.

The results in ELA, shown in Table B.2, also show that the DOK of PARCC assessments exceed those of MCAS at both grade levels. This difference is large in both grades, with 51 percent of PARCC fifth-grade score points at DOK 3 as compared to 27 percent of MCAS fifth-grade score points. Neither fifth-grade exam includes any DOK 4 score points, though 8 percent of the standards content at that grade is on DOK 4. The gap between PARCC and MCAS is even larger at eighth grade, however, with 67 percent of score points on PARCC at DOK 3+ and 36 percent of MCAS score points on DOK 3+. Again, the MCAS distribution, while generally including lower DOK than PARCC at eighth grade, is more aligned with the DOK reflected in the Common Core.

[20] Porter, A. C. (2002). Measuring the content of instruction: uses in research and practice. *Educational Researcher, 31*(7), 3-14.

[21] Yuan, L., & Le, V. (2014). *Measuring deeper learning through cognitively demanding test items: Results from the analysis of six national and international exams.* Santa Monica, CA: RAND.

Note that a 2012 study by RAND Corporation evaluated the cognitive demand of items that had been used on seventeen states' math and reading assessments, grades 3–8 and high school. The specific states were selected because prior literature suggested that their "state achievement tests were more cognitively demanding" than those of other states. Across this select group, RAND found: 1) In mathematics, fewer than 2 percent of items were at DOK 3 and none were at DOK 4; and 2) In reading, about 20 percent of items were at DOK 3 and fewer than 2 percent at DOK 4.[22]

Table B.1. Depth of knowledge of mathematics assessments

Math	DOK 1	DOK 2	DOK 3	DOK 4
MCAS 5	40.3%	57.7%	2.0%	0.0%
PARCC 5	34.4%	54.5%	11.1%	0.0%
CCSS 5	43.3%	49.6%	7.1%	0.0%
MCAS 8	40.2%	45.8%	14.0%	0.0%
PARCC 8	13.3%	62.0%	24.2%	0.5%
CCSS 8	50.9%	39.8%	9.3%	0.0%
AP	30.7%	54.2%	14.0%	1.0%
IB	21.0%	50.0%	28.0%	1.0%
NAEP	38.6%	54.0%	6.7%	0.6%
PISA	32.8%	50.4%	16.8%	0.0%
TIMSS	52.0%	46.5%	1.5%	0.0%

Note. Results for AP, IB, NAEP, PISA, TIMSS from Yuan and Le (2014)—more detail on which assessments were analyzed is available there. PARCC and MCAS results are based on percentages of score points, whereas other assessment results are based on percentages of items.

[22] The items were gleaned from tests administered between 2007 and 2010, with the exception of KY which was administered in 1998-1999; only publicly released items were evaluated. See: Yuan, Kun and Vi-Nhuan Le, "Estimating the Percentage of Students Who Were Tested on Cognitively Demanding Items Through the State Achievement Tests" (Santa Monica, CA: RAND Corporation), 2012, http://www.rand.org/pubs/working_papers/WR967.

Table B.2. Depth of knowledge of ELA assessments

Reading	DOK 1	DOK 2	DOK 3	DOK 4
MCAS 5	9.6%	63.5%	26.9%	0.0%
PARCC 5	4.5%	45.0%	50.5%	0.0%
CCSS 5	17.6%	36.8%	37.5%	8.1%
MCAS 8	4.8%	58.7%	33.7%	2.9%
PARCC 8	1.6%	29.1%	46.4%	22.9%
CCSS 8	10.1%	44.2%	41.8%	3.8%
AP	11.0%	33.0%	56.0%	0.0%
NAEP	20.4%	45.2%	33.3%	1.1%
PISA	37.3%	26.2%	36.5%	0.0%
PIRLS	55.2%	17.8%	26.1%	0.9%

Note. Results for AP, NAEP, PISA, PIRLS from Yuan and Le (2014)—more detail on which assessments were analyzed is available there. IB results not provided because IB does not have a reading assessment. AP, NAEP, PISA, and PIRLS results are for reading assessments, whereas MCAS, PARCC, and CCSS results are for combined ELA assessments. PARCC and MCAS results are based on percentages of score points, whereas other assessment results are based on percentages of items.

Acknowledgments

The Fordham Institute extends sincere thanks to the staff at the Massachusetts Department of Elementary and Secondary Education and to PARCC for their significant contributions to this study to date. From providing access to operational test forms, program documentation, and materials, to designing and delivering initial training on the MCAS and PARCC test design to the panelists, to fielding countless reviewer questions throughout the review process, we greatly appreciate all of your time and effort.

Author Biographies

Nancy Doorey, project manager and report co-author, has been deeply involved in educational reform for more than thirty years, serving as a teacher, state and district policymaker, program director, and consultant in the areas of assessment, teacher quality, and leadership. She has authored reports for several national organizations regarding advances in educational assessment, the six federally funded assessment consortia, and education governance. In 2009, Nancy co-led the creation of the Center for K-12 Assessment & Performance Management at ETS, which was charged with serving as a catalyst for advances in K–12 testing to support student learning, and authored its widely utilized series “Coming Together to Raise Achievement: New Assessments for the Common Core State Standards.” As the director of programs, she formulated the agendas and managed five national research symposia and six webcasts regarding advances and challenges in K–12 assessment. Nancy received a master’s degree in elementary education from Simmons College and a Certificate of Advanced Studies in computer science from Harvard University Extension; she also completed doctoral studies in educational leadership at Columbia University (ABD).

Morgan Polikoff, alignment expert and report co-author, is an assistant professor of education at the University of Southern California’s Rossier School of Education. His areas of expertise include K–12 education policy; college and career-ready standards; assessment policy; alignment; and the measurement of classroom instruction. Recent work has investigated teachers' instructional responses to content standards and critiqued the design of school and teacher accountability systems. Ongoing work focuses on the implementation of college and career-ready standards and the influence of curriculum materials and assessments on implementation. He is an associate editor of *American Educational Research Journal* and serves on the editorial boards for AERA Open and *Educational Administration Quarterly*. His research is currently supported by the National Science Foundation, the WT Grant Foundation, and the Institute of Education Sciences, among other sources. Polikoff received his doctorate from the University of Pennsylvania’s Graduate School of Education in 2010 with a focus on education policy and his bachelor’s in mathematics and secondary education from the University of Illinois at Urbana-Champaign in 2006.