Foreword

By Amber M. Northern and Michael J. Petrilli

We at the Thomas B. Fordham Institute have been evaluating the quality of state academic standards for nearly twenty years. **Our very first study**, published in the summer of 1997, was an appraisal of state English standards by Sandra Stotsky. Over the last two decades, we've regularly reviewed and reported on the quality of state K–12 standards for **mathematics**, **science**, **U.S. history**, **world history**, **English language arts**, and **geography**, as well as the **Common Core**, **International Baccalaureate**, **Advanced Placement** and other influential standards and frameworks (such as those used by **PISA, TIMSS, and NAEP**). In fact, evaluating academic standards is probably what we're best known for.

For most of those two decades, we've also dreamed of evaluating the tests linked to those standards—mindful, of course, that in most places the tests are the real standards. They're what schools (and sometimes teachers and students) are held accountable to and they tend to drive actual curricula and instruction. (That's probably the reason we and other analysts have never been able to demonstrate a close relationship between the quality of standards per se and changes in student achievement.) We wanted to know how well aligned the assessments were to the standards, whether they were of high quality, and what type of cognitive demands they placed on students.

But with fifty-one different sets of tests, such an evaluation was out of reach—particularly since any bona fide evaluation of assessments must get under the hood (and behind the curtain) to look at a sizable chunk of actual test items. Getting dozens of states—and their test vendors—to allow us to take a peek was nigh impossible.

So when the opportunity came along to conduct a groundbreaking evaluation of Common Core-aligned tests, we were captivated. We were daunted too, both by the enormity of the task and by the knowledge that our unabashed advocacy of the standards would likely cause any number of doubters and critics to sneer at such an evaluation coming from us, regardless of its quality or impartiality.

So let's address that first. It's true that we continue to believe that children in most states are better off with the Common Core standards than without them. If you don't care for the standards (or even the concept of "common" standards), or perhaps you come from a state that never adopted these standards or has since repudiated them, you should probably ignore this study. Our purpose here is not to re-litigate the Common Core debate. Rather, we want to know, for states that are sticking with the common standards, whether the "next generation assessments" that have been developed to accompany the standards deliver what they promised by way of strong content, quality, and rigor.

It is also true that the study was funded by a number of foundations that care about assessment quality and the Common Core (seven, in fact, including the Fordham Institute's own foundation). If you think that big private foundations are ruining public education, this study is also not for you.

Now is an especially opportune time to look closely at assessments, since the national testing landscape is in a state of flux. According to the Education Commission of the States, as of October 2015, six states and the District of Columbia planned to administer the Partnership for Assessment of Readiness for College and Careers (PARCC)

test in 2015–16 and fifteen states will deploy the Smarter Balanced Assessment Consortium (Smarter Balanced) test.[1] At least twenty-five others will administer state-specific assessments in math and English language arts. Some (Florida, Ohio, and Utah) will use tests developed by the American Institutes for Research (AIR); others (Indiana, Kentucky, and Virginia) are using Pearson-developed products; still others are choosing "blended" versions of consortium and state-developed items (Michigan and Massachusetts). A handful are undecided and currently in the midst of evaluating test vendors through their RFP process (Maine, Louisiana, and South Carolina). About half the states also require an additional assessment for college admissions, such as the ACT or SAT, which is generally administered in grade 11 (and sometimes statewide). And let's not forget that the new SAT will be unveiled in **March 2016**.

Hence there's no way any single study could come close to evaluating all of the products in use and under development in today's busy and fluid testing marketplace. But what we were able to do was to provide an in-depth appraisal of the content and quality of three "next generation" assessments—ACT Aspire, PARCC, and Smarter Balanced—and one best-in-class state test, the Massachusetts Comprehensive Assessment System (MCAS, 2014). In total, over thirteen million children (about 40 percent of the country's students in grades 3–11) took one of these four tests in spring 2015. Of course it would be good to encompass even more. Nevertheless, our study ranks as a major accomplishment—as well as possibly the most complex and ambitious single project ever undertaken by Fordham.

After we agreed to myriad terms and conditions, we and our team of nearly forty reviewers (more about them below) were granted secure access to operational items and test forms for grades 5 and 8 (the elementary and middle school capstone grades that are this study's focus).[2]

This was an achievement in its own right. It's no small thing to receive access to examine operational test forms. This is especially true in a divisive political climate where anti-testing advocates are looking for *any* reason to throw the baby out with the bathwater and where market pressure gives test developers ample reason to be wary of leaks, spies, and competitors. Each of the four testing programs is to be commended for allowing this external scrutiny of their "live" tests—tests that cost them much by way of blood, sweat, tears, and cash to develop and bring to market. They could have easily said "thanks, but no thanks." But they didn't.

Part of the reason they said yes was the care we took in recruiting smart, respected individuals to help with this project. Our two lead investigators, Nancy Doorey and Morgan Polikoff, together bring a wealth of experience in educational assessment and policy, test alignment, academic standards, and accountability. Nancy has authored reports for several national organizations on advances in educational assessment and she co-piloted the Center for K–12 Assessment and Performance Management at ETS. Morgan is assistant professor of education at the University of Southern California and a well-regarded analyst of the implementation of college and career readiness standards and the influence of curricular materials and tests on that implementation. He is an associate editor of the *American Educational Research Journal*, serves on the editorial board for *Educational Administration Quarterly*, and is the top finisher in the RHSU 2015 Edu-Scholar rankings for junior faculty.[3]

Nancy and Morgan were joined by two well-respected content experts who facilitated and reviewed the work of the ELA/Literacy and math review panels. Charles Perfetti, Distinguished University Professor of Psychology at University of Pittsburgh, served as the ELA/Literacy content lead, and Roger Howe, Professor of Mathematics at Yale, served as the math content lead.

1. J. Woods, "State Summative Assessments: 2015–2016 School Year" (Denver, CO: Education Commission of the States, 2015), http://www.ecs.org/ec-content/uploads/12141.pdf. According to ECS, fifteen states are members of the Smarter Balanced Assessment Consortia, and all but one plan to administer the full assessment in grades 3–8 math and English language arts.

2. The study targets "summative," not "formative," assessments, though most of these same test developers also make the latter available.

3. R. Hess, "2016 RHSU Edu-Scholar Public Influence: Top Tens," *Education Week* (blog), January 7, 2016, http://blogs.edweek.org/edweek/rick_hess_straight_up/2016/01/2016_rhsu_edu-scholar_public_influence_top_tens.html.

Given the importance and sensitivity of the task at hand, we spent months recruiting and vetting the individuals who would eventually comprise the panels led by Dr. Perfetti and Dr. Howe. We began by soliciting recommendations from each participating testing program and other sources, including content and assessment experts, individuals with experience in prior alignment studies, and several national and state organizations. Finalists were asked to submit CVs and detailed responses to a questionnaire regarding their familiarity with the Common Core, their prior experience in conducting alignment evaluations, and any potential conflicts of interest. Individuals currently or previously employed by participating testing organizations and writers of the Common Core were not considered. Given that most card-carrying experts in content and assessment have earned their experience by working on prior alignment or assessment-development studies, and that it's nearly impossible to find experts with zero conflicts, we prioritized balance and fairness. In the end, we recruited at least one reviewer recommended by each testing program to serve on each panel; this strategy helped to ensure fairness by equally balancing reviewer familiarity with the various assessments. (Their bios can be found in Appendix E.)

Which brings us to the matter at hand: How did our meticulously assembled panels go about evaluating the tests—and what did they find? You can read plenty on both questions in the Executive Summary and report itself, which includes ample detail about the study design, testing programs, criteria, and selection of test forms, and review procedures, among other topics.

But the short version is this: we deployed a brand new methodology developed by the Center for Assessment to evaluate the four tests—a methodology that was itself based on the Council of Chief State School Officers' 2014 **"Criteria for Procuring and Evaluating High-Quality Assessments."** Those criteria, say their authors, are "intended to be a useful resource for any state procuring and/or evaluating assessments aligned to their college and career readiness standards." This includes, of course, tests meant to accompany the Common Core standards.

About Those Criteria...

The CCSSO Criteria address the "content" and "depth" of state tests in both English language arts and mathematics. For ELA, "content" spans topics such as whether students are required to use evidence from texts; for math, they are concerned with whether the assessments focus strongly on the content most needed for success in later mathematics. The "depth" criteria for both subjects include whether the tests required a range of "cognitively demanding," high-quality items that make use of various item types (e.g., multiple choice, constructed response, etc.), among other things.

The Center for Assessment took these criteria and transformed them into a number of measurable elements that reviewers addressed. In the end, the newly minted methodology wasn't perfect. Our rock-star reviewers improved upon it and wanted others following in their footsteps to benefit from their learned experience. So we made adjustments along the way (see Section I, *Methodology Modifications* for more).

The panels essentially evaluated the extent of the match between the assessment and a key element of the CCSSO document. They assigned one of four ratings to each ELA and math-specific criterion, such that tests received one of four "match" ratings: Excellent, Good, Limited/Uneven, or Weak Match. To generate these marks, each panel reviewed the ratings from the grade 5 and grade 8 test forms, considered the results from the analysis of the program's documentation (which preceded the item review), and came to consensus on the rating.

What did they ultimately find? The summary findings appear below.

TABLE F-1

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

As shown, the PARCC and Smarter Balanced assessments earned an Excellent or Good Match to the subject-area CCSSO Criteria for both ELA/Literacy and mathematics. This was the case with both Content and Depth.

ACT Aspire and MCAS (along with the others) also did well regarding the quality of their items and the depth of knowledge assessed (both of which are part of the Depth rating). But the panelists also found that they did not adequately assess—or in some cases did not really assess at all—some of the priority content in both ELA/Literacy and mathematics at one or both grade levels in the study (Content).

What do we make of these bottom-line results? Simply put, developing a test—like all major decisions and projects in life—is full of trade-offs. PARCC and Smarter Balanced are a better match to the CCSSO criteria, which is not surprising, given that they were both developed with the Common Core in mind. ACT Aspire, on the other hand, was not developed for that explicit purpose. In a paper on their website, ACT officials Sara Clough and Scott Montgomery explain that ACT Aspire was

> *under development prior to the release of the Common Core State Standards [and] not designed to directly measure progress toward those standards. However, since ACT data, empirical research, and subject matter expertise about what constitutes college and career readiness was lent to the Common Core development effort, significant overlap exists between the Common Core State Standards and the college and career readiness constructs that ACT Aspire and the ACT measure.*[4]

Our reviewers also found some "overlap" in MCAS given that the state had added new Common Core items to its 2014 test. Yet the Bay State's intention was not a full redesign, particularly since it was then in the midst of deciding between MCAS and PARCC as its test of choice (the state ultimately decided on a hybrid).[5] To the extent that states want their tests to reflect the grade-level content in the new standards, they should choose accordingly.

The CCSSO Criteria do not consider testing time, cost, or comparability. But those are nonetheless key considerations for states as they make assessment decisions. Although PARCC and Smarter Balanced are a better match to the Criteria, they also take longer to administer and are more expensive. The estimated testing time for

4. S. Clough and S. Montgomery, "How ACT Assessments Align with State College and Career Readiness Standards" (Iowa City, IA: ACT, 2015), http://www.discoveractaspire.org/pdf/ACT_Alignment-White-Paper.pdf.

5. J. Fox, "Education Board Votes to Adopt Hybrid MCAS-PARCC Test," *Boston Globe*, November 17, 2015, https://www.bostonglobe.com/metro/2015/11/17/state-education-board-vote-whether-replace-mcas/aex1nGyBYZW2sucEW2082L/story.html. To the extent that states want their tests to reflect the grade-level content in the new standards, they should choose accordingly.

students in grades 5 and 8, on average, to complete both the ELA/Literacy and mathematics assessments for all four programs is as follows:

- ACT Aspire: three to three and one-quarter hours for all four tests (English, reading, writing, and mathematics)
- MCAS 2014: three and a half hours
- PARCC: seven to seven and a half hours[6]
- Smarter Balanced: five and a half hours

The longer testing times for PARCC and Smarter Balanced are due primarily to their inclusion of extended performance tasks. Both programs use these tasks to assess high-priority skills within the CCSS, such as the development of written compositions in which a claim is supported with evidence drawn from sources; research skills; and solving complex multi-step problems in mathematics. In addition to requiring more time than multiple-choice items, these tasks are also typically costlier to develop and score.[7]

Another trade-off pertains to inter-state comparability. Some states want the autonomy and uniqueness that come with having their own state test developed by their own educators. Other states prioritize the ability to compare their students with those in other states via a multi-state test. We think the extra time and money,[8] plus the comparability advantage, are trade-offs worth making, but we can't pretend that they're not tough decisions in a time of tight budgets and widespread anxiety about testing burden.

Of course we're mindful—as anyone in this field would be—of the recent backlash to testing and the so-called "opt-out movement." We understand that some local and state officials are wary of adopting longer tests. We also suspect that most of the concerns that parents have isn't with the length of one test in May, but with the pressure that educators feel to teach to the test and narrow the curriculum.

If we're right and that's the real problem, the answer is stronger tests, which encourage better, broader, richer instruction, and which make traditional "test prep" ineffective. Tests that allow students of all abilities, including both at-risk and high-achieving youngsters, to demonstrate what they know and can do. More rigorous tests that challenge students more than they've been challenged in the past. But, again, those tests tend to take a bit longer (say, five hours rather than two and a half hours) and cost a bit more. Our point is not to advocate for any particular tests but to root for those that have qualities that enhance, rather than constrict, a child's education and give her the best opportunity to show what she's learned.

A discussion of such qualities, and the types of trade-offs involved in obtaining them, are precisely the kinds of conversations that merit honest debate in states and districts.

******

We at Fordham don't plan to stay in the test-evaluation business. The challenge of doing this well is simply too overwhelming for a small think tank like ours. But we sincerely hope that others will pick up the baton, learn from

6. The 2015–16 PARCC revisions will reduce this time by an estimated one and a half hours.

7. That said, Matthew Chingos, in a 2012 study on state assessment spending, found that "collaborating with other states achieves cost savings simply by spreading fixed costs over more students…" (page 22). See M. Chingos, "Strength in Numbers: State Spending on K–12 Assessment Systems" (Washington, D.C.: Brookings Institution, November 29, 2012), http://www.brookings.edu/research/reports/2012/11/29-cost-of-ed-assessment-chingos.

8. Note that the per-pupil costs for PARCC, Smarter Balanced, and ACT Aspire are in the same ballpark, ranging from roughly $22 to $25 depending on the tested subjects. The MCAS, typically viewed as a higher-quality state test, costs $42 per student. The costs associated with many of the prior state tests were considerably lower than these figures so changing tests represented an increase for them. See M. Chingos, "Strength in Numbers." Cost estimates for PARCC and Smarter Balanced can be found here: http://www.parcconline.org/news-and-video/press-releases/248-states-select-contractor-to-help-develop-and-implement-parcc-tests; http://www.smarterbalanced.org/faq/7-what-does-it-cost-for-a-state-to-participate-in-smarter-balanced-assessment-consortium/. Per MCAS, "The approximate cost of the legacy MCAS assessment is $42 per student for ELA and mathematics per estimates presented to the Massachusetts State Board of Elementary and Secondary Education in fall 2015" (personal email communication with Michol Stapel, January 22, 2016). Per ACT Aspire, "The estimated price for 2016 is $25 per student and includes English, Mathematics, Reading, Writing, and Science subject tests" (personal email communication with Elizabeth Sullivan, January 21, 2016).

our experience, and provide independent evaluations of the assessments in use in the states that have moved away from PARCC, Smarter Balanced, and ACT Aspire.

Not only will such reviews provide critical information for state and local policymakers, as well as educators, curriculum developers and others, they might also deter the U.S. Department of Education from pursuing a dubious plan to make states put their new assessments through a federal evaluation system. In October 2015, the Department issued procedures for the "peer review" process that had been on hold for the last three years. The guidelines specify that states must produce evidence that they "used sound procedures in design and development to state tests aligned to academic standards, and for test administration and security." Count us among those who think renewed federal vetting of state tests invites unwanted meddling from Uncle Sam (and could spark another round of backlash akin to what happened to the Common Core itself a few years back.) Besides, twelve years during which the Department already had such guidance in place did little to improve the quality of state tests—hence the recent moves to improve them.

Parting Thoughts

We are living in a time of political upheaval, divisiveness, and vitriol. The public's faith in government and other large institutions is at an all-time low. So we're glad to be the bearers of good news for a change. All four tests we evaluated boasted items of high technical quality. Further, the next generation assessments that were developed with the Common Core in mind have largely delivered on their promises. Yes, they have improvements to make (you'll see that our reviewers weren't shy in spelling those out). But they tend to reflect the content deemed essential in the Common Core standards and demand much from students cognitively. They are, in fact, the kind of tests that many teachers have asked state officials to build for years.

Now they have them.