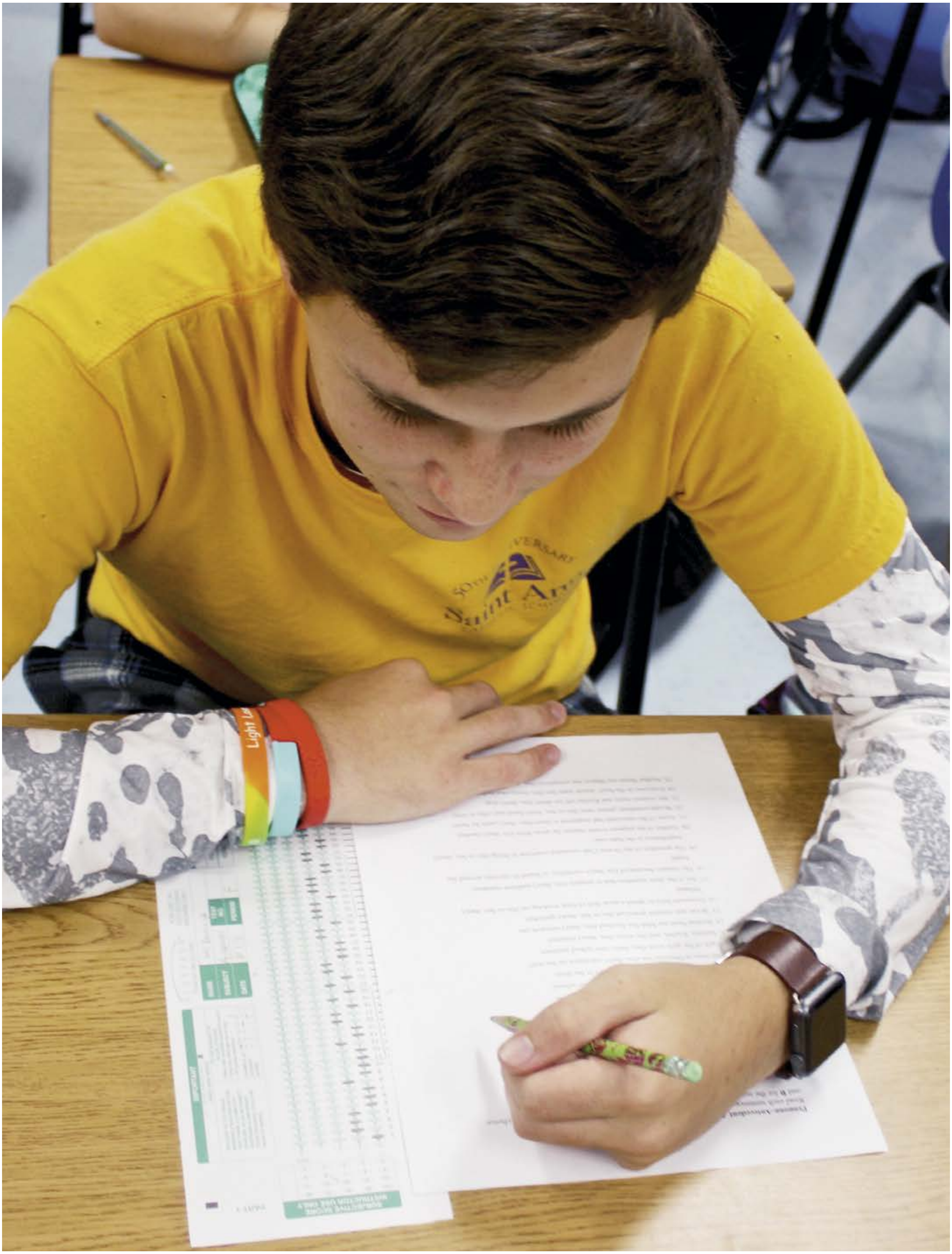


PHOTOGRAPH / BARBARA.CHAPMAN06; TWENTY20

TESTING COLLEGE READINESS

Massachusetts compares the validity of two standardized tests

THE STATE OF MASSACHUSETTS introduced a system of standardized testing in its public schools three years before the federal No Child Left Behind Act of 2001 mandated such practices for all 50 states. Although the tests have evolved over time, the Massachusetts Comprehensive Assessment System (MCAS) has been in place ever since. But after Massachusetts adopted the Common Core State Standards in 2010, its education leaders faced a decision: whether to stick with MCAS, which it had already revised to align with the Common Core, or switch to a "next-generation" test that was specifically designed for the Common Core—and to assess students' readiness for college. More than 40 other states have signed on to Common Core, and many face similar decisions about their student assessment systems.

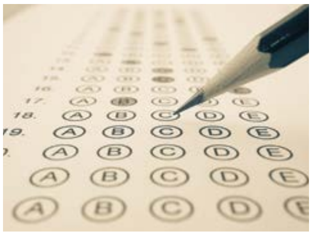
As a member of the multistate Partnership for Assessment of Readiness for College and Careers (PARCC) consortium, Massachusetts had a ready alternative in the new PARCC assessments. As of 2010, 45 states had joined either PARCC or the Smarter Balanced Assessment Consortium that was likewise developing new assessments seeking to better gauge students' higher-level thinking skills, but the number of states participating in both consortia has since fallen.

The stated goal of the PARCC exam is to measure whether students are on track to succeed in college, while the MCAS test aims to measure students' proficiency relative to statewide curriculum standards. But whether the PARCC test actually does a better job of measuring college preparedness was an open question prior to the fall of 2015. The Massachusetts Executive Office of Education commissioned this study in hopes of uncovering timely, rigorous evidence on how accurately the two tests assess college readiness.

This is the first study of its kind. Prior to its authorization, there was no reliable evidence that could demonstrate whether the new Common Core–aligned assessments (PARCC or Smarter Balanced) provide accurate information about which students are prepared for success in college.

by IRA NICHOLS-BARRER, KATE PLACE, ERIN DILLON, and BRIAN GILL

Ultimately, we found that the PARCC and MCAS 10th-grade exams do equally well at predicting students' college success, as measured by first-year grades and by the probability that a student needs remediation after entering college. Scores on both tests, in both math and English language arts (ELA), are positively correlated with students' college outcomes, and the differences between the predictive validity of PARCC and MCAS scores are modest. However, we found one important difference between the two exams: PARCC's cutoff scores for college-and career-readiness in math are set at a higher level than the MCAS proficiency cutoff and are better aligned with what it takes to earn "B" grades in college math. That is, while more students fail to meet the PARCC cutoff, those who do meet PARCC's college-readiness standard have better college grades than students who meet the MCAS proficiency standard.



By examining rigorous evidence about the validity of these two standardized tests, Massachusetts provides a model for other states facing difficult choices about whether and how to upgrade their assessment systems.

These results likely played a role in the November 2015 decision of the Massachusetts Board of Elementary and Secondary Education to adopt neither MCAS nor PARCC, but rather to develop a hybrid assessment that will aim to draw on the best of both tests. Our analysis cannot speak to the wisdom of that choice, which will become clear only with time. Nor should one assume that our study's results are applicable to other states facing similar decisions: Massachusetts has been a national leader in establishing high-quality learning standards for its students, and MCAS is widely regarded as one of the country's more sophisticated assessment systems. We do not have evidence on whether PARCC outperforms the assessment systems used in other states.

By examining rigorous evidence about the validity of both of these tests, however, Massachusetts provides a model for other states facing difficult choices about whether and how to upgrade their assessment systems.

An Experimental Test

Whether the PARCC test succeeds in measuring college preparedness better than the MCAS is an empirical question; answering it requires a rigorous, independent analysis of which test better predicts college outcomes. Our study sought to provide such an analysis. Our primary focus was the strength of association between students' MCAS or PARCC scores and their first-year college grades. We also examined how well each test predicts whether students are assigned to remedial coursework in their freshman year.

At the end of the 2014–15 academic year, Massachusetts arranged to have a sample of 866 college freshmen take the 10th-grade MCAS and PARCC assessments. (Our final analytic sample was 847 after the scores of 19 students were removed for technical reasons—for instance, because the students did not complete the exam or their scores showed evidence of low effort.) The students were enrolled at 11 public higher-education campuses in Massachusetts. Each student was randomly assigned to complete one component of either the MCAS or the PARCC exam. This approach ensured that the students taking the PARCC assessments were not systematically different from those taking the MCAS tests.

We collected college transcript data for all students in the sample, allowing us to examine the relationship between exam scores and several different outcomes, including grade point average (GPA) and enrollment in remedial courses. By examining whether high-scoring students perform better in college than low-scoring students, we can determine whether or not the exam scores have validity in predicting college outcomes. We were also able to examine whether students who meet designated standards on the tests ("proficient" on MCAS and "college-ready" on PARCC) are likely to be prepared for college as indicated by their need for remedial coursework and by their ability to earn "C" or "B" grades in college.

This methodology has its limitations. Ideally, a study of predictive validity would be longitudinal, tracking the outcomes of students over at least three years, from the point when they complete each exam (in 10th grade) to the end of their first year in college. But Massachusetts could not wait that long before choosing its assessment. By testing college freshmen, we could immediately provide evidence regarding

PHOTOGRAPH / SHUTTERSTOCK

the college outcomes of students relative to their performance on the MCAS or PARCC exams. Our own data suggest that this approach is an acceptable proxy for a longitudinal study: the relationship between our sample's scores on MCAS when the students took it in 10th grade and their college GPA is very similar to the relationship between their 2015 MCAS scores and their college GPA.

Our study was also limited to college students at public institutions in the state, a group that is not representative of the statewide population of public high-school students. Although our slate of participating institutions (six community colleges, three state universities, and two University of Massachusetts campuses) roughly mirrors public higher education across the state, our sample did not include students who dropped out of college before the spring semester or who enrolled in private or out-of-state institutions.

Nonetheless, the students in our sample do not differ greatly in terms of their exam performance from students statewide: students in the sample had average MCAS scores that were only slightly different than statewide averages among all 10th graders.

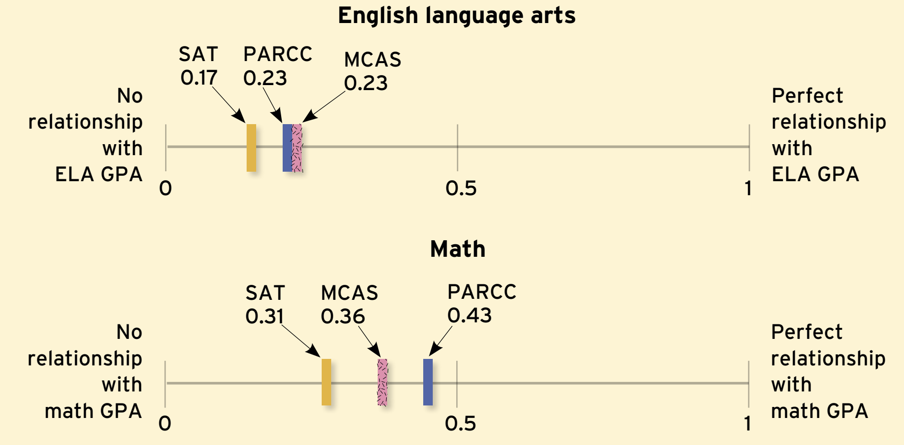
Predicting College Performance

We first assessed the extent to which students' scores on the PARCC and MCAS assessments are related to their college performance (as measured by GPA) and college readiness (as measured by placement in remedial courses). We report the results of these analyses below as correlation coefficients, a statistical measure that summarizes the strength of the relationship between two variables. Correlations have a minimum possible value of -1 (indicating a perfect negative relationship) and a maximum value of 1 (indicating a perfect positive relationship). A correlation of 0 indicates that there is no linear relationship between the two variables. As a benchmark, in our study data, students' high-school SAT scores had a correlation of 0.27 with their college grades. Given the size of the sample, our study would have been able to detect as statistically significant any differences in the correlations of the two tests as small as 0.2.

Exam scores as predictors of college performance. The primary indicator of college success that we examined is GPA. One challenge in working with this outcome is that course grades can reflect the difficulty of a subject and rigor of an institution's grading standards. More-demanding grading standards at some institutions, for example, could lead to lower overall grades among those schools' students, irrespective of the students' general college preparedness. Similarly, particular subject areas might be more challenging, leading to lower GPAs for students who take more courses in those subjects. Failing to account for these differences could have biased the study's findings. We therefore used a two-step process to establish consistency across the study sample before examining the two tests' relationship with GPAs. First, we adjusted grades based on whether or not they were from a remedial course. Second, we adjusted grades for the institution and course subject. (The results did not change when we tested alternative methods for standardizing GPAs, such as omitting remedial course grades or accounting for students' 10th-grade test scores.)

Correlating Test Scores and College Grades (Figure 1)

The correlations between MCAS and PARCC scores and college grades are similar to each other, and both exams are at least as correlated with college grades as are SAT scores, a widely used measure of college readiness.



SOURCE: Authors' calculations

PARCC and MCAS scores performed about equally well in predicting college GPA: the correlations are not statistically distinguishable. In English language arts, the two correlations are identical: scores on both tests have a 0.23 correlation with grades in ELA courses. The math correlations are a bit higher for both assessments, at 0.36 for MCAS and 0.43 for PARCC; the difference between the two is not statistically significant.

Figure 1 shows the subject-specific correlations on a scale.

To provide an additional benchmark, the figure also displays the correlation between students' SAT scores and GPA in the given subject. As seen in the figure, MCAS and PARCC correlations are similar to each other, and both exams are at least as correlated with college grades as are SAT scores.

Taken together, these results allow us to conclude that the scores on the PARCC and MCAS exams are similar in their relationship to college GPA.

Exam scores as indicators of college readiness. PARCC and MCAS also do equally well at predicting which students will need remedial coursework in college, a sign that the students are not fully prepared for college-level work. In ELA, the correlation between MCAS scores and not needing remedial coursework in any subject (0.36) is very similar to the correlation between scores on PARCC's ELA tests and not needing remediation (0.35). Likewise, in math, there is no significant difference between the MCAS (0.35) and PARCC (0.28) correlations with an indicator of which students do not enroll in remedial courses (in any subject) during their first year of college. When we examined whether students enroll in remedial courses in the tested subject specifically, we again found no statistically significant differences in the predictive ability of either test.

The SAT exam provides another measure of college preparedness. The SAT consists of three tests: reading, math, and writing. In total, 737 of the 847 students in our sample had scores in all three components of the exam, allowing us to analyze the relationships between their SAT scores and their MCAS or PARCC scores. In keeping with our other results, we found no clear pattern of differences between the MCAS and PARCC tests with respect to their relationship to SAT scores.

Scores on the PARCC and the MCAS exams do equally well at predicting students' success in college—an important characteristic of any state high-school assessment.

PHOTOGRAPH / SHUTTERSTOCK

Comparing Performance Categories

In addition to assessing the predictive validity of the MCAS and PARCC test scores, we also evaluated the utility of the cutoff scores that define performance levels on each exam. Massachusetts has traditionally used MCAS to assign students to one of four performance categories in each subject. High school students are required to achieve, at minimum, a "needs-improvement" (level two) score in both math and ELA in order to graduate from high school. The percentage of students achieving "proficiency" (level three) also has consequences for schools under federal and state accountability regimes. In our sample of first-year college students, 75 percent of MCAS students scored as proficient or better in math, and 66 percent scored as proficient or better in ELA.

The PARCC exam has defined five different performance categories and specifies that students scoring in the two highest performance categories (level four or five) should be considered college-and-career ready in that subject. In our study data, 60 percent of PARCC students scored as college-and-career ready in math and 66 percent scored as such in ELA.

PARCC's college-and-career readiness standard is meant to identify students who have at least a 75 percent chance of earning a "C" average in college. We examined whether the PARCC standard meets this goal by modeling the relationship between PARCC scores and the likelihood of obtaining a GPA of 2.0 (equivalent to a "C") or better, and then calculating this likelihood at the PARCC cutoff score for college-and-career readiness.

We find that the PARCC exam's college-ready standard not only meets but exceeds its stated target. In ELA, students at the college-ready cutoff score have an 89 percent probability of earning at least a "C" average, and in math, students at the cutoff score have an 85 percent probability of earning a "C" average or better.

In comparison, students at the MCAS cutoff score for proficiency have an 89 percent probability of earning at least a "C" average in ELA, but only a 62 percent probability of earning at least a "C" average in math. This indicates that meeting the PARCC college-ready standard in math provides a better signal that a student is indeed prepared for college-level work than does achieving proficiency on the math MCAS. At the same time, a higher share of students who were not deemed college-ready on the PARCC math test would have nonetheless been able to earn a "C" average.

More differences between the MCAS and PARCC performance levels come to light when we examine students' average college GPAs and the percentage of students earning at least a "B" average (see Figure 2). Students in the proficient category on the MCAS ELA assessment earned an average GPA of 2.66 in their first-year college English classes. This was not statistically distinguishable from the 2.76 GPA earned by students in the college-ready category on the PARCC ELA assessment. In contrast, students who were rated proficient on the MCAS math exam had a significantly lower math GPA (2.39) than students scoring in the college-and-career ready group for PARCC in math (2.81); this margin is equivalent to the difference between a "C+" and a "B-."

A similar pattern emerges in the percentages of students achieving a "B" average. In ELA, students in PARCC's college-ready performance category were about 8 percentage points more likely to achieve a 3.0 GPA than students rated as proficient on MCAS, but the difference is not statistically significant. In math, however, the difference is larger: in the PARCC college-ready group, students were 24 percentage points more likely to achieve "B" grades than students rated as proficient on the MCAS math test, and the difference is statistically significant.

We also compared the validity of these performance categories by examining the percentage of students who needed remedial coursework in their freshman year despite meeting a test's key performance threshold. This reveals a similar pattern to our results for college GPA: in math, the percentage of proficient MCAS students who enrolled in remedial courses (23.9 percent) exceeds the percentage of college-ready PARCC students who took remedial courses (12.6 percent). For the ELA performance threshold, the remediation rate for proficient MCAS students (22.5 percent) is also higher than the remediation rate for college-ready PARCC students (15.0 percent), though this difference is smaller and is not statistically significant.

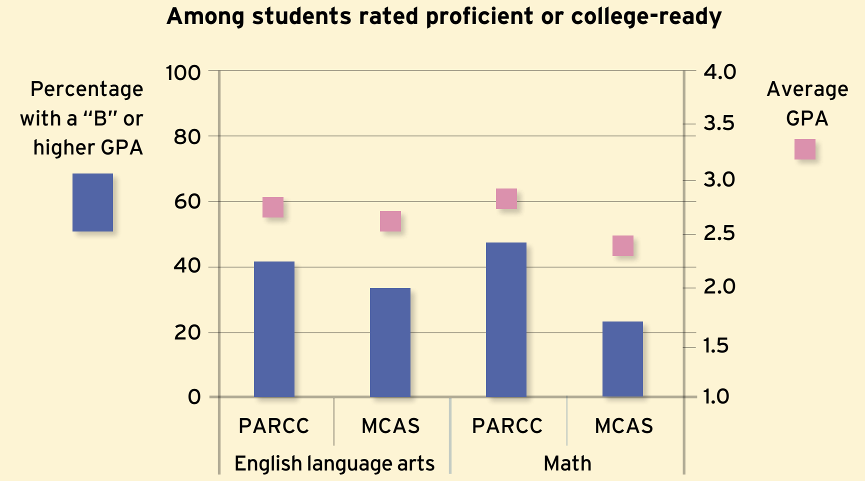
It is helpful to remember that the definitions of the PARCC and MCAS performance categories are not directly comparable: the PARCC exam explicitly seeks to identify students who are prepared for college, whereas the MCAS performance levels are more narrowly targeted to measure proficiency relative to state curriculum standards.

In addition, the differences in the success rates of students meeting key performance levels on each test are not due to differences in the tests' underlying ability to predict college outcomes. Because the underlying predictive ability of the scores is similar, performance levels could be defined in a comparable way for MCAS and PARCC, thereby making them equally predictive of college outcomes.

Better Grades for Those Rated College-Ready on PARCC

(Figure 2)

Students who were rated proficient on the MCAS math exam had a significantly lower average math GPA than students scoring in the college-and-career ready group for PARCC in math, and were significantly less likely to achieve "B" or higher grades. In English language arts, the differences were not statistically significant.



SOURCE: Authors' calculations

Implications for Massachusetts

This is the first study in any state to compare the predictive validity of one of the next-generation, consortium-developed assessments with that of the state assessment it would replace. As such, it provided timely evidence to education officials who were deciding which evaluation system to use in Massachusetts. The study's results demonstrate that scores on PARCC and MCAS do equally well at predicting students' success in college—an important characteristic of any state high-school assessment.

Results regarding the performance standards of each exam are also relevant to the decisionmaking process. In mathematics, the PARCC exam has defined a higher performance standard for college-and-career readiness than the current MCAS standard for proficiency, making the PARCC performance standards better aligned with college grading standards and remediation needs.

But because the underlying scores on the MCAS and PARCC assessments are equally predictive, Massachusetts policymakers had more than one option to align high-school mathematics-test standards with college readiness: one possibility would have been to adopt the PARCC exam, but another option would have been to continue using the MCAS test while simply setting a higher score threshold for college readiness. Either of these options would have achieved the goal of ensuring that the state's high-school assessments provide better information about college readiness to students, parents, educators, and policymakers.

While our study provides valuable evidence on predictive validity, this, of course, is not the only consideration that should inform a state's decision as to its preferred assessment. Exams may differ on a variety of other dimensions that are relevant to the state's choice. For example, the content knowledge and problem-solving skills measured by the PARCC and MCAS tests are not identical, and the tests might differ in the extent to which they align with specific high-school curricular reform goals or teaching standards. Differences in the content of the tests could also prompt changes in curriculum or instruction that might later produce differences in college success; our study cannot assess this possibility. These additional considerations, as well as a desire to maintain independent control of its assessment program outside of the constraints of a multistate consortium, likely played into the Massachusetts state board of education's ultimate decision to develop and adopt a hybrid test.

Performance levels could be defined in a comparable way for MCAS and PARCC, thereby making them equally predictive of college outcomes.

Broader Implications

This study provided timely evidence to decisionmakers in Massachusetts seeking to choose an examination system. For those who might be tempted to use these results to draw conclusions about the Common Core standards themselves, it's worth repeating that the MCAS exam has in recent years been revised to align with those standards. In other words, this was a test of two Common Core–aligned exams, not a Common Core–aligned exam and a starkly different alternative.

Furthermore, over most of the past decade, the standards for student proficiency that Massachusetts has set on the MCAS exam have far exceeded those established by most other state testing programs. If the current Massachusetts proficiency standards fall *somewhat* short of identifying students who are fully prepared to succeed at college-level math, it is likely that the proficiency standards used in other state assessment systems under No Child Left Behind fell *far* short of identifying college readiness. Between 2013 and 2015, however, many states dramatically raised their proficiency standards—in some cases by adopting new assessments aligned to the Common Core (see "After Common Core, States Set Rigorous Standards," *features,* Summer 2016).

Even though we cannot directly compare other states' assessments with the PARCC test, our study provides useful evidence for any state considering adopting PARCC. In particular, it demonstrates that the PARCC exam performs at least as well as the SAT in predicting students' outcomes in college. It also demonstrates that PARCC chose demanding thresholds for deeming a student "college-ready," giving students good information about whether they are prepared to succeed in college courses. This is particularly important, because individual states using PARCC have the discretion to set their performance levels lower than those specified by the test developers. In Ohio, before dropping out of the PARCC consortium, the state chose to adopt a lower standard of proficiency on the PARCC exam. Ohio's decision amounted to a state policy of grade inflation. It may have made students, parents, and educators happy in the short run, but it gave students unrealistically optimistic signals about their true readiness for college.

The bottom line is that, as many states weigh difficult choices about whether to keep or replace current statewide assessments, there is no substitute for examining rigorous evidence about the validity of the alternatives under consideration. By commissioning this study, Massachusetts has again provided a model for the nation.

Ira Nichols-Barrer is a researcher at Mathematica Policy Research, where Erin Dillon and Kate Place are analysts and Brian Gill is a senior fellow.

PHOTOGRAPH / SHUTTERSTOCK