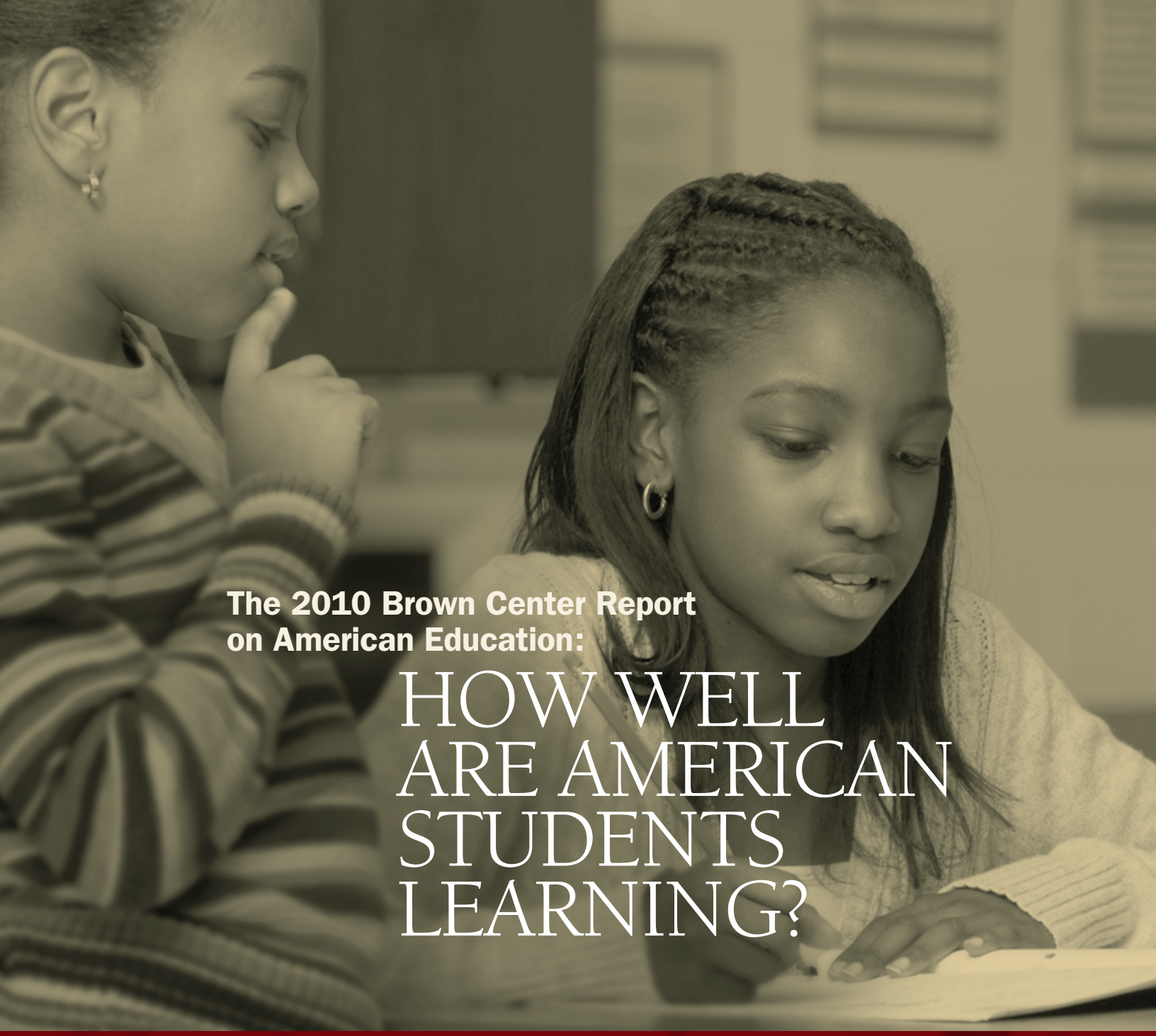


The 2010 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With Sections on International Tests, Who's Winning the Real Race to the Top, and NAEP and the Common Core State Standards

BROWN CENTER on Education Policy at BROOKINGS

ABOUT BROOKINGS

The Brookings Institution is a private nonprofit organization devoted to independent research and innovative policy solutions. For more than 90 years, Brookings has analyzed current and emerging issues and produced new ideas that matter—for the nation and the world.

ABOUT THE BROWN CENTER ON EDUCATION POLICY

Raising the quality of education in the United States for more people is imperative for society's well-being. With that goal in mind, the purpose of the Brown Center on Education Policy at Brookings is to examine the problems of the American education system and to help delineate practical solutions. For more information, see our website, www.brookings.edu/brown.

This report was made possible by the generous financial support of The Brown Foundation, Inc., Houston.

The 2010 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With Sections on International Tests, Who's Winning the Real Race to the Top, and NAEP and the Common Core State Standards

February 2011
Volume II, Number 5

by:
TOM LOVELESS
Senior Fellow, the Brown Center on Education Policy

TABLE OF CONTENTS

Research assistance by:

MICHELLE CROFT
Brown Center on Education Policy

THE 2010 BROWN CENTER REPORT ON AMERICAN EDUCATION

This edition of the Brown Center Report marks the tenth issue of the series and the final issue of Volume II. The publication began in 2000 with Bill Clinton in the White House and the Bush-Gore presidential campaign building toward its dramatic conclusion. That first report was organized in a three-part structure that all subsequent Brown Center Reports followed. Part I presents the latest results from state, national, or international assessments and alerts readers to important trends in the data. Part II explores an education issue in depth, sometimes by investigating different sources of empirical evidence than previous research, sometimes by posing a conventional question in an unconventional way. Part III analyzes a current or impending question regarding education policy. In all three sections, the studies strive to ask clear questions, gather the best available evidence, and present findings in a nonpartisan, jargon-free manner.

Part I of this year's Brown Center Report focuses on international assessments. The latest data from the Programme for International Student Assessment (PISA) were released in December 2010. The performance of the United States was mediocre, and although notching gains in all three subjects, the country scored near the international average in reading literacy and scientific literacy and below average in mathematical literacy. The term "literacy" is a signal that PISA covers different content than most achievement tests, and, indeed, assesses different skills than are emphasized in the school curriculum. As the 2006 PISA Framework states, the knowledge and skills tested on PISA "are defined not primarily in terms of a common denominator of national school curricula but in terms of what skills are deemed to be essential for future life."[1]

Two myths of international assessments are debunked—the first, that the United States once led the world on international tests of achievement. It never has. The second myth is that Finland leads the world in education, with China and India coming on fast. Finland has a superb school system, but, significantly, it scores at the very top only on PISA, not on other international assessments. Finland also has a national curriculum more in sync with a "literacy" thrust, making PISA a friendly judge in comparing Finnish students with students from other countries. And what about India and China? Neither country has ever participated in an international assessment. How they would fare is unknown.

Part II of the report looks at state test scores on the National Assessment of Educational Progress (NAEP) in light of the recent Race to the Top competition. The federal program encouraged states to apply for $4.35 billion in new money by promising to pursue a reform agenda backed by the Obama administration. Twelve states (for this discussion, the District of Columbia will be called a state) won the grants. But are the states that won the grants the same states that have accomplished the greatest gains in student learning? Not necessarily.

Who's winning the real race to the top? Both short- and long-term gains on NAEP are calculated with statistical controls for changes in the demographic characteristics of each state's students. Eight states—Florida, Maryland, Massachusetts, District of Columbia, Kentucky, New Jersey, Hawaii, and Pennsylvania—stand out for making superior gains. At the other end of the distribution, Iowa, Nebraska, West Virginia, and Michigan stand out for underperforming. Five of the eight impressive states won grants, but three did not. And a few states won grants even though they are faring poorly in the race to boost student achievement. Some of the reasons why a program called Race to the Top could distribute grant money in this manner are discussed.

Part III looks at NAEP. In June 2010, the Common Core State Standards Initiative released grade-by-grade standards for reading and mathematics. Two consortia were awarded $330 million to write tests aligned to the standards, and a total of 46 states have signed to at least one group. As the only assessment administered to representative samples of American students, NAEP has called itself "the Nation's Report Card" for decades.

How well does NAEP match up with the Common Core? We examined 171 public release items from the eighth-grade NAEP math test and coded them based on the grade level the Common Core recommends that the content be taught. The items registered, on average, two to three years below the eighth-grade mathematics recommended by the Common Core. More than 90 percent of the items from the "number" strand (content area) cover material below the eighth grade. Almost 80 percent of the items assessing "algebra" are, in fact, addressing content in the curriculum that is taught before eighth grade. With Common Core assessments on tap to begin in the 2014–2015 school year, policymakers and analysts alike need to start thinking now about how NAEP and the Common Core assessments can be reconciled so as to inform, not to confuse, the public about student achievement.

An overarching theme of this year's report is that events in the field of education are not always as they appear to be—and especially so with test scores. Whether commentators perpetrating myths of international testing, states winning races while evidencing only mediocre progress, or an eighth-grade test dominated by content below the eighth grade, the story is rarely as simple as it appears on first blush. This report tried to dig beneath the surface and uncover some of the complexities of these important issues.

Part I INTERNATIONAL TESTS

THE RESULTS OF THE 2009 PROGRAMME FOR INTERNATIONAL Student Assessment (PISA) were released in December 2010. The test is given every three years to students approximately 15 years of age. In the United States, most students taking PISA are in the fall semester of tenth grade. PISA measures reading literacy, mathematical literacy, and scientific literacy. The term "literacy" refers to the ability to apply knowledge and skills, whether learned in or out of school, to real-world situations. The three subjects alternate as the focus of the assessment, with reading literacy the main subject of the 2009 test. The first PISA test was given in 2000. Sixty-five countries participated in 2009:

the 34 members of the Organization for Economic Cooperation and Development (OECD), the parent organization of PISA, and an additional 31 partner nations.[2]

How did the United States do? Mediocre to poor. Table 1-1 displays the 2009 scores for all three subjects. The United States scored slightly above average in reading and science and below average in math. The top and bottom ten nations are also shown (for the sake of discussion, all participants, even if sub-national, will be referred to as "countries" or "nations"). The United States is far below the top-scoring nations in all three subjects, especially in math, where Shanghai-China outdistanced the United States by 113 points.[3] In all three subjects, Shanghai-China, South Korea, Singapore, Hong Kong-China, and Japan do very well. It is not just Asian countries that rank high on the PISA. Finland and Canada also do very well. As is evident in Table 1-1, top-scoring nations are among the most economically robust nations in the world, and those at the bottom of the rankings tend to be economically developing nations.

Good news for the United States can be found in Table 1-2. United States scores were up from the last PISA in each subject. Scores increased 5 points in reading, 13 points in math, and 13 points in science. The math and science gains are statistically significant. They may even have economic significance if a recent study by Stanford

University's Eric Hanushek is to be believed.[4] The study was conducted prior to the release of 2009 scores. Hanushek estimated that an increase of 25 points on PISA over the next 20 years would boost United States GDP by $41 trillion. If the gains from 2006 to 2009 are duplicated when the PISA is next given in 2012, the goal of making 25-point gains in math and science will be met far ahead of schedule. Scores bounce up and down, so making such gains is not guaranteed. As Table 1-2 shows, the 2009 gains in reading and math followed losses in both subjects during the previous intervals.

Some of the reaction to the scores was curious. United States gains were mostly ignored. The *New York Times* focused on Shanghai's performance.[5] Chester E. Finn Jr. said, "Wow, I'm kind of stunned. I'm thinking Sputnik."[6] President Obama also invoked Sputnik, the 1957 Russian satellite launch that galvanized a national reform movement in math and science education. This time the competitor is China. According to the *New York Times*, in a North Carolina speech, the president warned Americans that "with billions of people in India and China 'suddenly plugged into the world economy,' nations with the most educated workers will prevail."[7] The headline in *Time* magazine declared "China Beats Out Finland for Top Marks in Education."[8] In the *Christian Science Monitor*, Bob Wise, president of the Alliance for Excellent Education and former governor of West Virginia, expressed relief. "The good news is that the free-fall seems to have stopped—and it was a free-fall for a while."[9]

These reactions are misleading. They reinforce myths that Americans have about international testing.

Top Ten and Bottom Ten Countries on PISA 2009

Table 1-1

Reading		Math		Science	
Country	**Scale Score**	**Country**	**Scale Score**	**Country**	**Scale Score**
Shanghai-China	556	Shanghai-China	600	Shanghai-China	575
South Korea	539	Singapore	562	Finland	554
Finland	536	Hong Kong-China	555	Hong Kong-China	549
Hong Kong-China	533	South Korea	546	Singapore	542
Singapore	526	Chinese Taipei	543	Japan	539
Canada	524	Finland	541	South Korea	538
New Zealand	521	Liechtenstein	536	New Zealand	532
Japan	520	Switzerland	534	Canada	529
Australia	515	Japan	529	Estonia	528
Netherlands	508	Canada	527	Australia	527
United States	**500**	**International Avg**	**496**	**United States**	**514**
International Avg	**493**	**United States**	**487**	**International Avg**	**501**
Tunisia	404	Jordan	387	Argentina	401
Indonesia	402	Brazil	386	Tunisia	401
Argentina	398	Colombia	381	Kazakhstan	400
Kazakhstan	390	Albania	377	Albania	391
Albania	385	Tunisia	371	Indonesia	383
Qatar	372	Indonesia	371	Qatar	379
Panama	371	Qatar	368	Panama	376
Peru	370	Peru	365	Azerbaijan	373
Azerbaijan	362	Panama	360	Peru	369
Kyrgyz Republic	314	Kyrgyz Republic	331	Kyrgyz Republic	330

United States PISA Scores

Table 1-2

	PISA 2000	PISA 2003	PISA 2006	PISA 2009
Reading	504	495	–	500
Mathematics	–	483	474	487
Science	–	–	489	502

From *Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States*, p. 26.

Comparison of Countries in the First International Mathematics Study (FIMS)

Table 1-3

	FIMS 1964	Last TIMSS (1995–2007)*	PISA 2009
Israel	0.62 (1)	-0.51 (12)	-0.58 (12)
Japan	0.55 (2)	0.56 (1)	0.25 (2)
Belgium	0.49 (3)	0.23 (3)	0.11 (4)
Finland	0.23 (4)	0.06 (5)	0.37 (1)
Germany	0.16 (5)	-0.05 (7)	0.09 (6)
England	0.05 (6)	-0.01 (6)	-0.13 (10)
Scotland	-0.05 (7)	-0.27 (11)	-0.06 (7)
Netherlands	-0.11 (8)	0.22 (4)	0.22 (3)
France	-0.13 (9)	0.24 (2)	-0.08 (8)
Australia	-0.27 (10)	-0.18 (9)	0.10 (5)
United States	-0.35 (11)	-0.06 (8)	-0.18 (11)
Sweden	-0.51 (12)	-0.23 (10)	-0.11 (9)
Scale or Raw Score:			
United States	17.8	508	487
All Countries (mean)	23	500	496
12 Listed Countries (mean)	23	514	505

*Year of last TIMSS participation. TIMSS 2007: Israel, Japan, England, Scotland, Australia, United States, and Sweden. TIMSS 2003: Belgium (Flemish) and Netherlands. TIMSS 1999: Finland. TIMSS 1995: Germany and France.

Note: z-score is computed using the twelve-nation mean and SD of 15 for FIMS and 100 for TIMSS and PISA.

The Two Biggest Myths of International Testing

Myth #1. The United States once led the world on international tests.

Many Americans believe that students in the United States ranked number one in the world on international tests several decades ago and that after years of bad policies fell to the bottom of the pack. Typical is a September 2010 story in *Newsweek* magazine, which states, "U.S. students, who once led the world, currently rank 21st in the world in science and 25th in math."[10]

This is a myth. The United States never led the world. It was never number one and has never been close to number one on international math tests. Or on science tests, for that matter. For the sake of simplicity, let's stick to math. It is more accurate to say that the United States has always trailed the world on math tests. And, despite Bob Wise's comments, there has been no sharp decline—in either the short or long run. The United States performance on PISA has been flat to slightly up since the test's inception, and it has improved on Trends in Mathematics and Science Study (TIMSS) since 1995.

The First International Math Study (FIMS) was conducted in 1964.[11] Twelve countries participated: Australia, England, Belgium, Finland, France, Germany, Israel, Japan, Netherlands, Scotland, Sweden, and the United States. The project was carried out by the International Association for the Evaluation of Educational Achievement (IEA), an organization founded in 1958 by several prominent educational researchers from around the world, including Benjamin Bloom and Robert Thorndike of the United States and Torsten Husén of Sweden.[12] They believed that nations could be compared on academic achievement by testing a random sample of students on the same test, a novel idea at the time. The IEA continues today, administering several assessments that are descendants of that first test, including TIMSS and the Progress in International Reading Literacy Study (PIRLS).[13]

How did the United States do on FIMS? Table 1-3 shows the results for what was called "population 1B," the grade attended by the majority of thirteen-year-olds in each country (eighth grade in the United States). All scores have been converted into z-scores, a measure that expresses scores relative to a test's mean. For the TIMSS and PISA scores in the table's adjacent columns, the mean of the FIMS nations is used for calculating z-scores, keeping the comparison group consistent over time. Z-scores are calibrated in standard

deviation units. Positive scores indicate a country scoring above average (set at 0.00); negative scores are below average. Rank among the FIMS countries is also given.

In 1964, the United States ranked eleventh out of twelve countries, with a z-score of -0.35. Only Sweden scored lower. Other grades were also tested, and the results for the United States were equally disappointing. Today, on both TIMSS and PISA, the United States does better, scoring close to average, with a z-score of -0.06 on TIMSS and -0.18 on PISA. Among all participants on TIMSS and PISA, the United States scores slightly above average on TIMSS and slightly below average on PISA (when compared on PISA to OECD nations only).

Myth #2. Finland has the best educational system in the world, with India and China coming on strong.

Finland's vaunted reputation comes from its fine performance on PISA (see Table 1-3). Advocates of several education policies hold up Finland as a model for the world, especially, of course, when the Finnish system embraces a policy that the advocates favor. Linda Darling-Hammond, in *The Flat World and Education*, argues that the United States should follow Finland's lead in distributing educational resources more equitably, paying teachers higher salaries, decentralizing authority over most educational decisions, and eschewing high-stakes standardized testing.[14] Andreas Schleicher, the head of PISA, points to Finland's emphasis on equity as the principal reason that variation in performance among Finnish schools is low.[15] A study from University of Colorado researchers attributes Finland's success to the abolition of streaming (grouping students between schools by ability).[16] They see a lesson for the American practice of tracking (students of varying abilities attending the same schools but then grouped between classes by ability).

Finland's multi-decade participation in the IEA tests rarely receives mention in these analyses. As displayed in Table 1-3, Finland ranked fourth out of the twelve participating countries in FIMS—a solid showing (z-score of 0.23). But by 1999, Finland slipped to only a little above average in TIMSS (z-score of 0.06), ranking fifth of the original twelve countries and fourteenth of all countries taking the test. One complicating factor is age. Finland's students were younger than the rest of the eighth graders in TIMSS 1999, averaging 13.8 years compared with an international mean of 14.4 years (and 14.2 for the participating FIMS nations).[17] Finland's average age in FIMS was 13.9, approximately the same as the international mean of 13.8.[18] Age is positively related to performance on international tests (as is the number of grades students have attended school). An older Finnish cohort in 1999 probably would have produced a higher score.[19]

Finland stopped participating in TIMSS after the 1999 test. Since 2000, math scores from Finland come from only one test—PISA. On PISA 2009, Finland ranked first among the FIMS countries, a lofty ranking it has held throughout the decade. When the OECD launched PISA in 2000, many small countries believed participating in two international assessments would be repetitive and a potential burden in both money and time.[20] Finland chose to participate in PISA alone, at least for a few years. It is scheduled to participate in TIMSS again in 2011 and plans on testing both seventh and eighth graders, allowing for an investigation of age and grade effects.

Why is Finland successful? The policy explanations given by the analysts cited above—teacher professionalism, decentralizing authority, policies promoting equity,

In 1964, the United States ranked eleventh out of twelve countries.

The content of PISA is a better match with Finland's curriculum than is the TIMSS content.

de-streaming—are tenable reasons. Being cross-sectional, however, PISA and TIMSS data provide only weak evidence on the causes of Finland's success. As the FIMS scores indicate, Finnish students did quite well in 1964, several decades before many of the policies targeting professionalism, equity, decentralization, and de-streaming were adopted. This suggests that cultural and societal factors, which predate and are intertwined with the policies in question, may be the real drivers of success. Yet Finland also scores higher on PISA than on TIMSS. Why is that?

A plausible hypothesis stems from differences in the content of the two tests. The content of PISA is a better match with Finland's curriculum than is the TIMSS content. The objective of TIMSS is to assess what students have learned in school. Thus, the content of the test reflects topics in mathematics that are commonly taught in the world's school systems. Traditional domains of mathematics—algebra, geometry, operations with numbers—are well represented on TIMSS.

The objective of PISA, in contrast, is not to assess achievement "in relation to the teaching and learning of a body of knowledge."[21] As noted above, that same objective motivates attaching the term "literacy" to otherwise universally recognized school subjects. Jan de Lange, the head of the mathematics expert group for PISA, explains, "Mathematics curricula have focused on school-based knowledge whereas mathematical literacy involves mathematics as it is used in the real world."[22] PISA's Schleicher often draws a distinction between achievement tests (presumably including TIMSS) that "look back at what students were expected to have learned" and PISA, which "looks ahead to how well they can extrapolate from what they have learned and apply their knowledge and skills in novel settings."[23]

The emphasis on learner-centered, collaborative instruction and a future-oriented, relevant curriculum that focuses on creativity and problem solving has made PISA *the* international test for reformers promoting constructivist learning and 21st-century skills.[24] Finland implemented reforms in the 1990s and early 2000s that embraced the tenets of these movements. Several education researchers from Finland have attributed their nation's strong showing to the compatibility of recent reforms with the content of PISA.[25]

The reforms have not avoided controversy. When PISA results showed Finland to be the top country in the world in math, a group of more than two hundred university mathematicians in Finland petitioned the Finnish education ministry to complain that, regardless of what PISA was indicating, students increasingly were arriving in their classrooms unprepared in mathematics. Knowledge of fractions and algebra were singled out as particularly weak areas. Two signers of the petition posed the question, "[A]re the Finnish basic schools stressing too much numerical problems of the type emphasized in the PISA study, and are other countries, instead, stressing algebra, thus guaranteeing a better foundation for mathematical studies in upper secondary schools and in universities and polytechnics."[26] One Finnish researcher, analyzing national data, compared the math skills of 15- and 16-year-olds on tests given in 1981 and 2003. Sharp declines were registered on calculations involving whole numbers, fractions, and exponents. The explanation: "'Problem Solving' and putting emphasis on calculators have taken time from explaining the basic principles and ideas in mathematics."[27]

In sum, Finland appears to have an excellent school system; however, its performance varies by test. It scores highest on the

international test reflecting contemporary theories of mathematics (PISA), not as high on the international test tied to curriculum (TIMSS), and not as well as it once did on national tests assessing knowledge of traditional mathematics topics. India and China? No one really knows how they perform on international assessments. They have never participated in them as nations. There is no reliable score to compare academic achievement in China and India with that of other nations.[28]

Doesn't Shanghai's performance on PISA 2009 at least give a clue as to how China would score? No, it does not. For centuries, Shanghai has been the jewel of Chinese schooling, far ahead of its urban peers and light-years ahead of rural schools. Shanghai's municipal website reports that 83.8 percent of high school graduates enter college; the national figure is 24.0 percent.[29]

Within nations, achievement can vary dramatically, even between districts in close geographical proximity. In 1999, a group of school systems in suburban Chicago participated in TIMSS. Known as the First in the World Consortium, the group's eighth graders scored 560 in math, 58 points above the United States national score of 502.[30] Singapore, the top nation, scored 604. Naperville, another Chicago suburban district, participated on its own and scored 569. The Chicago Public Schools also took part, scoring 462. At the time, no one mistook Naperville as being representative of the United States as a whole. And no one seemed surprised by the 107-point difference between Chicago and Naperville despite the mere 30 miles that physically separate them.

Summary

This section of the Brown Center Report reviewed results from PISA 2009. The performance of the United States was mixed. Scores were up in all three subjects—reading literacy, mathematical literacy, and scientific literacy—but the United States still scores either about average or slightly below average for OECD nations. Two myths were addressed. The first is that the United States once led the world on achievement tests but then fell precipitously from its high standing. The truth is that the United States has never led the world on international tests. It scored eleventh out of twelve nations taking the first international assessment in 1964. American performance has remained steady or shown some improvement since then; it has not fallen.

The second myth is that Finland is the top-scoring nation in the world on international tests, with India and China rising. Finland scores at the top on one of the two international tests, PISA, and performs very well—but has never ranked number one—on the other, TIMSS. No one can say for sure how China and India compare with the rest of the world. They have never participated as nations in an international assessment. Finland's high PISA score in math may reflect a national curriculum that mirrors PISA's emphasis on problem solving, real-world mathematics, and 21st-century skills. Critics contend that such an emphasis has weakened Finnish students' knowledge of more traditional topics in mathematics, including fractions and algebra.

The lesson is that international test scores must be interpreted cautiously. Much of what one may hear or read about them is misleading. The content of a test—what it actually measures—matters. The next two sections of the Brown Center Report also examine test scores, with section three looking closely at content. The National Assessment of Educational Progress (NAEP) provides the data for the analyses.

There is no reliable score to compare academic achievement in China and India with that of other nations.

Part II WHO'S WINNING THE REAL RACE TO THE TOP?

THE OBAMA ADMINISTRATION'S KEY EDUCATION INITIATIVE TO date, Race to the Top, encouraged states to compete for federal grants by requiring them to promise several reforms. The program accomplished three objectives. First, it distributed billions of dollars to hard-strapped states desperate for revenue. Second, it incentivized states to embrace reform strategies favored by the administration. And third, it anointed twelve states (for this discussion, the District of Columbia will be called a state) as leaders in the race to improve America's schools.

But are they? State NAEP tests have been administered since 1990. Who is winning the real race to the top, the one measured by student achievement gains, not by the ability to secure federal grants?

The obvious way to answer that question is to examine the latest ranking of states on NAEP. Who scores the highest? But as the first section of the Brown Center Report illustrated with international test scores, simple rankings at one point in time do not always tell the whole story. Factors other than school quality go into test scores. Analysts refer to these as "confounds," influences that muddy the waters. Consider family wealth (or socioeconomic status), which is highly correlated with test scores. In 2009, Massachusetts scored 234 on NAEP's fourth-grade reading test and Mississippi scored 211.[31] Was it because Massachusetts has better schools? Or was it because only about one-third (33%) of Massachusetts pupils come from families poor enough to qualify for free and reduced lunch (a proxy for poverty) but in Mississippi the proportion is more than twice that much (69%).[32]

Confounds are difficult to untangle. Another problem is "selection bias." In education research, selection bias is frequently lurking, even in well-known studies. Some years ago, a study looked at how computers are used for instruction and whether different uses are related to math achievement.[33] The study found that students who spend a lot of time using computers for basic skills instruction have lower achievement than students who study more complex, "higher-level" content. The researcher concluded that computers should not be used for teaching basic skills.

The conclusion is faulty because the students working on basic skills may have

Simple rankings at one point in time do not always tell the whole story.

Selection bias is frequently lurking, even in well-known studies.

been doing so *because* they were weak at math. The students' low achievement led teachers to assign them to a regimen of basic skills instruction. Imagine a survey attempting to correlate teenagers' dietary habits with weight, finding that overweight teens tend to drink diet soft drinks and eat a lot of celery, and concluding that diet soft drinks and celery should be avoided because they cause obesity. Selection bias.

Analysts are on guard when comparing the outcomes of two groups, paying close attention to the manner in which subjects are "selected" into the groups. When students are sorted (or "selected") into groups (students studying basic skills and those studying more difficult material) and a characteristic pivotal to the sorting (prior math achievement) is also related to the outcome of interest (later math achievement), simple comparisons can produce biased findings.[34] Random assignment to groups helps to reduce selection bias and allows for sounder comparisons, making it the preferred approach for research investigating causal connections. When random assignment is impossible, analysts often use gain scores—or change in test scores—as an outcome measure and control for concurrent changes in other factors that may influence the gain.

What We Did

We compiled a complete data set of NAEP scores and demographic variables for the fifty states and the District of Columbia. Seeking to identify states that have improved the most—those winning the real race to the top—we first calculated the study's outcome variable: change in NAEP scores. Beginning in 2003, all states have been required to take NAEP, but before then participation was optional. Many states started as early as 1990. Consequently, two separate analyses were conducted of change in NAEP scores: one using the states' different starting points in NAEP as individual baselines, the other using 2003 as a baseline for all states. The analysis with individual baselines (called Model 1) examines NAEP gains registered by each state from the first year it participated in the assessment to 2009. Each state, then, has its own time frame in Model 1. NAEP gains in fourth- and eighth-grade mathematics and reading (i.e., gains on four tests) are combined to form a composite gain score, which is then converted into an average annual gain (i.e., divided by the number of years).

Controls are employed for changes in three demographic variables—percentage of students qualifying for free lunch, percentage in special education, and percentage in English language learning programs—during the years participating in NAEP.[35] These three demographic variables are known to be correlated with test scores.[36]

The analysis using 2003 as a starting point (Model 2) examines composite NAEP gains from 2003 to 2009 for all states. The same demographic controls are employed as in Model 1. For both models, ordinary least squares regressions were run, residuals were computed, and the residuals were standardized with a mean of 0.00 and standard deviation of 1.00. The residuals provide a measure of relative performance. Positive residuals indicate which states made gains on NAEP that were greater than expected, based on their initial NAEP scores and changes in the three demographic variables; negative residuals indicate which states' gains were smaller than expected.

How does analyzing gains lessen the potential for bias? If one state has a cultural history of emphasizing education and another state does not, those predilections will be present in the baseline measures, "baked in the cake" of initial NAEP scores. Recall the example above of weight and dietary habits. A researcher who focuses

on weight change, instead of just weight, might avoid arriving at the spurious conclusion that diet drinks and celery contribute to obesity. The use of gains is not foolproof. Unobserved differences can still arise that affect achievement in some states more than in others. But considering the current study's question—who's winning the real race to the top?—this analytical strategy equalizes the starting point in the race for states that already had a big lead and those that initially lagged behind.

The two models have different strengths. The Model 1 analysis is superior in utilizing all state achievement data collected by NAEP. It analyzes trends over a longer period of time, up to nineteen years. But it may also produce biased estimates if states that willingly began participating in NAEP in the 1990s are different in some way than states that were compelled to join the assessment in 2003—and especially if that "some way" is systematically related to achievement. They will have baselines allowing those differences to shine through—a selection effect. For example, it is plausible that states joining the state NAEP assessment in its early years had the public's commitment to boosting academic achievement that nonparticipating states did not have. Add in the fact that many states made uncommonly large gains on NAEP between 1998 and 2003, and Model 1 might be susceptible to bias.

Model 2 has the virtue of placing all states on equal footing, time-wise, by limiting the analysis to 2003–2009. But that six-year period may be atypical in NAEP's history—No Child Left Behind dominated national policy discussions—and by discarding more than half of available NAEP data (all of the data collected before 2003), the model could produce misleading estimates of longer term correlations. Keeping these relative strengths in mind, the estimates from both models are examined below. The results from Model 1 can be interpreted as reporting longer term trends and the results from Model 2, shorter term and more recent trends.

Which States Are Doing the Best?

Table 2-1 shows the states' standardized gain scores. The top ten states for Model 1 are shaded in red and the bottom ten in gray. These shadings are carried over onto the Model 2 list to make it easy to see how these states' rankings change between the two models. Eight states—Florida, Maryland, Massachusetts, District of Columbia, Kentucky, New Jersey, Hawaii, and Pennsylvania—stand out for making the top ten of both lists. At the other end of the distribution, Iowa, Nebraska, West Virginia, and Michigan stand out for underperformance. They make the bottom ten in both models.

A clarification before proceeding. The two models report relative performance, not absolute performance. As noted, the Model 2 rankings are based on recent performance, but even a sharp drop in the rankings does not mean a state's achievement level has recently fallen. New York is a good example. Note that the state's gain of 0.58 in Model 1 slides to -1.21 in Model 2. Recall that the average gain for all states is pegged at 0.00. New York made above-average gains from the early 1990s to 2009 but in more recent years has made below-average gains.

Let's unpack New York's numbers to see how they work. The state was an early joiner of NAEP, starting with eighth-grade math in 1990, fourth-grade math and reading in 1992, and eighth-grade reading in 1998. On these four tests, New York gained an average of almost three-quarters of a scale point per year (0.74) from its first year of participation to 2009. The average for all states on the

Some high-achieving states going unrewarded and a few winners having spotty records of achievement.

Models of State Standardized NAEP Gain Scores

Table 2-1

Jurisdiction	Model 1 Residual (NAEP Change from Individual Baseline to 2009)	Jurisdiction	Model 2 Residual (NAEP Change from 2003 to 2009)
Florida	2.15	District of Columbia	2.49
Maryland	1.91	Pennsylvania	1.97
Delaware	1.51	Maryland	1.95
Massachusetts	1.40	Massachusetts	1.70
District of Columbia	1.33	Florida	1.69
New Jersey	1.03	Kentucky	1.25
Kentucky	1.01	Alabama	1.19
Hawaii	0.79	Arkansas	0.97
Mississippi	0.77	Hawaii	0.87
Pennsylvania	0.66	New Jersey	0.78
Texas	0.58	Georgia	0.52
New York	0.58	Rhode Island	0.51
Ohio	0.58	Ohio	0.41
Louisiana	0.53	Nevada	0.35
North Carolina	0.51	Vermont	0.35
Vermont	0.50	Missouri	0.29
Washington	0.49	Tennessee	0.29
Colorado	0.48	New Mexico	0.24
Missouri	0.44	Delaware	0.14
South Carolina	0.41	Indiana	0.05
Arkansas	0.38	Kansas	0.05
Illinois	0.31	Montana	0.04
Tennessee	0.30	North Dakota	0.02
Connecticut	0.23	Texas	0.00
Virginia	0.21	Washington	-0.02
Minnesota	0.18	Connecticut	-0.03
California	0.15	Idaho	-0.04
Georgia	0.14	Mississippi	-0.17
Alabama	0.06	California	-0.18
New Hampshire	0.00	Minnesota	-0.19
Indiana	-0.02	Colorado	-0.31
Nevada	-0.10	Oklahoma	-0.34
Rhode Island	-0.26	Wisconsin	-0.36
Oregon	-0.30	Maine	-0.36
Kansas	-0.34	Virginia	-0.42
Idaho	-0.52	Utah	-0.43
Wyoming	-0.56	Illinois	-0.44
Montana	-0.60	Arizona	-0.47
Alaska	-0.75	South Dakota	-0.48
South Dakota	-0.79	Louisiana	-0.58
Arizona	-0.81	New Hampshire	-0.64
Utah	-0.86	Nebraska	-0.68
North Dakota	-0.91	Oregon	-0.74
Wisconsin	-1.16	Alaska	-0.77
New Mexico	-1.33	Wyoming	-0.91
Michigan	-1.40	South Carolina	-1.12
West Virginia	-1.52	New York	-1.21
Oklahoma	-1.61	North Carolina	-1.36
Nebraska	-1.70	Michigan	-1.77
Maine	-1.92	Iowa	-1.85
Iowa	-2.16	West Virginia	-2.23

Top ten states from model 1

Bottom ten states from model 1

same statistic is 0.65. New York's long-term record is impressive. From 2003 to 2009, however, New York's NAEP scores increased at a slower rate, 0.38 points per year. The average for all states was 0.62, eclipsing the New York gain. On a relative basis, then, New York looks like a star performer in Model 1 and an underperformer in Model 2. But, as should be emphasized again, the state made absolute gains in both periods.

What do the data in Table 2-1 say about the Race to the Top grant winners? Five states in the top ten of both models won Race to the Top grants—Florida, Maryland, Massachusetts, District of Columbia, and Hawaii. Delaware, another Race to the Top winner, made the top ten in Model 1 but not Model 2. It is clear that past performance was an important criterion for selection. But not a deal breaker. New York and North Carolina, states in the bottom ten of Model 2, were also grant winners. Their below-average performance on NAEP in recent years did not disqualify them from the competition. Perhaps other elements in the New York and North Carolina applications made up for that. At the same time, Kentucky, New Jersey, and Pennsylvania have excellent records of boosting student achievement, both short and long term, but received no grant money.[37]

The pattern that emerges here—some high-achieving states going unrewarded and a few winners having spotty records of achievement—parallels a Brown Center study of a decade ago evaluating the federal government's Blue Ribbon Schools program.[38] That competition, which continues today, recognizes individual schools for excellence. At the time of the study, the Blue Ribbon selection criteria included a good record of academic achievement, but only as one of a dozen criteria. The application also asked schools whether they embraced

a number of trendy educational practices. Most of the practices lacked evidence of improving student learning. Not surprisingly, the study found that about one-fourth of the schools winning Blue Ribbons scored below average for schools with similar demographic characteristics. The Blue Ribbon program was subsequently changed to elevate the importance of academic achievement.

Another highlight from Table 2-1 concerns an old saw about state rankings on tests of student achievement. In the early 1990s, Senator Daniel Patrick Moynihan published a paper showing a correlation of 0.52 between eighth-grade math scores and the distance of state capitals from the Canadian border.[39] The correlation between math scores and per pupil spending was a much weaker 0.20. Moynihan famously recommended that states wishing to improve their educational systems save their money and simply move closer to Canada. With his customary wit, Moynihan was pointing out that test scores reflect much more than the efforts of schools, or the resources provided to schools, but also the behavior of families and communities and the quality of social environments in which children are raised. States near the Canadian border exhibit a broad collection of social characteristics supporting high achievement. As noted above, all of those influences get baked in the cake of student test scores, making it important to scrutinize more than state rankings alone on test data collected at a single point in time.

Let's look at a few northern states in Table 2-1. North Dakota, Michigan, Wisconsin, and Maine all score in the bottom ten states in Model 1 despite sharing a border with Canada. Maine scored 227 on fourth-grade reading in 1992, 12 points above the 215 national average for public schools. In 2009, that difference had shrunk to only 4 points (224 versus 220). A similar trend holds for eighth-grade reading (Maine's 10-point advantage in 1998 narrowed to 6 points), fourth-grade math (a 13-point advantage in 1992 shrank to 5 points), and eighth-grade math (a 12-point lead on the national average in 1992 shrank to 4 points).

Maine's ranking slipped but was still impressive in 2009. In 1992, as Moynihan crunched those first NAEP numbers and issued his amusing geographical prescription for school reform, Maine ranked first in fourth-grade math. In 2009, Maine was tied for eighth place.

Could there be a test score ceiling holding down the initially high-scoring states? No, Massachusetts shows that it is possible for a high-scoring state to make progress. It is among the top-performing states in both models of the current study. Many people might be surprised, considering the media attention Massachusetts receives for high test scores, that it is not the top state. They forget that Massachusetts has always scored near the top on NAEP tests. In 1992, it scored 227 on fourth-grade math, 8 scale score points above the national average. And 5 points below Maine. By 2009, Massachusetts' scores had risen to 252 and Maine's to 244. Even among the top-scoring states, Massachusetts shows, there is room for improvement. And, as evidenced by Maine, there is also the possibility of disappointment.

Summary and Discussion

This study has identified states that are winning the real race to the top, that is, the race to boost student achievement. Seven states—Florida, Hawaii, Kentucky, Maryland, Massachusetts, New Jersey, and Pennsylvania—along with the District of Columbia stand out for making larger-than-expected gains in NAEP scores. The study controlled for changes in demographic characteristics of students—percentage in poverty,

Massachusetts shows that it is possible for a high-scoring state to make progress.... And, as evidenced by Maine, there is also the possibility of disappointment.

Recognizing excellence, rewarding preferred policies and programs, and compensating for disadvantage send contradictory signals on what the system values and how scarce resources will be shared.

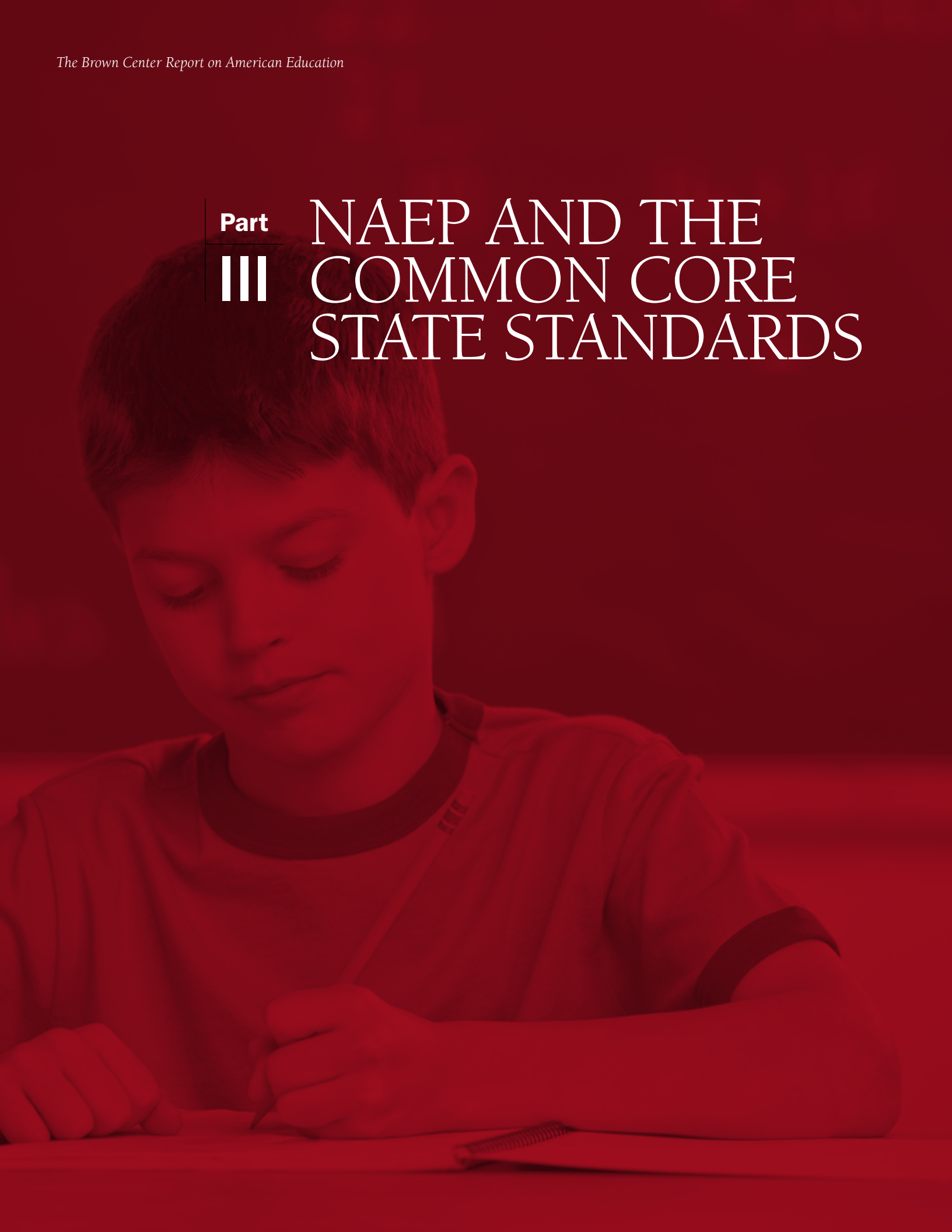
English language learners, and special education—and examined both long- and short-term gains. Four states (Iowa, Michigan, Nebraska, and West Virginia) stand out for underperforming.

Two lessons can be gleaned from the study. The most important is a reminder that NAEP scores must be interpreted carefully. Determining which states are doing well and which are doing poorly requires more than a glance at the latest rankings. Massachusetts has done well since the 1990s. That accomplishment is worthy of accolades. But Maryland, the District of Columbia, and Florida have done just as well, if not better. Their accomplishments are overlooked because these states are ranked lower on state NAEP scores. Many of the states bordering with Canada, famously recognized by Senator Daniel Patrick Moynihan in the early 1990s, have subsequently made slower progress than many other states. Rankings obscure their tepid gains because influences other than school performance go into test data collected at any single point in time. These states enjoy social environments that support student achievement.

The second lesson addresses contradictions in the educational system's incentive structure, contradictions that are reflected in Race to the Top. Rewards are a mainstay of American schooling. The title "Race to the Top" is a metaphor that alludes to recognizing and rewarding extraordinary accomplishments in student learning, an objective with broad political appeal. But that aim is in tension with other popular criteria for distributing resources. The Race to the Top program also rewards states for pursuing policies favored by the administration. Whether this is a reward for engaging in promising practices (if you believe in the policies) or simply a political payoff for preferred behavior (if you are skeptical), grants such as Race to the Top are one of the few tools in the possession of upper-level policymakers to shape the system below. Compensatory grants are yet another reason for granting revenue—giving additional funds to local districts based on need. These three distributional criteria—rewards for real accomplishments, inducements for particular behaviors, and compensation for disadvantage—are very different.

The point here is not to endorse one criterion over the others. The argument is that they do not cohere as a system of incentives. Recognizing excellence, rewarding preferred policies and programs, and compensating for disadvantage send contradictory signals on what the system values and how scarce resources will be shared. Far from settling such inconsistencies, democratic institutions—voters, school boards, state legislatures, Congress, and state and federal departments—probably deepen and prolong them. Persuasive arguments can be made for programs based on each of these objectives, and Race to the Top is an excellent example of how even a single program may simultaneously embrace more than one. The casual observer should not be fooled, however, into thinking that the states that won the Race to the Top are the states winning the real race to improve student learning.

Part III NAEP AND THE COMMON CORE STATE STANDARDS



UNLIKE MOST COUNTRIES, THE UNITED STATES DOES NOT have national education standards, no single set of expectations for what all American teachers should teach and all American students should learn. It never has. A question that the rest of the world considers foundational to its national school systems—deciding the content of the curriculum—sits in the hands of local authorities. That is because the United States has 50 state school systems. Heterogeneity extends to the deepest levels of schooling. Even students transferring from one teacher to another within the same school may, as a consequence, learn a different curriculum than their former classmates.

So it was an historical event when the Common Core State Standards in mathematics and reading were released in June 2010. Launched by the National Governors Association and the Council of Chief State School Officers, the Common Core Standards project brought together experts in both reading and math to develop a set of standards that would be, in what became a mantra, both "higher and fewer in number" than existing state standards.[40] The standards are voluntary—states choose whether to participate—but for the first time most American students will study a uniform curriculum through at least the eighth grade. A draft of the experts' work circulated for several months, and, based on input from other experts and the general public, the standards were finalized.[41] In September 2010, two consortia were awarded federal grants totaling $330 million to develop annual assessments aligned with the Common Core standards, and as of December 2010, 43 states and the District of Columbia have signed on to those efforts.[42] The tests are due to be given for the first time in the 2014–2015 school year.[43]

The nation currently monitors the math achievement of fourth, eighth, and twelfth graders on the National Assessment of Educational Progress (NAEP).[44] Since 1990, the main NAEP has assessed mathematics proficiency in five content strands—number properties and operations, algebra, geometry, measurement, and data analysis/

statistics/probability.[45] How well does NAEP match up with the Common Core standards in mathematics?

We tackled this question by analyzing NAEP items from the eighth-grade assessment. NAEP items are periodically released to the public to give an idea of the content of the test. For the current study, we coded all public release items from the algebra and number strands[46] based on the grade at which the Common Core recommends teaching the mathematics assessed by the item. The 2009 NAEP Framework in Mathematics calls for number and algebra items to comprise half of the eighth-grade assessment.[47] A total of 171 items were available, 98 from the number strand and 73 from algebra.[48] We were unable to code four items (two from each strand) because they assess skills not found in the Common Core.

A precursor to this study can be found in the 2004 Brown Center Report.[49] In that study, we coded the grade level of public release items labeled as "problem solving," one of NAEP's process strands (different from the content strands). Only problems involving the application of arithmetic were analyzed. At what grade level are students taught the arithmetic required to answer NAEP problem-solving items? We discovered that the mean fourth-grade NAEP item registered at 3.2 and the mean eighth-grade item at 3.7, suggesting that the typical item could be answered using arithmetic taught by the end of third grade. Primarily, this finding stems from NAEP's reliance on whole number arithmetic in word problems. We found that approximately 70 percent of the eighth-grade items focused on whole numbers. Problems with fractions, decimals, or percents—forms of rational numbers taught after third grade—are not common on NAEP.[50]

The 2004 study used the Singapore Math program as a rubric to code the grade level of items, assigning a value according to the grade and semester in which the arithmetic of the item was taught. By using the Common Core and evaluating the entire context of items, the current study's rubric produces higher grade-level estimates for items. Problems involving only simple arithmetic are classified at a higher grade level if they are posed in the context of more sophisticated topics that are taught at a later grade (e.g., coordinate plane, equations with two variables). Selected NAEP items are shown below.

Grade Level of NAEP Items in the Common Core (Number, 8th-Grade Test)

Table 3-1

	Number		
Grade	**Total (N)**	**Calculator (N)**	**Average Percent Correct**
2	1	0	64.0%
3	9	0	79.3%
4	27	4	72.1%
5	17	7	61.2%
6	23	12	53.4%
7	15	8	37.5%
8	6	4	31.5%
TOTAL	**98**	**35**	**58.6%**

Note: Mean grade level: 5.2, median grade level: 5

Findings

Table 3-1 displays data on items from the number strand. In terms of grade level on the Common Core, the items assess mathematics found at second through eighth grades. The number strand of the eighth-grade NAEP is best described as pitched at the fifth-grade level if calibrated by the Common Core. The average grade level for the items is 5.2. The median item also registers at the fifth-grade level, meaning that about half of the items cover material from the fifth grade or earlier and half from the fifth grade or later. More than 90 percent of the items (92 out of 98) cover material

More than 90 percent of the items cover material below the eighth grade.

Grade Level of NAEP Items in the Common Core (Algebra, 8th-Grade Test)

Table 3-2

Algebra			
Grade	**Total (N)**	**Calculator (N)**	**Average Percent Correct**
2	1	0	86.0%
3	0	0	–
4	6	0	74.2%
5	8	1	60.5%
6	27	8	54.4%
7	16	9	48.7%
8	15	3	45.4%
Total	**73**	**21**	**54.0%**

Note: Mean grade level: 6.3, median grade level: 6

below the eighth grade. Note that this does not make the test easy for eighth graders. The average item is answered correctly by 58.6 percent of eighth graders nationally, and for items pitched at the sixth-grade level and later, the percentage answering correctly is only 45.0 percent.

Calculators are an interesting factor. According to the NAEP framework, calculators are provided to students on approximately one-third of the eighth-grade test.[51] As indicated in Table 3-1, the number items in public release reflect a similar proportion, with 35.7 percent involving a calculator. Calculators are more likely to be provided on items with content from higher grades (sixth grade and above) than from lower grades. About half of the items coded as sixth to eighth grades allow calculators, compared with one-fifth of the items from earlier grades. The more advanced the grade level of the NAEP item, the more likely that a calculator is allowed.

Table 3-2 presents data on algebra items. They appear about one grade more challenging than number items, with a mean grade level of 6.3 and median of sixth grade. Performance on the algebra items is similar to that in the number strand. The average item is answered correctly by 54.0 percent of students. Performance on items encompassing material from the sixth to eighth grades averages 50.5 percent. And, again, calculators tend to be provided on items from higher rather than lower grades.

Frankly, most of the skills measured in the algebra strand, especially those appearing before eighth grade in the Common Core, assess algebraic reasoning, not content from a formal algebra course.

Let's examine a few problems considered "algebra" on NAEP.

Sample NAEP Items

One of the items from the algebra strand is coded at the second-grade level. What does second-grade algebra look like? Here is the item:

Block M5, Question 6 (2009)

□ – 8 = 21

What number should be put in the box to make the number sentence above true?

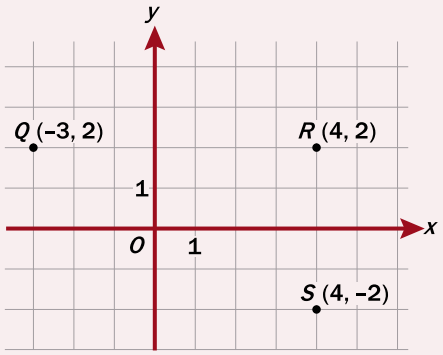
Answer: (29)

The item was answered correctly by 85.6 percent of eighth graders. It is almost a first-grade item. In first grade, the Common Core recommends problem solving with addition and subtraction using numbers within 20. The skill is extended to numbers within 100 in second grade,[52] as noted here:

> Use addition and subtraction within 100 to solve one- and two-step word problems involving situations of adding to, taking from, putting together, taking apart, and comparing, with unknowns in all positions, e.g., by using drawings and equations with a symbol for the unknown number to represent the problem. (Page 19, Operations and Algebraic Thinking 2.0A)

A more difficult item is:

Block M3, Question 13 (2005)



If the points *Q*, *R*, and *S* shown above are three of the vertices of rectangle *QRST*, which of the following are the coordinates of *T* (not shown) ?

A. (4, –3)
B. (3, –2)
C. (–3, 4)
D. (–3, –2)
E. (–2, –3)

This is a sixth-grade problem. It was answered correctly by 60.0 percent of eighth graders.

One must know something about a rectangle (that opposite sides are parallel and equal in length) and some basic knowledge of coordinates—in this case, that *T* will have the x value of *Q* and the y value of *S*. The coordinate plane is introduced in fifth grade, but initially students work with only the first quadrant and learn how to locate individual points. In sixth grade, the Common Core extends study to all four quadrants, incorporates the construction of polygons, and recommends teaching the following skills:

> Draw polygons in the coordinate plane given coordinates for the vertices; use coordinates to find the length of a side joining points with the same first coordinate or the same second coordinate. Apply these techniques in the context of solving real-world and mathematical problems. (Page 45, Geometry 6.G)

A still more difficult problem follows:

Block M6, Question 27 (2003)

X	Y
0	–3
1	–1
2	1

Which of the following equations is true for the three pairs of *x* and *y* values in the table above?

A. $3x + 2 = y$
B. $3x - 2 = y$
C. $2x + 3 = y$
D. $2x - 3 = y$
E. $x - 3 = y$

The item was answered correctly by 45 percent, incorrectly by 52 percent, and was omitted by 3 percent. The item was difficult to code using the Common Core. We wound up labeling it as recommended for eighth grade. An eighth-grade standard exists that is close to capturing the above task, but the standard demands a more complex understanding of functions:

> Determine the rate of change and initial value of the function from a description of a relationship or from two (*x*, *y*) values, including reading these from a table or from a graph. (Page 55, Functions 8.F)

The item does not require students to calculate rate of change, although this is an elementary linear equation with a slope first-year algebra students would be expected to calculate. The task is to identify a simple

The coordinate plane is introduced in fifth grade, but initially students work with only the first quadrant…

…the Common Core embraces the notion that students can learn to see algebra as generalized arithmetic if they are taught concepts sequentially…

two-variable equation that matches a table of values. Students who use a "plug and chug" strategy with the first pair (0, –3) will eliminate A, B, and C, thereby narrowing potentially the correct answer to D or E. The second pair (1, –1) eliminates E and leaves D as the only possible correct answer. Supporting the theory that plug and chug is a popular approach, E is the incorrect item most often selected (17 percent), but not by much—C (15 percent), B (14 percent), and A (7 percent). Students who plug and chug using the third pair (2, 1) will arrive at the correct answer in one step.

The next item was easier to classify, but not for students to answer:

Block M12, Question 3 (2005)

Which of the following is equal to 6 (x + 6) ?

A. x + 12
B. 6x + 6
C. 6x + 12
D. 6x + 36
E. 6x + 66

This is a sixth-grade item. It was answered correctly by 44 percent of eighth graders. It assesses the understanding of an important concept, the distributive property of multiplication over addition:

> Apply the properties of operations to generate equivalent expressions. For example, apply the distributive property to the expression 3 (2 + x) to produce the equivalent expression 6 + 3x; apply the distributive property to the expression 24x + 18y to produce the equivalent expression 6 (4x + 3y); apply properties of operations to y + y + y to produce the equivalent expression 3y. (Page 44, Expressions and Equations, 6.EE [italics omitted])

Students first encounter the distributive property in third and fourth grades as they learn multiplication with whole numbers, but, as this standard illustrates, the concept is generalized to include unknowns in sixth grade. When it comes to properties, the Common Core embraces the notion that students can learn to see algebra as generalized arithmetic if they are taught some of the structure behind arithmetic operations that will later be used in algebra and if care is given to developing fluency with numbers and engaging students in a variety of applications.

Summary and Discussion

This study coded the grade level of 171 items from the number and algebra strands of the eighth-grade NAEP test. The Common Core standards in math were used as the coding rubric. Items from the number strand range from the first to eighth grade, with a median level of fifth grade and mean of 5.2. Items from the algebra strand range from the second to eighth grades, with a median level of sixth grade and mean of 6.3. In both strands, calculators are provided about 33 percent of the time overall and 2½ times more often on items from upper grades (sixth through eighth) compared with items from lower grades.

Sample NAEP algebra items were presented. The items would all come from a pre-algebra course or earlier in a student's mathematics education and would not be part of a formal algebra course. The items support two criticisms. Critics of the Common Core have complained that the eighth-grade standards reflect mathematics learned prior to algebra, undermining the contemporary movement to provide "algebra for all" in eighth grade.[53] Critics of NAEP have similarly pointed out that the eighth-grade assessment contains problems

called "algebra" that are in fact pre-algebra in origin.[54]

The Common Core and NAEP share common ground—and some would say a common weakness—in how they test algebra. And yet they seem to diverge on the crucial question of content. The public release items of the eighth-grade NAEP are, on average, two to three years below the eighth-grade mathematics recommended by the Common Core.

The discrepancy arises because of varying definitions of an "eighth-grade" math test. Two very different models are in play. One kind, which NAEP typifies, assesses all of the mathematics learned through eighth grade. Eighth-grade skills and knowledge are the most difficult content on such a test, but they comprise a portion of the items, perhaps corresponding to only a single grade's share of the K–8 grade span. Consequently, the average item on NAEP registers significantly below eighth grade, and approximately 90 percent of eighth-grade NAEP items are taught before that grade. Several items on the eighth-grade NAEP test are also on the fourth-grade NAEP test.

The second model is an eighth-grade test that assesses what is learned in eighth grade. The tests keyed to the Common Core appear to be heading in that direction. Tests will be administered at each grade level in grades 3–8 and reflect the skills and knowledge that the Common Core recommends for that particular grade. Items out of grade level, either below or above, may be included but are rare on such a test. The average item falls near the middle of the grade being tested. End-of-course exams and Advanced Placement (AP) tests are examples of this kind of test, although anchored to a particular course rather than grade level. AP tests are not interested in what a student learned in fifth grade. Nor will eighth-grade Common Core tests be interested in such content. That will be the job of the fifth-grade test.

Both models can legitimately be called an "eighth-grade test," and yet they assess different mathematics. Consider floors and ceilings. The first (NAEP) has a low floor (a primary grade) and tight ceiling (end of eighth grade) and assesses several years of mathematics curriculum. The second (Common Core) has a high floor (beginning of eighth grade) and tight ceiling (end of eighth grade) and assesses a single year's curriculum.

So what does the future hold for NAEP and the Common Core? As currently planned, the two programs will assess different mathematics and might report different results. Even if they report similar results, each score will reveal something different about American students' math skills. Bear in mind that the two programs serve different purposes. NAEP is a survey. It is "top-down" and draws a random sample of students from which inferences are drawn. It monitors national and state progress and, except for several large urban districts, reports no score below the state level. The Common Core, meanwhile, will be "bottom-up," testing all students. It promises to produce student-level scores that can be aggregated to yield performance measures for classes, schools, districts, and states—even a national score if all states eventually participate. It also can generate data during the school year, providing useful feedback to teachers on the effectiveness of instruction and curricular materials.

In the beginning, the two programs will overlap in issuing state scores. And that could cause confusion, especially in states receiving contradictory signals from the two tests about their students' performance. Factor in the confusion from reporting the

The discrepancy arises because of varying definitions of an "eighth-grade" math test. Two very different models are in play.

Much work remains to bring the Common Core standards to life in a real assessment.

percentage of students performing at different levels (i.e., basic, proficient, advanced, and the like) on tests of vastly different content, and conflicts are bound to arise. One option is to ratchet up the difficulty of NAEP items, bringing the test in harmony with the Common Core. That could merely achieve test redundancy, however, and lead some to question the necessity of continuing one or the other program.

Another possibility is that adaptive testing will bridge the chasm between the two tests. Adaptive testing delivers computer-based assessments. It enhances the capability of delivering items that are sensitive to students' individual achievement profiles and would expand the scales of both assessments by including more lower level and advanced items. While taking the same test, struggling math students can get items that are below grade level and precocious math students can get items more suited to their advanced standing. If it becomes a feature of both assessments, adaptive testing may bring NAEP and the Common Core assessments in closer alignment.

Of course, all of this admittedly is crystal ball gazing. Much work remains to bring the Common Core standards to life in a real assessment. Once that happens, an education process will be needed that informs the public and political leaders on what the NAEP and Common Core measure, what they have in common, and what differentiates their results. A similar challenge exists with the main and long-term trend NAEP assessments, but unfortunately, even after 20 years of shared history between these two tests, very few observers who comment on their results in the press seem aware of the tests' key differences. The same is true of the main NAEP and state assessments.

A new era is dawning for NAEP. The program has supplied the nation with progress reports on student learning since 1969. Now, Common Core assessments are on the way. Whether the new assessments push NAEP aside, succeed in augmenting the information provided by NAEP, or force a redefinition of NAEP's role in monitoring student learning will be at the top of the NAEP policy agenda in the years ahead.

NOTES

1 *Assessing Scientific, Reading, and Mathematical Literacy: A Framework for PISA 2006* (OECD, 2006) p. 11.

2 H. L. Fleischman, P. J. Hopstock, M. P. Pelczar, and B. E. Shelley, *Highlights From PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context* (NCES 2011–004), U.S. Department of Education, National Center for Education Statistics (Washington, DC: U.S. Government Printing Office, 2010).

3 To put in context, in 2009 the international average was 496 and the standard deviation was 100.

4 Eric A. Hanushek and Ludger Woessmann, *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes* (OECD, 2010).

5 Sam Dillon, "Top Test Scores From Shanghai Stun Educators," *New York Times*, December 7, 2010, p. A1.

6 Id.

7 Id.

8 Peter Gumbel, "China Beats Out Finland for Top Marks in Education," *Time*, December 7, 2010.

9 Amanda Paulson, "US Students Halt Academic 'Free-Fall,' but Still Lag in Global Testing," *Christian Science Monitor*, December 7, 2010.

10 Pat Wingert, "Good Thing Kids Can't Vote: Obama Backs the Idea of a Longer School Year," *Newsweek*, September 27, 2010.

11 A pilot study was conducted prior to FIMS.

12 *International Study of Achievement in Mathematics: A Comparison of Twelve Countries (Vols. 1–2)*, edited by T. Husén (New York: John Wiley & Sons, 1967).

13 "Brief History of IEA," (http://www.iea.nl/brief_history_of_iea.html).

14 Linda Darling-Hammond, *The Flat World and Education: How America's Commitment to Equity Will Determine Our Future* (New York: Teachers College Press, 2010).

15 Sean Cavanagh, "Poverty's Effect on U.S. Scores Greater Than for Other Nations," *Education Week*, December 12, 2007, pp. 1, 13.

16 Carol C. Burris, Kevin G. Welner, and Jennifer W. Bezoza, *Universal Access to a Quality Education: Research and Recommendations for the Elimination of Curricular Stratification* (Boulder, CO: Education and the Public Interest Center & Education Policy Research Unit, 2009).

17 Ina V. S. Mullis et al., *TIMSS 1999 International Mathematics Report* (Boston: The International Study Center and The International Association for the Evaluation of Educational Achievement, 2000).

18 Original data for Finland are reported from FIMS. After publication of the Finnish results, it was discovered that scores from Population 1B students attending elementary schools had been omitted. Including them appears to lower the national average; see unweighted data in Appendix III in Volume II of *International Study of Achievement in Mathematics: A Comparison of Twelve Countries (Vols. 1–2)*, edited by T. Husén (New York: John Wiley & Sons, 1967).

19 See Jan-Eric Gustafsson, "Understanding Causal Influences on Educational Achievement through Analysis of Differences over Time within Countries," in *Lessons Learned: What International Assessments Tell Us about Math Achievement*, edited by Tom Loveless (Washington, DC: The Brookings Institution Press, 2007), on age-grade effects on achievement.

20 Unlike in large countries, a huge proportion of a small country's students must take the test to attain statistical power.

21 OECD, *Measuring Student Knowledge and Skills: A New Framework for Assessment* (OECD, 1999).

22 Jan de Lange, "Mathematics for Literacy," in *Quantitative Literacy: Why Numeracy Matters for Schools and Colleges*, edited by Bernard L. Madison and Lynn Arthur Steen (Princeton, NJ: National Council on Education and the Disciplines, 2003), p. 80.

23 See Learning and Technology World Forum, "Re-Imagining Education: Andreas Schleicher Keynote Presentation," 2009 (http://www.latwf.org/en-gb/About/2009-Highlights1/Video-hightlights/Andreas-Schleicher/)

24 For example, see Tony Wagner, *The Global Achievement Gap: Why Even Our Best Schools Don't Teach the New Survival Skills Our Children Need—And What We Can Do About It* (New York: Basic Books, 2008).

25 Jouni Välijärvi, Pirjo Linnakylä, Pekka Kupari, Pasi Reinikainen, and Inga Arffman, *The Finnish Success in PISA—And Some Reasons Behind It* (Finland, Institute for Educational Research, University of Jyväskylä, 2002).

26 Kyösti Tarvainen and Simo K. Kivelä, "Severe Shortcomings in Finnish Mathematics Skills," *Helsingin Sanomat*, March 10, 2005.

27 L. Näveri, "Understanding Computations," *Dimensio 3* (2005), pp. 49–52 (in Finnish). Cited and discussed in "Mathematics Curriculum Development in Finland—Unexpected Effects," Olli Martio, University of Helsinki.

28 Selected precincts in China have participated in past assessments, but a nationally representative sample has not been tested.

29 "Basic Facts, Science and Education, Regular Education" (www.shanghai.gov.cn/shanghai/node23919/node24059/). Also see *Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States* (OECD, 2010).

30 Ina V. S. Mullis et al., *Mathematics Benchmarking Report TIMSS 1999—Eighth Grade* (IEA, 2001).

31 NCES, U.S. Department of Education, *Reading 2009: National Assessment of Educational Progress at Grades 4 and 8* (NCES 2010–458). (U.S. Department of Education, 2010).

32 See the NCES website on the Common Core of Data, http://nces.ed.gov/ccd/bat/.

33 Harold Wenglinsky, *Does It Compute? The Relationship Between Educational Technology and Student Achievement in Mathematics* (Princeton, NJ: ETS, 1998).

34 One important benefit of random assignment is that unmeasured characteristics should be equally distributed between treatment and control groups.

35 The year prior to NAEP participation is used as the start point.

36 See Sharif M. Shakrani and Edward Roeber, *Assessment Accommodations for Students with Disabilities and English Language Learners Used by States and NAEP* (National Assessment Governing Board, 2009) and David Grissmer, *Improving NAEP for Research and Policymaking* (National Assessment Governing Board, 2002).

37 Sean Cavanagh, "N.J. Schools Chief Fired Over Race to Top Gaffe," *Education Week*, September 1, 2010, p. 16.

38 Tom Loveless, *The Brown Center Report on American Education: How Well Are American Students Learning?* (Washington: The Brookings Institution, 2000), pp. 26–30.

39 Daniel Patrick Moynihan, *Miles to Go: A Personal History of Social Policy*. (Cambridge, MA: Harvard University Press, 1996), p. 148.

40 "Common Core State Standards Development Work and Feedback Group Announced," News Release (Washington: National Governors Association Center for Best Practices and the Council of Chief State School Officers, July 1, 2009).

41 See the Common Core State Standards Initiative website on About the Standards, http://www.corestandards.org/about-the-standards.

42 Catherine Gewertz, "Common-Standards Watch: South Dakota Makes 44," *Curriculum Matters*, Education Week, November 29, 2010.

43 "Beyond the Bubble Tests: The Next Generation of Assessments," Prepared Remarks of U.S. Secretary of Education Arne Duncan to State Leaders at Achieve's American Diploma Project (ADP) Leadership Team Meeting, Alexandria, VA, September 2, 2010.

44 The long-term trend NAEP test assesses students at ages 9, 13, and 17.

45 See the NCES website on the NAEP Mathematics Framework, http://nces.ed.gov/nationsreportcard/mathematics/whatmeasure.asp.

46 The number strand refers to the number sense, properties, and operations strand for the 1990–2003 NAEP mathematics framework and the number properties and operations strand in the current mathematics framework.

47 National Assessment Governing Board, U.S. Department of Education, *Mathematics Frameworks for the 2009 National Assessment of Educational Progress* (Washington: 2008).

48 See the NAEP Questions Tool, http://nces.ed.gov/nationsreportcard/itmrlsx/landing.aspx.

49 Tom Loveless, *The 2004 Brown Center Report on American Education: How Well Are American Students Learning?* (Washington: The Brookings Institution, 2004), pp. 5–17.

50 Theresa Smith Neidorf and others, *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006–029). (U.S. Department of Education, 2006).

51 The NAEP is organized by blocks of items, with calculators allowed on one-third of blocks.

52 The Common Core also allows for mentally adding or subtracting 10 with numbers within 100.

53 Catherine Gewertz, "Draft Common Standards Elicit Kudos and Criticism," *Education Week*, March 17, 2010, pp. 1, 14–15.

54 See David Klein, "What Do the NAEP Math Tests Really Measure?" *Notices of the AMS*, January 2011, pp. 53–55. Also see *2004 Brown Center Report*.

THE BROOKINGS INSTITUTION

STROBE TALBOTT
President

DARRELL WEST
Vice President and Director
Governance Studies Program

BROWN CENTER STAFF

GROVER "RUSS" WHITEHURST
Senior Fellow and Director

TOM LOVELESS
Senior Fellow

MICHELLE CROFT
Research Analyst

MATTHEW CHINGOS
Non-resident Fellow

PAUL T. HILL
Non-resident Senior Fellow

DIANE RAVITCH
Non-resident Senior Fellow

Views expressed in this report are solely those of the author.

BROWN CENTER on
Education Policy
at BROOKINGS

BROOKINGS
1775 Massachusetts Avenue, NW • Washington, D.C. 20036
Tel: 202–797–6000 • Fax: 202–797–6004
www.brookings.edu

The Brown Center on Education Policy
Tel: 202–797–6090 • Fax: 202–797–2480
www.brookings.edu/brown