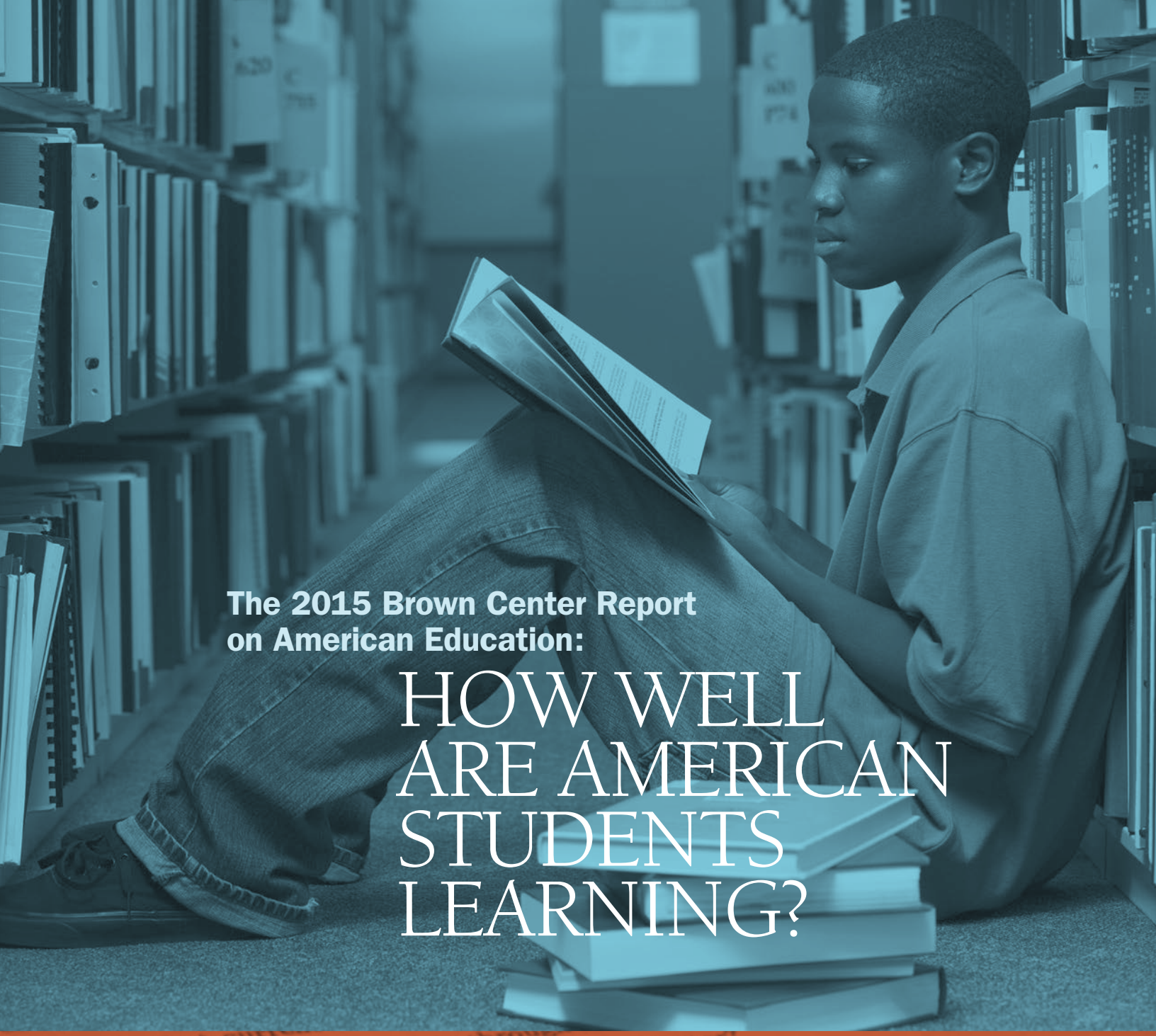


The 2015 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on the gender gap in reading, effects of the Common Core, and student engagement

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/about/centers/brown.

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

The 2015 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on the gender gap in reading, effects of the Common Core, and student engagement

March 2015
Volume 3, Number 4

by:
TOM LOVELESS
Nonresident Senior Fellow, The Brown Center on Education Policy,
The Brookings Institution

TABLE OF CONTENTS

3 Introduction

PART I
8 Girls, Boys, and Reading

PART II
18 Measuring Effects of the Common Core

PART III
26 Student Engagement

36 Notes

Data verification by: Katharine Lindquist

THE 2015 BROWN CENTER REPORT ON AMERICAN EDUCATION

The 2015 Brown Center Report (BCR) represents the 14th edition of the series since the first issue was published in 2000. It includes three studies. Like all previous BCRs, the studies explore independent topics but share two characteristics: they are empirical and based on the best evidence available. The studies in this edition are on the gender gap in reading, the impact of the Common Core State Standards – English Language Arts on reading achievement, and student engagement.

Part one examines the gender gap in reading. Girls outscore boys on practically every reading test given to a large population. And they have for a long time. A 1942 Iowa study found girls performing better than boys on tests of reading comprehension, vocabulary, and basic language skills. Girls have outscored boys on every reading test ever given by the National Assessment of Educational Progress (NAEP)—the first long term trend test was administered in 1971—at ages nine, 13, and 17. The gap is not confined to the U.S. Reading tests administered as part of the Progress in International Reading Literacy Study (PIRLS) and the Program for International Student Assessment (PISA) reveal that the gender gap is a worldwide phenomenon. In more than sixty countries participating in the two assessments, girls are better readers than boys.

Perhaps the most surprising finding is that Finland, celebrated for its extraordinary performance on PISA for over a decade, can take pride in its high standing on the PISA reading test solely because of the performance of that nation's young women. With its 62 point gap, Finland has the largest gender gap of any PISA participant, with girls scoring 556 and boys scoring 494 points (the OECD average is 496, with a standard deviation of 94). If Finland were only a nation of young men, its PISA ranking would be mediocre.

Part two is about reading achievement, too. More specifically, it's about reading and the English Language Arts standards of the Common Core (CCSS-ELA). It's also about an important decision that policy analysts must make when evaluating public policies—the determination of when a policy begins. How can CCSS be properly evaluated?

Two different indexes of CCSS-ELA implementation are presented, one based on 2011 data and the other on data collected in 2013. In both years, state education officials were surveyed about their Common Core implementation efforts. Because forty-six states originally signed on to the CCSS-ELA—and with at least forty still on track for full implementation by 2016—little variability exists among the states in terms of standards policy. Of course, the four states that never adopted CCSS-ELA can serve as a small control group. But variation is also found in how the states are implementing CCSS. Some states are pursuing an array of activities and aiming for full implementation earlier rather than later. Others have a narrow, targeted implementation strategy and are proceeding more slowly.

The analysis investigates whether CCSS-ELA implementation is related to 2009-2013 gains on the fourth grade NAEP reading test. The analysis cannot verify causal relationships between the two variables, only correlations. States that have aggressively implemented CCSS-ELA (referred to as "strong" implementers in the study) evidence a one to one and one-half point larger gain on the NAEP scale compared to non-adopters of the standards. This association is similar in magnitude to an advantage found in a study of eighth grade math achievement in last year's BCR. Although positive, these effects are quite small. When the 2015 NAEP results are released this winter, it will be important for the fate of the Common Core project to see if strong implementers of the CCSS-ELA can maintain their momentum.

Part three is on student engagement. PISA tests fifteen-year-olds on three subjects—reading, math, and science—every three years. It also collects

a wealth of background information from students, including their attitudes toward school and learning. When the 2012 PISA results were released, PISA analysts published an accompanying volume, *Ready to Learn: Students' Engagement, Drive, and Self-Beliefs,* exploring topics related to student engagement.

Part three provides secondary analysis of several dimensions of engagement found in the PISA report. Intrinsic motivation, the internal rewards that encourage students to learn, is an important component of student engagement. National scores on PISA's index of intrinsic motivation to learn mathematics are compared to national PISA math scores. Surprisingly, the relationship is negative. Countries with highly motivated kids tend to score lower on the math test; conversely, higher-scoring nations tend to have less-motivated kids.

The same is true for responses to the statements, "I do mathematics because I enjoy it," and "I look forward to my mathematics lessons." Countries with students who say that they enjoy math or look forward to their math lessons tend to score lower on the PISA math test compared to countries where students respond negatively to the statements. These counterintuitive finding may be influenced by how terms such as "enjoy" and "looking forward" are interpreted in different cultures. Within-country analyses address that problem. The correlation coefficients for within-country, student-level associations of achievement and other components of engagement run in the anticipated direction—they are positive. But they are also modest in size, with correlation coefficients of 0.20 or less.

Policymakers are interested in questions requiring analysis of aggregated data—at the national level, that means between-country data. When countries increase their students' intrinsic motivation to learn math, is there a concomitant increase in PISA math scores? Data from 2003 to 2012 are examined. Seventeen countries managed to increase student motivation,

but their PISA math scores fell an average of 3.7 scale score points. Fourteen countries showed no change on the index of intrinsic motivation—and their PISA scores also evidenced little change. Eight countries witnessed a decline in intrinsic motivation. Inexplicably, their PISA math scores increased by an average of 10.3 scale score points. Motivation down, achievement up.

Correlation is not causation. Moreover, the absence of a positive correlation—or in this case, the presence of a negative correlation—is not refutation of a possible positive relationship. The lesson here is not that policymakers should adopt the most effective way of stamping out student motivation. The lesson is that the level of analysis matters when analyzing achievement data. Policy reports must be read warily—especially those freely offering policy recommendations. Beware of analyses that exclusively rely on within- or between-country test data without making any attempt to reconcile discrepancies at other levels of analysis. Those analysts could be cherry-picking the data. Also, consumers of education research should grant more credence to approaches modeling change over time (as in difference in difference models) than to cross-sectional analyses that only explore statistical relationships at a single point in time.

Part I GIRLS, BOYS, AND READING

GIRLS SCORE HIGHER THAN BOYS ON TESTS OF READING

Gability. They have for a long time. This section of the Brown Center Report assesses where the gender gap stands today and examines trends over the past several decades. The analysis also extends beyond the U.S. and shows that boys' reading achievement lags that of girls in every country in the world on international assessments. The international dimension—recognizing that U.S. is not alone in this phenomenon—serves as a catalyst to discuss why the gender gap exists and whether it extends into adulthood.

Background

One of the earliest large-scale studies on gender differences in reading, conducted in Iowa in 1942, found that girls in both elementary and high schools were better than boys at reading comprehension.[1] The most recent results from reading tests of the National Assessment of Educational Progress (NAEP) show girls outscoring boys at every grade level and age examined. Gender differences in reading are not confined to the United States. Among younger children—age nine to ten, or about fourth grade—girls consistently outscore boys on international assessments, from a pioneering study of reading comprehension conducted in fifteen countries in the 1970s, to the results of the Program in International Reading Literacy Study (PIRLS) conducted in forty-nine nations and nine benchmarking entities in 2011. The same is true for students in high school. On the 2012 reading literacy test of the Program for International Student Assessment (PISA), worldwide gender gaps are evident between fifteen-year-old males and females.

As the 21st century dawned, the gender gap came under the scrutiny of reporters and pundits. Author Christina Hoff Sommers added a political dimension to the gender gap, and some say swept the topic into the culture wars raging at the time, with her 2000 book *The War Against Boys: How Misguided Feminism is Harming Our Young Men*.[2] Sommers argued that boys' academic inferiority, and in particular their struggles

with reading, stemmed from the feminist movement's impact on schools and society. In the second edition, published in 2013, she changed the subtitle to *How Misguided Policies Are Harming Our Young Men*. Some of the sting is removed from the indictment of "misguided feminism." But not all of it. Sommers singles out for criticism a 2008 report from the American Association of University Women.[3] That report sought to debunk the notion that boys fared poorly in school compared to girls. It left out a serious discussion of boys' inferior performance on reading tests, as well as their lower grade point averages, greater rate of school suspension and expulsion, and lower rate of acceptance into college.

Journalist Richard Whitmire picked up the argument about the gender gap in 2010 with *Why Boys Fail: Saving Our Sons from an Educational System That's Leaving Them Behind*.[4] Whitmire sought to separate boys' academic problems from the culture wars, noting that the gender gap in literacy is a worldwide phenomenon and appears even in countries where feminist movements are weak to nonexistent. Whitmire offers several reasons for boys' low reading scores, including poor reading instruction (particularly a lack of focus on phonics), and too few books appealing to boys' interests. He also dismisses several explanations that are in circulation, among them, video games, hip-hop culture, too much testing, and feminized classrooms. As with Sommers's book, Whitmire's culprit can be found in the subtitle: the educational system. Even if the educational system is not the original source of the problem, Whitmire argues, schools could be doing more to address it.

In a 2006 monograph, education policy researcher Sara Mead took on the idea that American boys were being shortchanged by schools. After reviewing achievement data from NAEP and other tests, Mead concluded that the real story of the gender gap wasn't one of failure at all. Boys and girls were both making solid academic progress, but in some cases, girls were making larger gains, misleading some commentators into concluding that boys were being left behind. Mead concluded, "The current boy crisis hype and the debate around it are based more on hopes and fears than on evidence."[5]

Explanations for the Gender Gap

The analysis below focuses on where the gender gap in reading stands today, not its causes. Nevertheless, readers should keep in mind the three most prominent explanations for the gap. They will be used to frame the concluding discussion.

Biological/Developmental: Even before attending school, young boys evidence more problems in learning how to read than girls. This explanation believes the sexes are hardwired differently for literacy.

School Practices: Boys are inferior to girls on several school measures—behavioral, social, and academic—and those discrepancies extend all the way through college. This explanation believes that even if schools do not create the gap, they certainly don't do what they could to ameliorate it.

Cultural Influences: Cultural influences steer boys toward non-literary activities (sports, music) and define literacy as a feminine characteristic. This explanation believes cultural cues and strong role models could help close the gap by portraying reading as a masculine activity.

Girls outscore boys in reading at every grade level and age.

The gender gap is widest in adolescence.

The U.S. Gender Gap in Reading

Table 1-1 displays the most recent data from eight national tests of U.S. achievement. The first group shows results from the National Assessment of Educational Progress Long Term Trend (NAEP-LTT), given to students nine, 13, and 17 years of age. The NAEP-LTT in reading was first administered in 1971. The second group of results is from the NAEP Main Assessment, which began testing reading achievement in 1992. It assesses at three different grade levels: fourth, eighth, and twelfth. The last two tests are international assessments in which the U.S. participates, the Progress in International Reading Literacy Study (PIRLS), which began in 2001, and the Program for International Student Assessment (PISA), first given in 2000. PIRLS tests fourth graders, and PISA tests 15-year-olds. In the U.S., 71 percent of students who took PISA in the fall of 2012 were in tenth grade.

Two findings leap out. First, the test score gaps between males and females are statistically significant on all eight assessments. Because the sample sizes of the assessments are quite large, statistical significance does not necessarily mean that the gaps are of practical significance—or even noticeable if one observed several students reading together. The tests also employ different scales. The final column in the table expresses the gaps in standard deviation units, a measure that allows for comparing the different scores and estimating their practical meaningfulness.

The second finding is based on the standardized gaps (expressed in SDs). On both NAEP tests, the gaps are narrower among elementary students and wider among middle and high school students. That pattern also appears on international assessments. The gap is twice as large on PISA as on PIRLS.[6] A popular explanation for the gender gap involves the different maturation rates of boys and girls. That theory will be discussed in greater detail below, but at this point in the analysis, let's simply note that the gender gap appears to grow until early adolescence—age 13 on the LTT-NAEP and grade eight on the NAEP Main.

Should these gaps be considered small or large? Many analysts consider 10 scale score points on NAEP equal to about a year of learning. In that light, gaps of five to 10 points appear substantial. But compared to other test score gaps on NAEP, the gender gap is modest in size. On the 2012 LTT-NAEP for nine-year-olds, the five point gap between boys and girls is about one-half of the 10 point gap between students living in cities and those living in suburbs.[7] The gap between students who are eligible for free and reduced lunch and those who are not is 28 points; between black and white students, it is 23 points; and between English language learners (ELL) and non-ELL students, it is 34 points.

Table 1-1 only shows the size of the gender gap as gauged by assessments at single points in time. For determining trends, let's take a closer look at the LTT-

U.S. Gender Gap in Literacy
(Results from Eight Tests)

Table 1-1

Test	Age/Grade	Male	Female	Gap	Standard Deviation	Gap in Standard Deviations
NAEP-LTT (2012)	Age 9	218	223	5*	38	0.13
	Age 13	259	267	8*	37	0.22
	Age 17	283	291	8*	42	0.19
NAEP-Main (2013)	4th Grade	219	225	6*	37	0.16
	8th Grade	263	273	10*	34	0.29
	12th Grade	284	293	9*	38	0.24
International Assessments	PIRLS (2011) 4th Grade	551	562	11*	73	0.15
	PISA (2012) Age 15	482	513	31*	92	0.34

*Significantly different from zero, p<.05.

NAEP, since it provides the longest running record of the gender gap. In Table 1-2, scores are displayed from tests administered since 1971 and given nearest to the starts and ends of decades. Results from 2008 and 2012 are both shown to provide readers an idea of recent fluctuations. At all three ages, gender gaps were larger in 1971 than they are today. The change at age nine is statistically significant, but not at age 13 (p=0.10) or age 17 (p=.07), although they are close. Slight shrinkage occurred in the 1980s, but the gaps expanded again in the 1990s. The gap at age 13 actually peaked at 15 scale score points in 1994 (not shown in the table), and the decline since then is statistically significant. Similarly, the gap at age 17 peaked in 1996 at 15 scale score points, and the decline since then is also statistically significant. More recently, the gap at age nine began to shrink again in 1999, age 13 began shrinking in the 2000s, and age 17 in 2012.

Table 1-3 decomposes the change figures by male and female performance. Sara Mead's point, that the NAEP story is one of both sexes gaining rather than boys falling behind, is even truer today than when she made it in 2006. When Mead's analysis was published, the most recent LTT-NAEP data were from 2004. Up until then, girls had made greater reading gains than boys. But that situation has reversed. Boys have now made larger gains over the history of LTT-NAEP, fueled by the gains that they registered from 2004 to 2012. The score for 17-year-old females in 2012 (291) was identical to their score in 1971.

Trends in the U.S. Gender Gap, NAEP LTT Reading Scores, 1971–2012
(Amount Girls Outscored Boys, in NAEP Scale Score Points)

Table 1-2

	1971	1980	1990	1999	2008	2012
Age 9	13	10	11	6	7	5*
Age 13	11	8	13	12	8	8
Age 17	12	7	12	13	11	8

* 2012 gap significantly different than 1971 gap (p<.05). For age 9, the difference between 1971 and 2012 does not correspond to change in gap reported in Table 1-3 because of rounding.

Change in NAEP LTT Reading Scores, 1971–2012

Table 1-3

	Male	Female	Change in Gap
Age 9	+17	+10	−7
Age 13	+9	+6	−3
Age 17	+4	0	−4

International Perspective

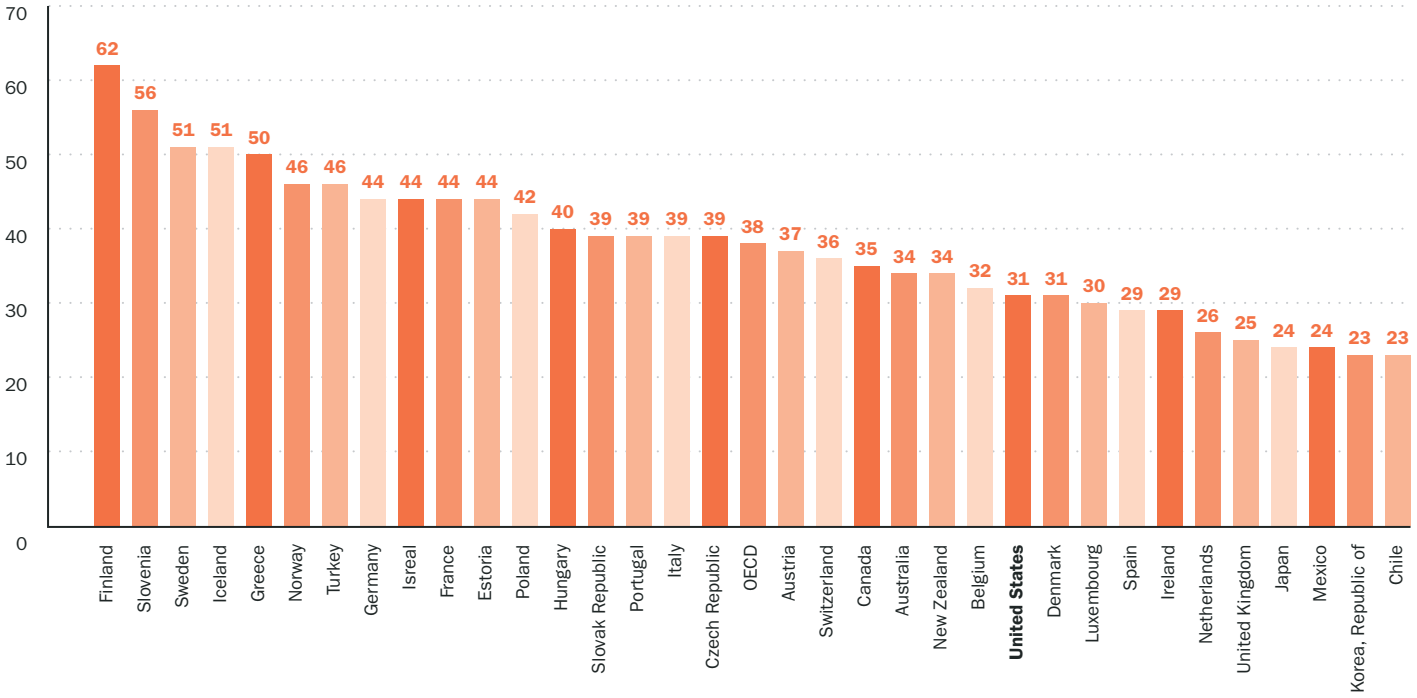
The United States is not alone in reading's gender gap. Its gap of 31 points is not even the largest (see Figure 1-1). On the 2012 PISA, all OECD countries exhibited a gender gap, with females outscoring males by 23 to 62 points on the PISA scale (standard deviation of 94). On average in the OECD, girls outscored boys by 38 points (rounded to 515 for girls and 478 for boys). The U.S. gap of 31 points is less than the OECD average.

Finland had the largest gender gap on the 2012 PISA, twice that of the U.S., with females outscoring males by an astonishing 62 points (0.66 SDs). Finnish girls scored 556, and boys scored 494. To put this gap in perspective, consider that Finland's renowned superiority on PISA tests is completely dependent on Finnish girls. Finland's boys' score of 494 is about the same as the international average of 496, and not much above the OECD average for males (478). The reading performance of Finnish boys is not statistically significantly different from boys in the U.S. (482) or from the average U.S. student, both boys and girls (498).

On the 2012 PISA, all OECD countries exhibited a gender gap.

International Gender Gap in Reading Literacy, 2012 PISA

Fig. 1-1



Finnish superiority in reading only exists among females.

There is a hint of a geographical pattern. Northern European countries tend to have larger gender gaps in reading. Finland, Sweden, Iceland, and Norway have four of the six largest gaps. Denmark is the exception with a 31 point gap, below the OECD average. And two Asian OECD members have small gender gaps. Japan's gap of 24 points and South Korea's gap of 23 are ranked among the bottom four countries. The Nordic tendency toward large gender gaps in reading was noted in a 2002 analysis of the 2000 PISA results.[8] At that time, too, Denmark was the exception. Because of the larger sample and persistence over time, the Nordic pattern warrants more confidence than the one in the two Asian countries.

Back to Finland. That's the headline story here, and it contains a lesson for cautiously interpreting international test scores. Consider that the 62 point gender gap in Finland is only 14 points smaller than the U.S. black-white gap (76 points) and 21 points larger than the Hispanic-white gap (41 points) on the same test. Finland's gender gap illustrates the superficiality of much of the commentary on that country's PISA performance. A common procedure in policy analysis is to consider how policies differentially affect diverse social groups. Think of all the commentators who cite Finland to promote particular policies, whether the policies address teacher recruitment, amount of homework, curriculum standards, the role of play in children's learning, school accountability, or high stakes assessments.[9] Advocates pound the table while arguing that these policies are obviously beneficial. "Just

look at Finland," they say. Have you ever read a warning that even if those policies contribute to Finland's high PISA scores—which the advocates assume but serious policy scholars know to be unproven—the policies also may be having a negative effect on the 50 percent of Finland's school population that happens to be male?

Would Getting Boys to Enjoy Reading More Help Close the Gap?

One of the solutions put forth for improving boys' reading scores is to make an effort to boost their enjoyment of reading. That certainly makes sense, but past scores of national reading and math performance have consistently, and counterintuitively, shown no relationship (or even an inverse one) with enjoyment of the two subjects. PISA asks students how much they enjoy reading, so let's now investigate whether fluctuations in PISA scores are at all correlated with how much 15-year-olds say they like to read.

The analysis below employs what is known as a "differences-in-differences" analytical strategy. In both 2000 and 2009, PISA measured students' reading ability and asked them several questions about how much they like to read. An enjoyment index was created from the latter set of questions.[10] Females score much higher on this index than boys. Many commentators believe that girls' greater enjoyment of reading may be at the root of the gender gap in literacy.

When new international test scores are released, analysts are tempted to just look at variables exhibiting strong correlations with achievement (such as amount of time spent on homework), and embrace them as potential causes of high achievement. But cross-sectional correlations can be deceptive. The direction of causality cannot be determined, whether it's doing a lot of homework that leads to high achievement, or simply that good students tend to take classes that assign more homework. Correlations in cross-sectional data are also vulnerable to unobserved factors that may influence achievement. For example, if cultural predilections drive a country's exemplary performance, their influence will be masked or spuriously assigned to other variables unless they are specifically modeled.[11] Class size, between-school tracking, and time spent on learning are all topics on which differences-in-differences has been fruitfully employed to analyze multiple cross-sections of international data.

Another benefit of differences-in-differences is that it measures statistical relationships longitudinally. Table 1-4 investigates the question: Is the rise and fall of reading enjoyment correlated with changes in reading achievement? Many believe that if boys liked reading more, their literacy test scores would surely increase. Table 1-4 does not support that belief. Data are available for 27 OECD countries, and they are ranked by how much they boosted males' enjoyment of reading. The index is set at the student-level with a mean of 0.00 and standard deviation of 1.00. For the twenty-seven nations in Table 1-4, the mean national change in enjoyment is -.02 with a standard deviation of .09.

Germany did the best job of raising boys' enjoyment of reading, with a gain of 0.12 on the index. German males' PISA scores also went up—a little more than 10 points (10.33). France, on the other hand, raised males' enjoyment of reading nearly as much as Germany (0.11), but French males' PISA scores declined by 15.26 points. A bit further down the column, Ireland managed to get boys to enjoy reading a little more (a gain of 0.05) but their reading performance fell a whopping 36.54 points. Toward the bottom end of the list, Poland's boys enjoyed

Many believe that if boys liked reading more, their literacy test scores would increase.

The correlation coefficient for change in enjoyment and change in reading score is -0.01, indicating no relationship between the two.

Relationship of Change in Males' Reading Enjoyment and Change in Males' Reading Score, 2000–2009
(Ranked by Change in Males' Index of Reading Enjoyment)

Table 1-4

International Average (OECD)	Change in Males' Enjoyment Index	Change in Males' PISA Reading Literacy Score
Germany	0.12	10.33
France	0.11	– 15.26
Japan	0.10	– 6.23
Korea, Republic of	0.08	4.00
Belgium	0.06	0.28
Ireland	0.05	– 36.54
Norway	0.04	– 5.50
New Zealand	0.03	– 8.26
Canada	0.02	– 11.74
Spain	0.01	– 14.39
Chile	0.01	42.06
Italy	0.01	– 5.37
Hungary	0.00	10.88
United States	– 0.01	– 1.89
Austria	– 0.01	– 26.47
Greece	– 0.02	3.11
OECD	– 0.03	– 4.28
Denmark	– 0.04	– 5.06
Switzerland	– 0.04	1.39
Australia	– 0.04	– 16.50
Luxembourg	– 0.08	23.90
Sweden	– 0.11	– 23.58
Portugal	– 0.12	12.18
Finland	– 0.14	– 11.73
Poland	– 0.14	14.29
Iceland	– 0.14	– 10.37
Mexico	– 0.16	1.17
Czech Republic	– 0.20	– 17.13
MEAN	– 0.02	– 3.42
SD	0.09	16.18

reading less in 2009 than in 2000, a decline of 0.14 on the index, but over the same time span, their reading literacy scores increased by more than 14 points (14.29). Among the countries in which the relationship goes in the expected direction is Finland. Finnish males' enjoyment of reading declined (-0.14) as did their PISA scores in reading literacy (-11.73). Overall, the correlation coefficient for change in enjoyment and change in reading score is -0.01, indicating no relationship between the two.

Christina Hoff Sommers and Richard Whitmire have praised specific countries for first recognizing and then addressing the gender gap in reading. Recently, Sommers urged the U.S. to "follow the example of the British, Canadians, and Australians."[12] Whitmire described Australia as "years ahead of the U.S. in pioneering solutions" to the gender gap. Let's see how those countries appear in Table 1-4. England does not have PISA data for the 2000 baseline year, but both Canada and Australia are included. Canada raised boys' enjoyment of reading a little bit (0.02) but Canadian males' scores fell by about 12 points (-11.74). Australia suffered a decline in boys' enjoyment of reading (-0.04) and achievement (-16.50). As promising as these countries' efforts may have appeared a few years ago, so far at least, they have not borne fruit in raising boys' reading performance on PISA.

Achievement gaps are tricky because it is possible for the test scores of the two groups being compared to both decline while the gap increases or, conversely, for scores of both to increase while the gap declines. Table 1-4 only looks at males' enjoyment of reading and its relationship to achievement. A separate differences-in-differences analysis was conducted (but not displayed here) to see whether changes in the enjoyment gap—the difference between boys' and girls' enjoyment of reading—are related to changes in reading achievement. They are not (correlation coefficient of 0.08). National PISA data simply do not support the hypothesis that the superior reading performance of girls is related to the fact that girls enjoy reading more than boys.

Discussion

Let's summarize the main findings of the analysis above. Reading scores for girls exceed those for boys on eight recent assessments of U.S. reading achievement. The gender gap is larger for middle and high school students than for students in elementary school. The gap was apparent on the earliest NAEP tests in the 1970s and has shown some signs of narrowing in the past decade. International tests reveal that the gender gap is worldwide. Among OECD countries, it even appears among countries known for superior performance on PISA's reading test. Finland not only exhibited the largest gender gap in reading on the 2012 PISA, the gap had widened since 2000. A popular recommendation for boosting boys' reading performance is finding ways for them to enjoy reading more. That theory is not supported by PISA data. Countries that succeeded in raising boys' enjoyment of reading from 2000 to 2009 were no more likely to improve boys' reading performance than countries where boys' enjoyment of reading declined.

The origins of the gender gap are hotly debated. The universality of the gap certainly supports the argument that it originates in biological or developmental differences between the two sexes. It is evident among students of different ages in data collected at different points in time. It exists across the globe, in countries with different educational systems, different popular cultures, different child rearing practices, and different conceptions of gender roles. Moreover, the greater prevalence of reading impairment among young boys—a ratio of two or three to one—suggests an endemic difficulty that exists before the influence of schools or culture can take hold.[13]

But some of the data examined above also argue against the developmental explanation. The gap has been shrinking on NAEP. At age nine, it is less than half of what it was forty years ago. Biology doesn't change that fast. Gender gaps in math and science, which were apparent in achievement

The gender gap exists across the globe, in countries with different educational systems, different popular cultures, different child rearing practices, and different conceptions of gender roles.

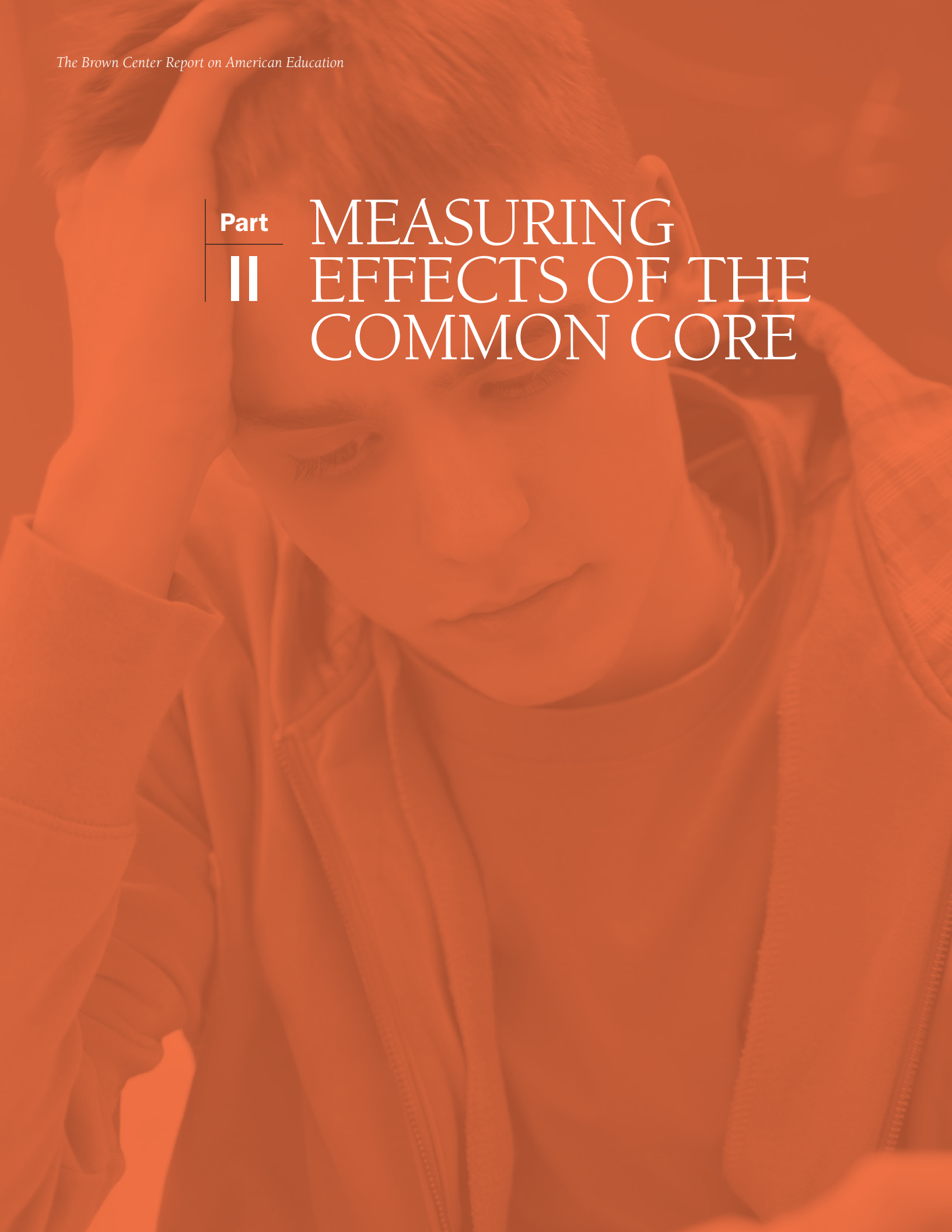
The gap on NAEP at age nine is less than half of what it was forty years ago.

data for a long time, have all but disappeared, especially once course taking is controlled. The reading gap also seems to evaporate by adulthood. On an international assessment of adults conducted in 2012, reading scores for men and women were statistically indistinguishable up to age 35—even in Finland and the United States. After age 35, men had statistically significantly higher scores in reading, all the way to the oldest group, age 55 and older. If the gender gap in literacy is indeed shaped by developmental factors, it may be important for our understanding of the phenomenon to scrutinize periods of the life cycle beyond the age of schooling.

Another astonishing pattern emerged from the study of adult reading. Participants were asked how often they read a book. Of avid book readers (those who said they read a book once a week) in the youngest group (age 24 and younger), 59 percent were women and 41 percent were men. By age 55, avid book readers were even more likely to be women, by a margin of 63 percent to 37 percent. Two-thirds of respondents who said they never read books were men. Women remained the more enthusiastic readers even as the test scores of men caught up with those of women and surpassed them.

A few years ago, Ian McEwan, the celebrated English novelist, decided to reduce the size of the library in his London townhouse. He and his younger son selected thirty novels and took them to a local park. They offered the books to passers-by. Women were eager and grateful to take the books, McEwan reports. Not a single man accepted. The author's conclusion? "When women stop reading, the novel will be dead."[14]

McEwan might be right, regardless of the origins of the gender gap in reading and the efforts to end it.



Part II MEASURING EFFECTS OF THE COMMON CORE

OVER THE NEXT SEVERAL YEARS, POLICY ANALYSTS WILL evaluate the impact of the Common Core State Standards (CCSS) on U.S. education. The task promises to be challenging. The question most analysts will focus on is whether the CCSS is good or bad policy. This section of the Brown Center Report (BCR) tackles a set of seemingly innocuous questions compared to the hot-button question of whether Common Core is wise or foolish. The questions all have to do with when Common Core actually started, or more precisely, when the Common Core started having an effect on student learning. And if it hasn't yet had an effect, how will we know that CCSS has started to influence student achievement?

The analysis below probes this issue empirically, hopefully persuading readers that deciding when a policy begins is elemental to evaluating its effects. The question of a policy's starting point is not always easy to answer. Yet the answer has consequences. You can't figure out whether a policy worked or not unless you know when it began.[15]

The analysis uses surveys of state implementation to model different CCSS starting points for states and produces a second early report card on how CCSS is doing. The first report card, focusing on math, was presented in last year's BCR. The current study updates state implementation ratings that were presented in that report and extends the analysis to achievement in reading. The goal is not only to estimate CCSS's early impact, but also to lay out a fair approach for establishing when the Common Core's impact began—and to do it now before data are generated that either critics or supporters can use to bolster their arguments. The experience of No Child Left Behind (NCLB) illustrates this necessity.

Background

After the 2008 National Assessment of Educational Progress (NAEP) scores were released, former Secretary of Education

Margaret Spellings claimed that the new scores showed "we are on the right track."[16] She pointed out that NAEP gains in the previous decade, 1999–2009, were much larger than in prior decades. Mark Schneider of the American Institutes of Research (and a former Commissioner of the National Center for Education Statistics [NCES]) reached a different conclusion. He compared NAEP gains from 1996–2003 to 2003–2009 and declared NCLB's impact disappointing. "The pre-NCLB gains were greater than the post-NCLB gains."[17] It is important to highlight that Schneider used the 2003 NAEP scores as the starting point for assessing NCLB. A report from FairTest on the tenth anniversary of NCLB used the same demarcation for pre- and post-NCLB time frames.[18] FairTest is an advocacy group critical of high stakes testing—and harshly critical of NCLB—but if the 2003 starting point for NAEP is accepted, its conclusion is indisputable, "NAEP score improvement slowed or stopped in both reading and math after NCLB was implemented."

Choosing 2003 as NCLB's starting date is intuitively appealing. The law was introduced, debated, and passed by Congress in 2001. President Bush signed NCLB into law on January 8, 2002. It takes time to implement any law. The 2003 NAEP is arguably the first chance that the assessment had to register NCLB's effects.

Selecting 2003 is consequential, however. Some of the largest gains in NAEP's history were registered between 2000 and 2003. Once 2003 is established as a starting point (or baseline), pre-2003 gains become "pre-NCLB." But what if the 2003 NAEP scores were influenced by NCLB? Experiments evaluating the effects of new drugs collect baseline data from subjects before treatment, not after the treatment has begun. Similarly, evaluating the effects of public policies require that baseline data are not influenced by the policies under evaluation.

Avoiding such problems is particularly difficult when state or local policies are adopted nationally. The federal effort to establish a speed limit of 55 miles per hour in the 1970s is a good example. Several states already had speed limits of 55 mph or lower prior to the federal law's enactment. Moreover, a few states lowered speed limits in anticipation of the federal limit while the bill was debated in Congress. On the day President Nixon signed the bill into law—January 2, 1974—the Associated Press reported that only 29 states would be required to lower speed limits. Evaluating the effects of the 1974 law with national data but neglecting to adjust for what states were already doing would obviously yield tainted baseline data.

There are comparable reasons for questioning 2003 as a good baseline for evaluating NCLB's effects. The key components of NCLB's accountability provisions—testing students, publicizing the results, and holding schools accountable for results—were already in place in nearly half the states. In some states they had been in place for several years. The 1999 iteration of *Quality Counts, Education Week's* annual report on state-level efforts to improve public education, entitled *Rewarding Results, Punishing Failure*, was devoted to state accountability systems and the assessments underpinning them. Testing and accountability are especially important because they have drawn fire from critics of NCLB, a law that wasn't passed until years later.

The Congressional debate of NCLB legislation took all of 2001, allowing states to pass anticipatory policies. Derek Neal and Diane Whitmore Schanzenbach reported that "with the passage of NCLB lurking on the horizon," Illinois placed hundreds of schools

Baseline data should not be influenced by the policies under evaluation.

Using 2003 as a baseline assumes that none of these activities—previous accountability systems, public lists of schools in need of improvement, anticipatory policy shifts—influenced achievement.

on a watch list and declared that future state testing would be high stakes.[19] In the summer and fall of 2002, with NCLB now the law of the land, state after state released lists of schools falling short of NCLB's requirements. Then the 2002–2003 school year began, during which the 2003 NAEP was administered. Using 2003 as a NAEP baseline assumes that none of these activities—previous accountability systems, public lists of schools in need of improvement, anticipatory policy shifts—influenced achievement. That is unlikely.[20]

The Analysis

Unlike NCLB, there was no "pre-CCSS" state version of Common Core. States vary in how quickly and aggressively they have implemented CCSS. For the BCR analyses, two indexes were constructed to model CCSS implementation. They are based on surveys of state education agencies and named for the two years that the surveys were conducted. The 2011 survey reported the number of programs (e.g., professional development, new materials) on which states reported spending federal funds to implement CCSS. Strong implementers spent money on more activities. The 2011 index was used to investigate eighth grade math achievement in the 2014 BCR. A new implementation index was created for this year's study of reading achievement. The 2013 index is based on a survey asking states when they planned to complete full implementation of CCSS in classrooms. Strong states aimed for full implementation by 2012–2013 or earlier.

Fourth grade NAEP reading scores serve as the achievement measure. Why fourth grade and not eighth? Reading instruction is a key activity of elementary classrooms but by eighth grade has all but disappeared. What remains of "reading" as an independent subject, which has typically morphed into the study of literature, is subsumed under the English-Language Arts curriculum, a catchall term that also includes writing, vocabulary, listening, and public speaking. Most students in fourth grade are in self-contained classes; they receive instruction in all subjects from one teacher. The impact of CCSS on reading instruction—the recommendation that non-fiction take a larger role in reading materials is a good example—will be concentrated in the activities of a single teacher in elementary schools. The burden for meeting CCSS's press for non-fiction, on the other hand, is expected to be shared by all middle and high school teachers.[21]

Changes in NAEP 4th Grade Reading, By 2011 Implementation Index

Table 2-1

Implementation Rating	2009–2011	2011–2013	2009–2013
Strong (n=19)	0.22	0.64	0.87
Medium (n=27)	0.17	0.81	0.99
Non-adopters (n=4)	– 0.78	0.53	– 0.24
All (n=50)	0.12	0.73	0.84

Note: Strong = adopted CCSS in ELA and pursued three implementation strategies (PD, new instructional materials, joined testing consortium). Medium = adopted CCSS-ELA standards but did not employ at least one of the implementation strategies. Non-adopters = did not adopt CCSS-ELA.

Source: Modified from: Webber, A., Troppe, P., Milanowski, A., Gutmann, G., Reisner, E. and Goertz, M. State (2014), Table H.1. Standards and Assessment Indicators by State, 2010–2011, in *State Implementation of Reforms Promoted Under the Recovery Act*, National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC.

Results

Table 2-1 displays NAEP gains using the 2011 implementation index. The four year period between 2009 and 2013 is broken down into two parts: 2009–2011 and 2011–2013. Nineteen states are categorized as "strong" implementers of CCSS on the 2011 index, and from 2009–2013, they outscored the four states that did not adopt CCSS by a little more than one scale score point (0.87 vs. -0.24 for a 1.11 difference).

The non-adopters are the logical control group for CCSS, but with only four states in that category—Alaska, Nebraska, Texas, and Virginia—it is sensitive to big changes in one or two states. Alaska and Texas both experienced a decline in fourth grade reading scores from 2009–2013.

The 1.11 point advantage in reading gains for strong CCSS implementers is similar to the 1.27 point advantage reported last year for eighth grade math. Both are small. The reading difference in favor of CCSS is equal to approximately 0.03 standard deviations of the 2009 baseline reading score. Also note that the differences were greater in 2009–2011 than in 2011–2013 and that the "medium" implementers performed as well as or better than the strong implementers over the entire four year period (gain of 0.99).

Table 2-2 displays calculations using the 2013 implementation index. Twelve states are rated as strong CCSS implementers, seven fewer than on the 2011 index.[22] Data for the non-adopters are the same as in the previous table. In 2009–2013, the strong implementers gained 1.27 NAEP points compared to -0.24 among the non-adopters, a difference of 1.51 points. The thirty-four states rated as medium implementers gained 0.82. The strong implementers on this index are states that reported full implementation of CCSS-ELA by 2013. Their larger gain in 2011–2013 (1.08 points) distinguishes them from the strong implementers in the previous table. The overall advantage of 1.51 points over non-adopters represents about 0.04 standard deviations of the 2009 NAEP reading score, not a difference with real world significance. Taken together, the 2011 and 2013 indexes estimate that NAEP reading gains from 2009–2013 were one to one and one-half scale score points larger in the strong CCSS implementation states compared to the states that did not adopt CCSS.

Changes in NAEP 4th Grade Reading, By 2013 Implementation Index

Table 2-2

Implementation Rating	2009–2011	2011–2013	2009–2013
Strong (n=12)	0.19	1.08	1.27
Medium (n=34)	0.20	0.62	0.82
Non-adopters (n=4)	– 0.78	0.53	– 0.24
All (n=50)	0.12	0.73	0.84

Note: Strong = 2012–13 academic year or earlier, state's timeline for classroom implementation of ELA-CSSS, Medium = after 2012–2013 academic year, state's timeline for implementation of ELA-CSSS, Non-adopters = states not adopting ELA-CCSS.

Data Source: Achieve (2013), "CCSS/CCR Standards Implementation Timeline," in *Closing the Expectations Gap 2013 Annual Report*, p. 39. Data based on 2013 survey of state educational agencies.

Common Core and Reading Content

As noted above, the 2013 implementation index is based on when states scheduled full implementation of CCSS in classrooms. Other than reading achievement, does the index seem to reflect changes in any other classroom variable believed to be related to CCSS implementation? If the answer is "yes," that would bolster confidence that the index is measuring changes related to CCSS implementation.

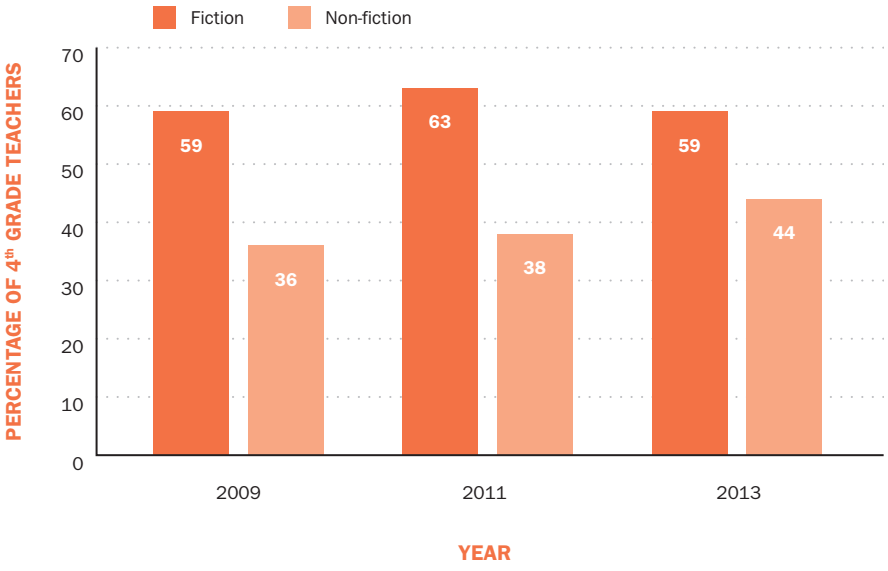
Let's examine the types of literature that students encounter during instruction. Perhaps the most controversial recommendation in the CCSS-ELA standards is the call for teachers to shift the content of reading materials away from stories and other fictional forms of literature in favor of more non-fiction. NAEP asks fourth grade teachers the extent to which they teach fiction and non-fiction over the course of the school year (see Figure 2-1).

Historically, fiction dominates fourth grade reading instruction. It still does. The percentage of teachers reporting that they teach fiction to a "large extent" exceeded the percentage answering "large extent" for non-

The 1.11 point advantage in reading gains for strong CCSS implementers is similar to the 1.27 point advantage for eighth grade math. Both are small.

4th Grade Teachers (percent) Teaching Fiction and Non-Fiction a "Great Extent," NAEP 2009–2013

Figure 2-1



Fiction's Prominence and the Implementation of CCSS, 2009–2013

Table 2-3

2013 Implementation Rating	Fiction's Prominence			Change		
	2009	2011	2013	2009–2011	2011–2013	2009–2013
Strong (n=12)	23.0	24.6	12.2	1.6	– 12.4	– 10.8
Medium (n=34)	22.0	22.6	14.5	0.6	– 8.1	– 7.5
Non-Adopters (n=4)	27.0	25.3	17.3	– 1.8	– 8.0	– 9.8

Note: Fiction's prominence equals the difference reported in Figure 2-1, the percentage of teachers saying they teach fiction "a great extent" minus the percentage saying they teach non-fiction "a great extent."

fiction by 23 points in 2009 and 25 points in 2011. In 2013, the difference narrowed to only 15 percentage points, primarily because of non-fiction's increased use. Fiction still dominated in 2013, but not by as much as in 2009.

The differences reported in Figure 2-1 are national indicators of fiction's declining prominence in fourth grade reading instruction. What about the states? We know that they were involved to varying degrees with the implementation of Common Core from 2009–2013. Is there evidence that fiction's prominence was more likely to weaken in states most aggressively pursuing CCSS implementation?

Table 2-3 displays the data tackling that question. Fourth grade teachers in strong implementation states decisively favored the use of fiction over non-fiction in 2009 and 2011. But the prominence of fiction in those states experienced a large decline in 2013 (-12.4 percentage points). The decline for the entire four year period, 2009–2013, was larger in the strong implementation states (-10.8) than in the medium implementation (-7.5) or non-adoption states (-9.8).

Conclusion

This section of the Brown Center Report analyzed NAEP data and two indexes of CCSS implementation, one based on data collected in 2011, the second from data collected in 2013. NAEP scores for 2009–2013 were examined. Fourth grade reading scores improved by 1.11 scale score points in states with strong implementation of CCSS compared to states that did not adopt CCSS. A similar comparison in last year's BCR found a 1.27 point difference on NAEP's eighth grade math test, also in favor of states with strong implementation of CCSS. These differences, although certainly encouraging to CCSS supporters, are quite small, amounting to (at most) 0.04 standard deviations (SD) on the NAEP scale. A threshold of 0.20 SD—five times larger—is often invoked as the minimum size for a test score change to be regarded as noticeable. The current study's findings are also merely statistical associa-

tions and cannot be used to make causal claims. Perhaps other factors are driving test score changes, unmeasured by NAEP or the other sources of data analyzed here.

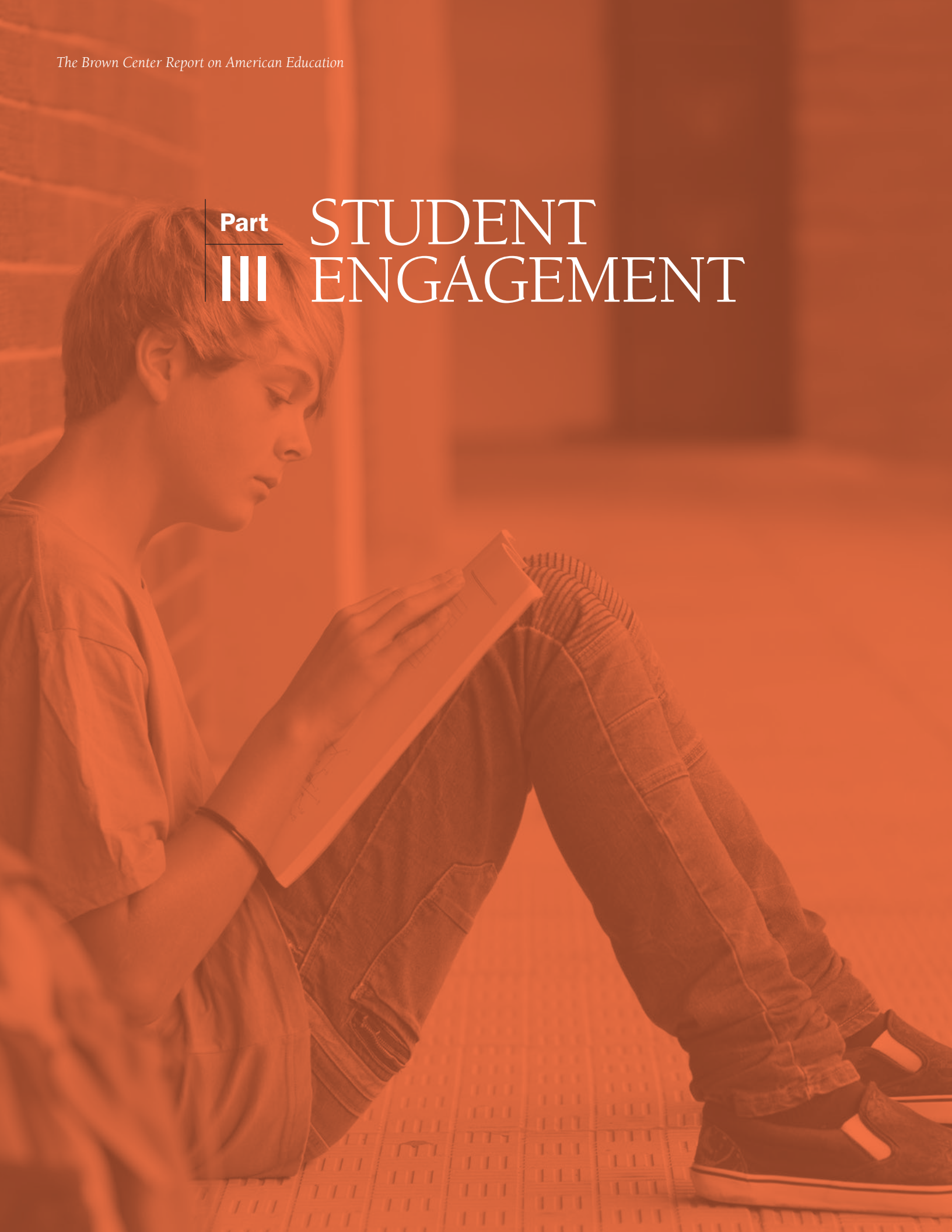
The analysis also found that fourth grade teachers in strong implementation states are more likely to be shifting reading instruction from fiction to non-fiction texts. That trend should be monitored closely to see if it continues. Other events to keep an eye on as the Common Core unfolds include the following:

1. The 2015 NAEP scores, typically released in the late fall, will be important for the Common Core. In most states, the first CCSS-aligned state tests will be given in the spring of 2015. Based on the earlier experiences of Kentucky and New York, results are expected to be disappointing. Common Core supporters can respond by explaining that assessments given for the first time often produce disappointing results. They will also claim that the tests are more rigorous than previous state assessments. But it will be difficult to explain stagnant or falling NAEP scores in an era when implementing CCSS commands so much attention.
2. Assessment will become an important implementation variable in 2015 and subsequent years. For analysts, the strategy employed here, modeling different indicators based on information collected at different stages of implementation, should become even more useful. Some states are planning to use Smarter Balanced Assessments tests, others are using The Parnership for Assessment of Readiness for College and Careers (PARCC), and still others are using their own home-grown tests. To capture variation among the states on this important dimension of implementation, analysts will need to use indicators that are up-to-date.
3. The politics of Common Core injects a dynamic element into implementation. The status of implementation is constantly changing. States may choose to suspend, to delay, or to abandon CCSS. That will require analysts to regularly re-configure which states are considered "in" Common Core and which states are "out." To further complicate matters, states may be "in" some years and "out" in others.

A final word. When the 2014 BCR was released, many CCSS supporters commented that it is too early to tell the effects of Common Core. The point that states may need more time operating under CCSS to realize its full effects certainly has merit. But that does not discount everything states have done so far—including professional development, purchasing new textbooks and other instructional materials, designing new assessments, buying and installing computer systems, and conducting hearings and public outreach—as part of implementing the standards. Some states are in their fifth year of implementation. It could be that states need more time, but innovations can also produce their biggest "pop" earlier in implementation rather than later. Kentucky was one of the earliest states to adopt and implement CCSS. That state's NAEP fourth grade reading score declined in both 2009–2011 and 2011–2013. The optimism of CCSS supporters is understandable, but a one and a half point NAEP gain might be as good as it gets for CCSS.

A one and a half point NAEP gain might be as good as it gets for CCSS.

Part III STUDENT ENGAGEMENT



STUDENT ENGAGEMENT REFERS TO THE INTENSITY WITH WHICH students apply themselves to learning in school. Traits such as motivation, enjoyment, and curiosity—characteristics that have interested researchers for a long time—have been joined recently by new terms such as, "grit," which now approaches cliché status. International assessments collect data from students on characteristics related to engagement. This study looks at data from the Program for International Student Assessment (PISA), an international test given to fifteen-year-olds. In the U.S., most PISA students are in the fall of their sophomore year. The high school years are a time when many observers worry that students lose interest in school.

Compared to their peers around the world, how do U.S. students appear on measures of engagement? Are national indicators of engagement related to achievement? This analysis concludes that American students are about average in terms of engagement. Data reveal that several countries noted for their superior ranking on PISA—e.g., Korea, Japan, Finland, Poland, and the Netherlands—score below the U.S. on measures of student engagement. Thus, the relationship of achievement to student engagement is not clear cut, with some evidence pointing toward a weak positive relationship and other evidence indicating a modest negative relationship.

The Unit of Analysis Matters

Education studies differ in units of analysis. Some studies report data on individuals, with each student serving as an observation. Studies of new reading or math programs, for example, usually report an average gain score or effect size representing the impact of the program on the average student. Others studies report aggregated data, in which test scores or other measurements are averaged to yield a group score. Test scores of schools, districts, states, or countries are constructed like that. These scores represent the performance of groups, with each group serving as a single observation, but they are really just data from individuals that have been aggregated to the group level.

Aggregated units are particularly useful for policy analysts. Analysts are interested in how Fairfax County or the state of Virginia or the United States is doing. Governmental bodies govern those jurisdictions and policymakers craft policy for all of the citizens within the political jurisdiction—not for an individual.

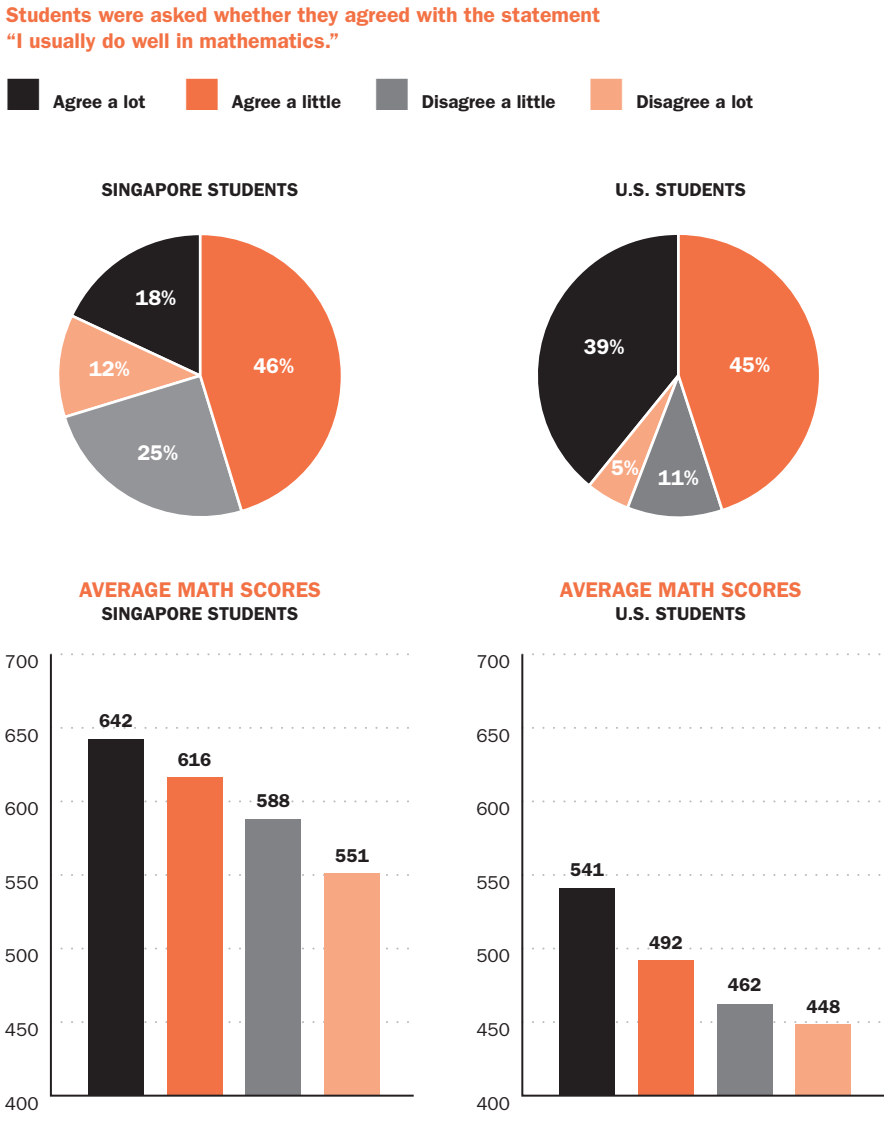
The analytical unit is especially important when investigating topics like student engagement and their relationships with achievement. Those relationships are inherently individual, focusing on the interaction of psychological characteristics. They are also prone to reverse causality, meaning that the direction of cause and effect cannot readily be determined. Consider self-esteem and academic achievement. Determining which one is cause and which is effect has been debated for decades. Students who are good readers enjoy books, feel pretty good about their reading abilities, and spend more time reading than other kids. The possibility of reverse causality is one reason that beginning statistics students learn an important rule: correlation is *not* causation.

Starting with the first international assessments in the 1960s, a curious pattern has emerged. Data on students' attitudes toward studying school subjects, when examined on a national level, often exhibit the opposite relationship with achievement than one would expect. The 2006 Brown Center Report (BCR) investigated the phenomenon in a study of "the happiness factor" in learning.[23] Test scores of fourth graders in 25 countries and eighth graders in 46 countries were analyzed. Students in countries with low math scores were more likely to report that they enjoyed math than students in high-scoring countries. Correlation coefficients for the association of enjoyment and achievement were -0.67 at fourth grade and -0.75 at eighth grade.

Confidence in math performance was also inversely related to achievement. Correlation coefficients for national achievement and the percentage of students responding affirmatively to the statement, "I

Relationship of Math Achievement with "I usually do well in mathematics," TIMSS 2003

Fig. 3-1



Note: Data refer only to eigth grade.

Sources: Figure recreated from Loveless, T. (2006) "The Happiness Factor in Student Learning," *The 2006 Brown Center Report on American Education: How Well are America Students Learning?*, p. 18–19, Washington, DC: The Brookings Institution. Data from: Martin, M. O. (Ed.) (2005) *TIMSS 2003 User Guide for the International Database*, p. 67, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

The unit of analysis must be considered when examining data on students' characteristics and their relationship to achievement.

usually do well in mathematics," were -0.58 among fourth graders and -0.64 among eighth graders. Nations with the most confident math students tend to perform poorly on math tests; nations with the least confident students do quite well.

That is odd. What's going on? A comparison of Singapore and the U.S. helps unravel the puzzle. The data in Figure 3-1 are for eighth graders on the 2003 Trends in Mathematics and Science Study (TIMSS). U.S. students were very confident—84% either agreed a lot or a little (39% + 45%) with the statement that they usually do well in mathematics. In Singapore, the figure was 64% (46% + 18%). With a score of 605, however, Singaporean students registered about one full standard deviation (80 points) higher on the TIMSS math test compared to the U.S. score of 504.

When within-country data are examined, the relationship exists in the expected direction. In Singapore, highly confident students score 642, approximately 100 points above the least-confident students (551). In the U.S., the gap between the most- and least-confident students was also about 100 points—but at a much lower level on the TIMSS scale, at 541 and 448. Note that the least-confident Singaporean eighth grader still outscores the most-confident American, 551 to 541.

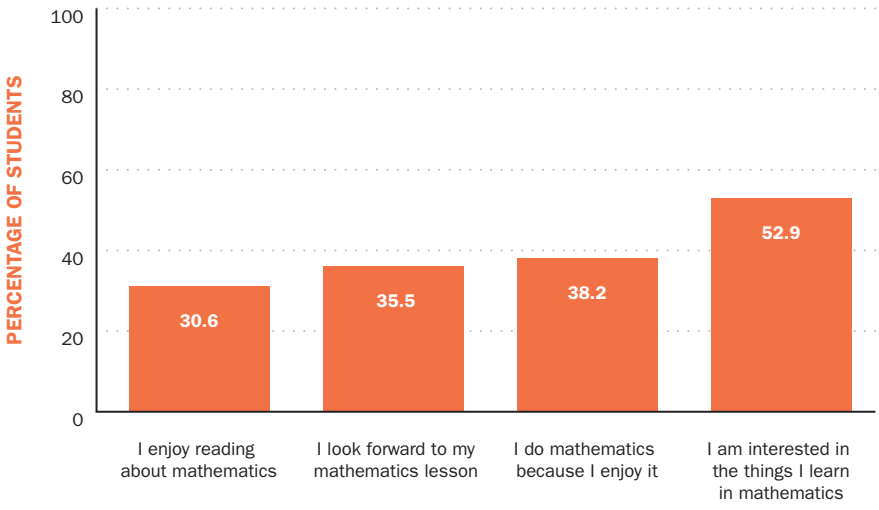
The lesson is that the unit of analysis must be considered when examining data on students' psychological characteristics and their relationship to achievement. If presented with country-level associations, one should wonder what the within-country associations are. And vice versa. Let's keep that caution in mind as we now turn to data on fifteen-year-olds' intrinsic motivation and how nations scored on the 2012 PISA.

Intrinsic Motivation

PISA's index of intrinsic motivation to learn mathematics comprises responses to four items on the student questionnaire: 1) I enjoy reading about mathematics; 2) I look forward to my mathematics lessons; 3) I do mathematics because I enjoy it; and 4) I am interested in the things I learn in mathematics. Figure 3-2 shows the percentage of students in OECD countries—thirty of the most economically developed nations in the world—responding that they agree or strongly agree with the statements. A little less than one-third (30.6%) of students responded favorably to reading about math, 35.5% responded favorably to looking forward to math lessons, 38.2% reported doing math because they enjoy it, and 52.9% said they were interested in the things they learn in math. A ballpark estimate, then, is that one-third to one-half of students respond affirmatively to the individual components of PISA's intrinsic motivation index.

Students' Intrinsic Motivation to Learn Mathematics
Percentage of students across OECD countries who reported that they "agree" or "strongly agree" with the following statements:

Fig. 3-2



Source: OECD (2013), Graph III.3.9. Students' intrinsic motivation to learn mathematics: Percentage of students across OECD countries who reported that they "agree" or "strongly agree" with the following statements:, in *PISA 2012 Results: Ready to Learn (Volume III)*, OECD Publishing, Paris.DOI: http://dx.doi.org/10.1787/9789264201170-graph27-en

Table 3-1 presents national scores on the 2012 index of intrinsic motivation to learn mathematics. The index is scaled with an average of 0.00 and a standard deviation of 1.00. Student index scores are averaged to produce a national score. The scores of 39 nations are reported—29 OECD countries and 10 partner countries.[24] Indonesia appears to have the most intrinsically motivated students in the world (0.80), followed by Thailand (0.77), Mexico (0.67), and Tunisia (0.59). It is striking that developing countries top the list. Universal education at the elementary level is only a recent reality in these countries, and they are still struggling to deliver universally accessible high schools, especially in rural areas and especially to girls. The students who sat for PISA may be an unusually motivated group. They also may be deeply appreciative of having an opportunity that their parents never had.

The U.S. scores about average (0.08) on the index, statistically about the same as New Zealand, Australia, Ireland, and Canada. The bottom of the table is extremely interesting. Among the countries with the least intrinsically motivated kids are some PISA high flyers. Austria has the least motivated students (-0.35), but that is not statistically significantly different from the score for the Netherlands (-0.33). What's surprising is that Korea (-0.20), Finland (-0.22), Japan (-0.23), and Belgium (-0.24) score at the bottom of the intrinsic motivation index even though they historically do quite well on the PISA math test.

Enjoying Math and Looking Forward to Math Lessons

Let's now dig a little deeper into the intrinsic motivation index. Two components of the index are how students respond to "I do mathematics because I enjoy it" and "I look forward to my mathematics lessons." These

Index of Intrinsic Motivation to Learn Mathematics, 2012
(Ranked by country index score)

Table 3-1

Country	Index	Standard Error
Indonesia	0.80	(0.02)
Thailand	0.77	(0.02)
Mexico	0.67	(0.01)
Tunisia	0.59	(0.02)
Turkey	0.44	(0.02)
Brazil	0.42	(0.01)
Denmark	0.35	(0.02)
Hong Kong-China	0.30	(0.02)
Russian Federation	0.29	(0.02)
Uruguay	0.27	(0.02)
Greece	0.21	(0.02)
Iceland	0.15	(0.02)
Macao-China	0.15	(0.01)
Sweden	0.12	(0.02)
Portugal	0.12	(0.02)
New Zealand	0.11	(0.02)
Australia	0.11	(0.01)
Liechtenstein	0.09	(0.08)
AVERAGE	**0.08**	**0.02**
United States	0.08	(0.03)
Ireland	0.06	(0.02)
Canada	0.05	(0.01)
Italy	0.01	(0.02)
France	– 0.02	(0.02)
Switzerland	– 0.02	(0.02)
Latvia	– 0.05	(0.02)
Germany	– 0.11	(0.02)
Spain	– 0.14	(0.01)
Norway	– 0.15	(0.02)
Poland	– 0.16	(0.02)
Luxembourg	– 0.16	(0.02)
Czech Republic	– 0.16	(0.02)
Hungary	– 0.18	(0.02)
Slovak Republic	– 0.19	(0.02)
Korea	– 0.20	(0.03)
Finland	– 0.22	(0.02)
Japan	– 0.23	(0.02)
Belgium	– 0.24	(0.02)
Netherlands	– 0.33	(0.02)
Austria	– 0.35	(0.02)

Relationship of Math Achievement with "I do mathematics because I enjoy it" and "I look forward to my mathematics lessons"
(Countries ranked by national percentage of students who "agree" or "strongly agree" with statements).

Table 3-2

	Enjoy			Looking Forward		
	Country	Percent	PISA Score	Country	Percent	PISA Score
	Indonesia	78.3	375	Indonesia	72.3	375
	Thailand	70.6	427	Mexico	70.6	413
	Tunisia	58.0	388	Thailand	68.8	427
	Denmark	56.9	500	Tunisia	54.4	388
	Liechtenstein	56.2	535	Denmark	51.5	500
	Brazil	55.8	391	Hong Kong-China	49.8	561
	Hong Kong-China	54.9	561	Turkey	48.9	448
	Mexico	52.8	413	New Zealand	46.1	500
	Turkey	52.7	448	Russian Federation	45.9	482
	Greece	51.7	453	United States	45.4	481
	Uruguay	50.6	409	Australia	45.3	504
	Switzerland	48.5	531	Brazil	43.9	391
	Iceland	47.7	493	Liechtenstein	42.3	535
	Italy	45.8	485	Macao-China	41.7	538
	Portugal	45.5	487	Uruguay	40.7	409
	Russian Federation	42.9	482	Ireland	40.2	501
	Macao-China	42.3	538	Canada	39.7	518
	France	41.5	495	Iceland	39.7	493
	Germany	39.0	514	Switzerland	38.9	531
MEDIAN POINT	Australia	39.0	504	Germany	36.9	514
	Latvia	38.6	491	Greece	36.8	453
	New Zealand	38.2	500	Sweden	36.2	478
	Spain	37.0	484	Luxembourg	35.7	490
	Ireland	37.0	501	Czech Republic	33.9	499
	Sweden	37.0	478	Japan	33.7	536
	United States	36.6	481	Norway	33.2	489
	Canada	36.6	518	Austria	32.6	506
	Poland	36.1	518	Portugal	32.6	487
	Luxembourg	35.3	490	Slovak Republic	30.8	482
	Netherlands	32.4	523	Hungary	30.3	477
	Norway	32.2	489	Italy	29.0	485
	Japan	30.8	536	Spain	25.7	484
	Korea	30.7	554	Finland	24.8	519
	Czech Republic	30.3	499	Belgium	24.2	515
	Belgium	28.8	515	France	23.8	495
	Finland	28.8	519	Korea	21.8	554
	Slovak Republic	27.9	482	Poland	21.3	518
	Hungary	27.5	477	Latvia	20.8	491
	Austria	23.8	506	Netherlands	19.8	523
	Mean	**42.5**	**487**	**Mean**	**38.7**	**487**
	Correlation Coefficient	**– 0.58**		**Correlation Coefficient**	**– 0.57**	

sentiments are directly related to schooling. Whether students enjoy math or look forward to math lessons is surely influenced by factors such as teachers and curriculum. Table 3-2 rank orders PISA countries by the percentage of students who "agree" or "strongly agree" with the questionnaire prompts. The nations' 2012 PISA math scores are also tabled. Indonesia scores at the top of both rankings, with 78.3% enjoying math and 72.3% looking forward to studying the subject. However, Indonesia's PISA math score of 375 is more than one full standard deviation below the international mean of 494 (standard deviation of 92). The tops of the tables are primarily dominated by low-performing countries, but not exclusively so. Denmark is an average-performing nation that has high rankings on both sentiments. Liechtenstein, Hong Kong-China, and Switzerland do well on the PISA math test and appear to have contented, positively-oriented students.

Several nations of interest are shaded. The bar across the middle of the tables, encompassing Australia and Germany, demarcates the median of the two lists, with 19 countries above and 19 below that position. The United States registers above the median on looking forward to math lessons (45.4%) and a bit below the median on enjoyment (36.6%). A similar proportion of students in Poland—a country recently celebrated in popular media and in Amanda Ripley's book, *The Smartest Kids in the World*,[25] for making great strides on PISA tests—enjoy math (36.1%), but only 21.3% of Polish kids look forward to their math lessons, very near the bottom of the list, anchored by Netherlands at 19.8%.

Korea also appears in Ripley's book. It scores poorly on both items. Only 30.7% of Korean students enjoy math, and less than that, 21.8%, look forward to studying the

subject. Korean education is depicted unflatteringly in Ripley's book—as an academic pressure cooker lacking joy or purpose—so its standing here is not surprising. But Finland is another matter. It is portrayed as laid-back and student-centered, concerned with making students feel relaxed and engaged. Yet, only 28.8% of Finnish students say that they study mathematics because they enjoy it (among the bottom four countries) and only 24.8% report that they look forward to math lessons (among the bottom seven countries). Korea, the pressure cooker, and Finland, the laid-back paradise, look about the same on these dimensions.

Another country that is admired for its educational system, Japan, does not fare well on these measures. Only 30.8% of students in Japan enjoy mathematics, despite the boisterous, enthusiastic classrooms that appear in Elizabeth Green's recent book, *Building a Better Teacher*.[26] Japan does better on the percentage of students looking forward to their math lessons (33.7%), but still places far below the U.S. Green's book describes classrooms with younger students, but even so, surveys of Japanese fourth and eighth graders' attitudes toward studying mathematics report results similar to those presented here. American students say that they enjoy their math classes and studying math more than students in Finland, Japan, and Korea.

It is clear from Table 3-2 that at the national level, enjoying math is not positively related to math achievement. Nor is looking forward to one's math lessons. The correlation coefficients reported in the last row of the table quantify the magnitude of the inverse relationships. The -0.58 and -0.57 coefficients indicate a moderately negative association, meaning, in plain English, that countries with students who enjoy math or look forward to math lessons tend to score below average on the PISA math test. And high-scoring nations tend to register below average on these measures of student engagement. Country-level associations, however, should be augmented with student-level associations that are calculated within each country.

Within-Country Associations of Student Engagement with Math Performance

The 2012 PISA volume on student engagement does not present within-country correlation coefficients on intrinsic motivation or its components. But it does offer within-country correlations of math achievement with three other characteristics relevant to student engagement. Table 3-3 displays

What is a Correlation Coefficient?

A Pearson correlation coefficient measures the strength of a linear relationship between two variables. The coefficient is always between -1.00 and +1.00. The closer a coefficient is to +/- 1.00 the stronger a relationship is between two variables. 1.00 signifies a perfect positive relationship while -1.00 signifies a perfect negative relationship.

Within-Country Associations of Student Engagement with Math Performance, PISA 2012
(Correlation coefficients, within-country calculations at student-level)

Table 3-3

	Sense of Belonging	Attitudes Toward School	Arriving Late for School
Range	– 0.02 to 0.18*	– 0.05* to .24*	– 0.23* to – 0.03
OECD Avg.	0.08*	0.11*	– 0.14*
U.S.	0.07*	0.14*	– 0.20*

* Significantly different from zero, p<.05.
Note: N = 39 nations (29 OECD and 10 partner nations).

Source: OECD (2013), Table III.2.9. Change between 2003 and 2012 in the association between students' engagement with school and mathematics performance, in *PISA 2012 Results: Ready to Learn (Volume III)*, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264201170-table127-en

Neither enjoying math nor looking forward to one's math lesson is positively related to math achievement.

Within-country correlations of math achievement and student engagement trend in the direction expected but they are small in magnitude.

statistics for students' responses to: 1) if they feel like they belong at school; 2) their attitudes toward school, an index composed of four factors;[27] and 3) whether they had arrived late for school in the two weeks prior to the PISA test. These measures reflect an excellent mix of behaviors and dispositions.

The within-country correlations trend in the direction expected but they are small in magnitude. Correlation coefficients for math performance and a sense of belonging at school range from -0.02 to 0.18, meaning that the country exhibiting the strongest relationship between achievement and a sense of belonging—Thailand, with a 0.18 correlation coefficient—isn't registering a strong relationship at all. The OECD average is 0.08, which is trivial. The U.S. correlation coefficient, 0.07, is also trivial. The relationship of achievement with attitudes toward school is slightly stronger (OECD average of 0.11), but is still weak.

Of the three characteristics, arriving late for school shows the strongest correlation, an unsurprising inverse relationship of -0.14 in OECD countries and -0.20 in the U.S. Students who tend to be tardy also tend to score lower on math tests. But, again, the magnitude is surprisingly small. The coefficients are statistically significant because of large sample sizes, but in a real world "would I notice this if it were in my face?" sense, no, the correlation coefficients are suggesting not much of a relationship at all.

The PISA report presents within-country effect sizes for the intrinsic motivation index, calculating the achievement gains associated with a one unit change in the index. One of several interesting findings is that intrinsic motivation is more strongly associated with gains at the top of the achievement distribution, among students at the 90th percentile in math scores, than at the bottom of the distribution, among students at the 10th percentile.

The report summarizes the within-country effect sizes with this statement: "On average across OECD countries, a change of one unit in the index of intrinsic motivation to learn mathematics translates into a 19 score-point difference in mathematics performance."[28] This sentence can be easily misinterpreted. It means that within each of the participating countries students who differ by one unit on PISA's 2012 intrinsic motivation index score about 19 points apart on the 2012 math test. It does not mean that a country that gains one unit on the intrinsic motivation index can expect a 19 point score increase.[29]

Let's now see what that association looks like at the national level.

National Changes in Intrinsic Motivation, 2003–2012

PISA first reported national scores on the index of intrinsic motivation to learn mathematics in 2003. Are gains that countries made on the index associated with gains on PISA's math test? Table 3-4 presents a score card on the question, reporting the changes that occurred in thirty-nine nations—in both the index and math scores—from 2003 to 2012. Seventeen nations made statistically significant gains on the index; fourteen nations had gains that were, in a statistical sense, indistinguishable from zero—labeled "no change" in the table; and eight nations experienced statistically significant declines in index scores.

The U.S. scored 0.00 in 2003 and 0.08 in 2012, notching a gain of 0.08 on the index (statistically significant). Its PISA math score declined from 483 to 481, a decline of 2 scale score points (not statistically significant).

Table 3-4 makes it clear that national changes on PISA's intrinsic motivation index are not associated with changes in math achievement. The countries registering gains on the index averaged a decline of

Change in National Index of Intrinsic Motivation to Learn Mathematics and PISA Math Score, 2003–2012

Table 3-4

National Change*	Change in Index Score	Change in PISA Math Score
Gain (n = 17)	+0.12	– 3.70
No Change (n = 14)	– 0.01	– 0.09
Decline (n = 8)	– 0.15	+10.30

* Gaining and declining nations registered statistically significant changes ($p < 0.05$) on index of intrinsic motivation to learn mathematics.

3.7 points on PISA's math assessment. The countries that remained about the same on the index had math scores that also remain essentially unchanged (-0.09). And the most striking finding: countries that declined on the index (average of -0.15) actually gained an average of 10.3 points on the PISA math scale. Intrinsic motivation went down; math scores went up. The correlation coefficient for the relationship over all, not shown in the table, is -0.30.

Conclusion

The analysis above investigated student engagement. International data from the 2012 PISA were examined on several dimensions of student engagement, focusing on a measure that PISA has employed since 2003, the index of intrinsic motivation to learn mathematics. The U.S. scored near the middle of the distribution on the 2012 index. PISA analysts calculated that, on average, a one unit change in the index was associated with a 19 point gain on the PISA math test. That is the average of within-country calculations, using student-level data that measure the association of intrinsic motivation with PISA score. It represents an effect size of about 0.20—a positive effect, but one that is generally considered small in magnitude.[30]

The unit of analysis matters. Between-country associations often differ from within-country associations. The current study used a difference in difference approach that calculated the correlation coefficient for two variables at the national level: the change in intrinsic motivation index from 2003–2012 and change in PISA score for the same time period. That analysis produced a correlation coefficient of -0.30, a negative relationship that is also generally considered small in magnitude.

Neither approach can justify causal claims nor address the possibility of reverse causality occurring—the possibility that high math achievement boosts intrinsic motivation to learn math, rather than, or even in addition to, high levels of motivation leading to greater learning. Poor math achievement may cause intrinsic motivation to fall. Taken together, the analyses lead to the conclusion that PISA provides, at best, weak evidence that raising student motivation is associated with achievement gains. Boosting motivation may even produce declines in achievement.

Here's the bottom line for what PISA data recommends to policymakers: Programs designed to boost student engagement—perhaps a worthy pursuit even if unrelated to achievement—should be evaluated for their effects in small scale experiments before being adopted broadly. The international evidence does not justify wide-scale concern

National changes on PISA's intrinsic motivation index are not associated with changes in math achievement.

The unit of analysis matters. Between-country associations often differ from within-country associations.

over current levels of student engagement in the U.S. or support the hypothesis that boosting student engagement would raise student performance nationally.

Let's conclude by considering the advantages that national-level, difference in difference analyses provide that student-level analyses may overlook.

1. They depict policy interventions more accurately. Policies are actions of a political unit affecting all of its members. They do not simply affect the relationship of two characteristics within an individual's psychology. Policymakers who ask the question, "What happens when a country boosts student engagement?" are asking about a country-level phenomenon.
2. Direction of causality can run differently at the individual and group levels. For example, we know that enjoying a school subject and achievement on tests of that subject are positively correlated at the individual level. But they are not always correlated—and can in fact be negatively correlated—at the group level.
3. By using multiple years of panel data and calculating change over time, a difference in difference analysis controls for unobserved variable bias by "baking into the cake" those unobserved variables at the baseline. The unobserved variables are assumed to remain stable over the time period of the analysis. For the cultural factors that many analysts suspect influence between-nation test score differences, stability may be a safe assumption. Difference in difference, then, would be superior to cross-sectional analyses in controlling for cultural influences that are omitted from other models.
4. Testing artifacts from a cultural source can also be dampened. Characteristics such as enjoyment are culturally defined, and the language employed to describe them is also culturally bounded. Consider two of the questionnaire items examined above: whether kids "enjoy" math and how much they "look forward" to math lessons. Cultural differences in responding to these prompts will be reflected in between-country averages at the baseline, and any subsequent changes will reflect fluctuations net of those initial differences.

NOTES

1 J.B. Stroud and E.F. Lindquist, "Sex differences in achievement in the elementary and secondary schools," *Journal of Educational Psychology*, vol. 33(9) (Washington, D.C.: American Psychological Association, 1942), 657–667.

2 Christina Hoff Sommers, *The War Against Boys: How Misguided Feminism Is Harming Our Young Men* (New York, NY: Simon & Schuster, 2000).

3 Christianne Corbett, Catherine Hill, and Andresse St. Rose, *Where the Girls Are: The Facts About Gender Equity in Education* (Washington, D.C.: American Association of University Women, 2008).

4 Richard Whitmire, *Why Boys Fail: Saving Our Sons from an Educational System That's Leaving Them Behind* (New York, NY: AMACOM, 2010).

5 Sara Mead, *The Evidence Suggests Otherwise: The Truth About Boys and Girls* (Washington, D.C.: Education Sector, 2006).

6 PIRLS and PISA assess different reading skills. Performance on the two tests may not be comparable.

7 NAEP categories were aggregated to calculate the city/suburb difference.

8 OECD, *Reading for Change: Performance and Engagement Across Countries* (Paris: OECD, 2002), 125.

9 The best example of promoting Finnish education policies is Pasi Sahlberg's *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* (New York: Teachers College Press, 2011).

10 The 2009 endpoint was selected because 2012 data for the enjoyment index were not available on the NCES PISA data tool.

11 The problem of unobserved variables is formally called omitted variable bias.

12 Christina Hoff Sommers, "The Boys at the Back," New York Times, February 2, 2013; Richard Whitmire, *Why Boys Fail* (New York, NY: AMACOM, 2010), 153.

13 J.L. Hawke, R.K. Olson, E.G. Willcutt, S.J. Wadsworth, & J.C. DeFries, "Gender ratios for reading difficulties," *Dyslexia* 15(3), (Chichester, England: Wiley, 2009), 239–242.

14 Daniel Zalewski, "The Background Hum: Ian McEwan's art of unease," The New Yorker, February 23, 2009.

15 These ideas were first introduced in a 2013 *Brown Center Chalkboard* post I authored, entitled, "When Does a Policy Start?"

16 Maria Glod, "Since NCLB, Math and Reading Scores Rise for Ages 9 and 13," *Washington Post*, April 29, 2009.

17 Mark Schneider, "NAEP Math Results Hold Bad News for NCLB," AEIdeas (Washington, D.C.: American Enterprise Institute, 2009).

18 Lisa Guisbond with Monty Neill and Bob Schaeffer, *NCLB's Lost Decade for Educational Progress: What Can We Learn from this Policy Failure?* (Jamaica Plain, MA: FairTest, 2012).

19 Derek Neal and Diane Schanzenbach, "Left Behind by Design: Proficiency Counts and Test-Based Accountability," NBER Working Paper No. W13293 (Cambridge: National Bureau of Economic Research, 2007), 13.

20 Careful analysts of NCLB have allowed different states to have different starting dates: see Thomas Dee and Brian A. Jacob, "Evaluating NCLB," *Education Next* 10, no. 3 (Summer 2010); Manyee Wong, Thomas D. Cook, and Peter M. Steiner, "No Child Left Behind: An Interim Evaluation of Its Effects on Learning Using Two Interrupted Time Series Each with Its Own Non-Equivalent Comparison Series," Working Paper 09–11 (Evanston, IL: Northwestern University Institute for Policy Research, 2009).

21 Common Core State Standards Initiative. "English Language Arts Standards, Key Design Consideration." Retrieved from: http://www.corestandards.org/ELA-Literacy/introduction/key-design-consideration/

22 Twelve states shifted downward from strong to medium and five states shifted upward from medium to strong, netting out to a seven state swing.

23 Tom Loveless, "The Happiness Factor in Student Learning," *The 2006 Brown Center Report on American Education: How Well are American Students Learning?* (Washington, D.C.: The Brookings Institution, 2006).

24 All countries with 2003 and 2012 data are included.

25 Amanda Ripley, *The Smartest Kids in the World: And How They Got That Way* (New York, NY: Simon & Schuster, 2013)

26 Elizabeth Green, *Building a Better Teacher: How Teaching Works (and How to Teach It to Everyone)* (New York, NY: W.W. Norton & Company, 2014).

27 The attitude toward school index is based on responses to: 1) Trying hard at school will help me get a good job, 2) Trying hard at school will help me get into a good college, 3) I enjoy receiving good grades, 4) Trying hard at school is important. See: OECD, PISA 2012 Database, Table III.2.5a.

28 OECD, *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs (Volume III)* (Paris: PISA, OECD Publishing, 2013), 77.

29 PISA originally called the index of intrinsic motivation the *index of interest and enjoyment in mathematics*, first constructed in 2003. The four questions comprising the index remain identical from 2003 to 212, allowing for comparability. Index values for 2003 scores were re-scaled based on 2012 scaling (mean of 0.00 and SD of 1.00), meaning that index values published in PISA reports prior to 2012 will not agree with those published after 2012 (including those analyzed here). See: OECD, *PISA 2012 Results: Ready to Learn: Students' Engagement, Drive and Self-Beliefs* (Volume III) (Paris: PISA, OECD Publishing, 2013), 54.

30 PISA math scores are scaled with a standard deviation of 100, but the average within-country standard deviation for OECD nations was 92 on the 2012 math test.

THE BROOKINGS INSTITUTION

STROBE TALBOTT
President

DARRELL WEST
Vice President and Director
Governance Studies Program

BROWN CENTER STAFF

BETH AKERS
Fellow

MATTHEW CHINGOS
Senior Fellow and Research Director

TOM LOVELESS
Nonresident Senior Fellow

GROVER "RUSS" WHITEHURST
Senior Fellow and Herman and George R. Brown Chair in Education Studies

ELLIE KLEIN
Center and Research Coordinator

KATHARINE LINDQUIST
Research Analyst

LIZ SABLICH
Communications Manager

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

B | 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