



**The 2016 Brown Center Report
on American Education:**

HOW WELL ARE AMERICAN STUDENTS LEARNING?

*With sections on Reading and Math
in the Common Core Era, Tracking
and Advanced Placement (AP), and
Principals as Instructional Leaders*

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 2016 Brown Center Report
on American Education:**

HOW WELL ARE AMERICAN STUDENTS LEARNING?

*With sections on Reading and Math
in the Common Core Era, Tracking
and Advanced Placement (AP), and
Principals as Instructional Leaders*

March 2016
Volume 3, Number 5

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

TABLE OF CONTENTS

3 Introduction

PART I

6 Reading and Math in the Common Core Era

PART II

16 Tracking and Advanced Placement

PART III

26 Principals As Instructional Leaders: An International Perspective

36 Notes

THE 2016 BROWN CENTER REPORT ON AMERICAN EDUCATION

The 2016 edition of the Brown Center Report (BCR) is number five in the third volume and the 15th issue overall. The series began in 2000, making this the fifth consecutive presidential election year in which an issue has been published. As is customary, this year's BCR contains three studies.

Part one is on the Common Core State Standards (CCSS) and instruction in math and reading. National Assessment of Educational Progress (NAEP) data indicate that nonfiction is receiving greater emphasis in both fourth and eighth grade reading instruction, data and geometry are receding in importance in fourth grade math, and course enrollments in eighth grade math are shifting away from advanced courses toward a single, general math course. The CCSS supports all of these trends.

Previous issues of the BCR presented models to classify states by their implementation of CCSS. States that are not followers of CCSS have been reluctant to embrace the changes in curriculum and instruction that are encouraged in those standards. The models also show that CCSS implementation is associated with a change of less than a single NAEP scale score point in both fourth grade reading and eighth grade math. Critics blamed Common Core for disappointing NAEP scores in 2015. The good news for Common Core supporters is that nothing in the analysis supports that charge. The bad news is that there also is no evidence that CCSS has made much of a difference during a six year period of stagnant NAEP scores.

The second section investigates whether tracking in eighth grade is related to Advanced Placement (AP) outcomes in high school. Tracking, the practice of grouping students into different classes based on ability or prior achievement, is a controversial topic. Critics argue that tracking creates or

reinforces social inequities. Middle school is when students first experience tracking, typically in mathematics. In eighth grade, the tracking question currently boils down to whether high achieving students who are ready for a formal algebra course will get one—or whether all students will take the same general math course.

Is middle school tracking related to either AP participation or test scores? State-level tracking data from 2009 and AP data from 2013 are used to tackle the question. States that had a larger percentage of eighth grade students in tracked math classes produced a larger percentage of high-scoring AP students four years later. The heightened AP performance held across racial subgroups—white, black, and Hispanic. There was no relationship between tracking and AP participation. Taken together, these findings suggest the heightened performance was not a result of increased selectivity into AP.

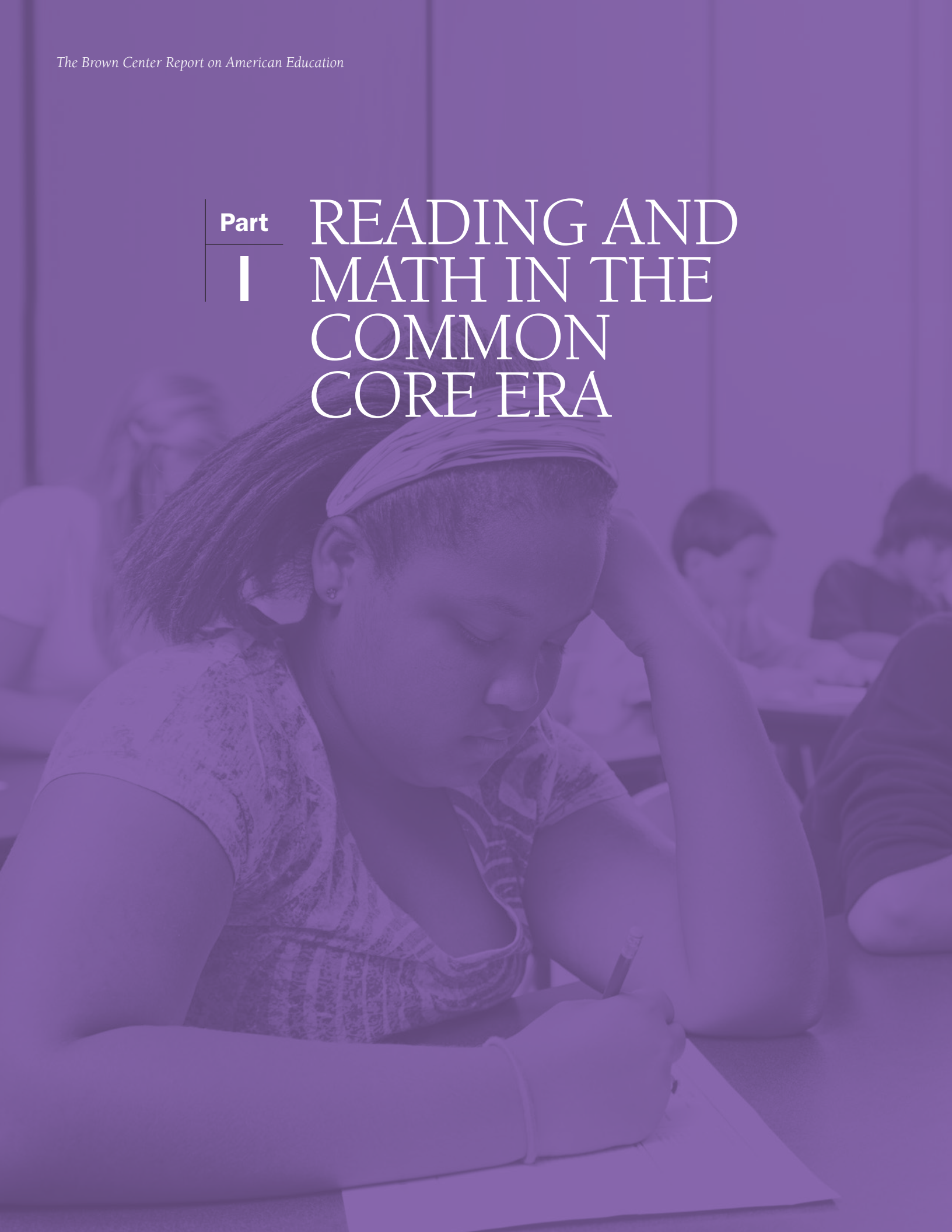
Part three examines school leadership from an international perspective. All around the world, school principals are called on to provide instructional leadership. Data from the Trends in International Mathematics and Science Study (TIMSS) are examined to see whether principals' instructional activities are associated with student achievement. No association was found.

What does instructional leadership look like internationally? Principals are most likely to influence instruction by developing and setting educational goals for their schools. They also monitor teachers' implementation of goals, but are less likely to give advice to teachers who have questions or problems with their teaching. Giving pedagogical guidance may be considered intrusive in some countries. In other countries, teachers may get advice from colleagues. In three consistently high achieving countries—Finland, Hong Kong, and Japan—principals are especially reluctant to give advice; however, principals in Korea, another perennially high achieving country, are more activist in offering instructional guidance.

Part

I

READING AND MATH IN THE COMMON CORE ERA



COMMON CORE STATE STANDARDS (CCSS) HAVE BEEN ADOPTED as the reading and math standards of more than 40 states. All but a few states scheduled full implementation of the standards, including assessments, by the end of the 2014–15 school year. Three states (Indiana, Oklahoma, and South Carolina) have rescinded previous adoptions of the Common Core, and others have made minor revisions. This section of the Brown Center Report (BCR) will exploit the variation in state implementation of CCSS to look at the association of the standards with reading and mathematics performance in grades four and eight, the two grades tested by the National Assessment of Educational Progress (NAEP).

Previous BCRs introduced two models for assessing the effects of CCSS. The models use different indicators and sources of data to create implementation indexes. The indexes sort states into three groups: non-adopters of CCSS, medium implementers, and strong implementers. The 2011 index (IMP11) is based on a 2011 survey of state agencies' use of federal Recovery Act funds to implement CCSS.¹ States were categorized as “strong” implementers if they reported joining a CCSS test consortium and spending money on professional development and new instructional materials (nineteen states). The 2013 index (IMP13) is based on a 2013 survey of state agencies, requesting their timeline for when CCSS would be fully implemented. The states that planned to achieve “classroom implementation of

ELA-CCSS” by the end of the 2012–2013 school year were categorized as strong implementers (12 states).

The analytical approach has both strengths and weaknesses. A strength is modeling data longitudinally, allowing each state to serve, in a sense, as its own control. Analyses of cross-sectional data, as opposed to longitudinal data, are particularly vulnerable to producing spurious correlations. A second strength, especially important in interpreting NAEP gains and losses in 2015, is that the indexes were constructed years prior to the release of the 2015 scores. Advocates and critics of CCSS have labored mightily to present the disappointing 2015 NAEP scores in the most favorable light for their cause. Making up rules for explaining test scores after the scores are known

introduces the usual pitfalls of post hoc analysis, and to do so while participating in a political debate should raise alarm bells about the indifference of the analysis.

As reported in the 2015 BCR, analyses employing the two indexes indicated that states engaged in strong implementation of CCSS registered a one to one and one-half point advantage in NAEP gains from 2009–2013 compared to nonadopters of CCSS. Both indexes were constructed based on the perspectives of state policy authorities, the elites of implementation efforts. They also only reflect the earliest stages of CCSS implementation. This year's study goes deeper by examining reports from teachers and principals on changes occurring inside schools. The data come from surveys conducted as part of NAEP.

Please note that the following analysis does not investigate whether changes in particular practices have caused gains or losses in student achievement. The practices examined here are simply being used as markers for indicating the degree to which CCSS recommendations have penetrated schools and classrooms. The 2011 and 2013 implementation indexes are based on information provided by state policymakers as to the breadth and ambitiousness of state implementation plans. The objective now is to see whether those indexes correspond with reports from practitioners on how CCSS implementation is actually proceeding in schools and classrooms. Scholars of public policy have long known that frontline implementers (sometimes referred to as “street level bureaucrats”) can profoundly shape the end results of top-down initiatives.² In the following analysis, responses of teachers and principals to NAEP questionnaires are aggregated to the state level to measure each state's implementation of CCSS. Before examining the state-level data, let's examine national trends.

Fourth and Eighth Grade Reading: The Dominance of Fiction Wanes

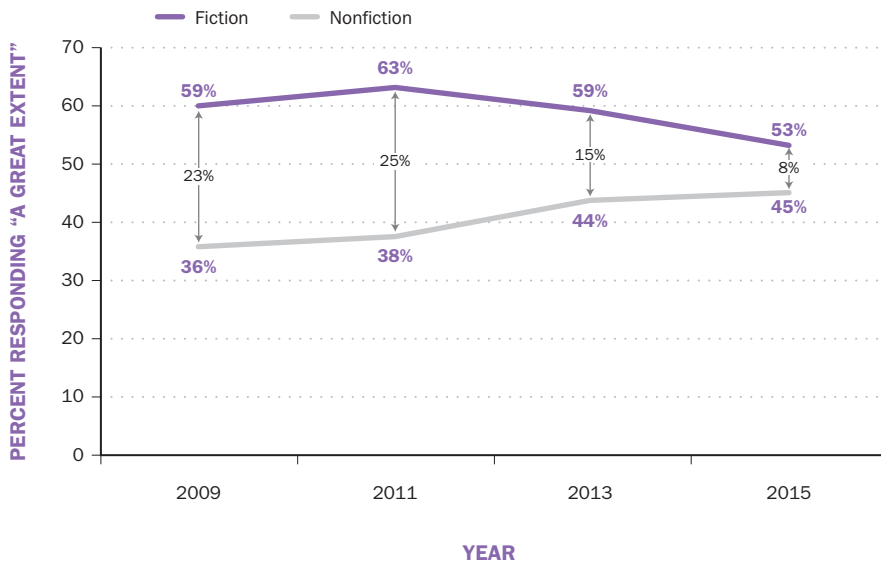
Fiction has long dominated reading instruction. The Common Core standards in English Language Arts (ELA) urge teachers to strike a greater balance between fiction and nonfiction. CCSS-ELA recommends that informational texts, including texts studied in classes other than English, should command 70 percent of students' reading materials by the end of high school. The recommendation proved to be quite controversial, with an extended debate breaking out in 2012.³ Mark Bauerlein and Sandra Stotsky pointed out that research does not support the superiority of studying nonfiction in improving reading skills. Moreover, as an implementation issue, the recommendation runs headlong into deeply entrenched practice. The question is whether teachers will readily abandon the lesson plans they've spent years honing through trial and error—and give up the texts on which those lessons are focused.

But it appears change is happening; the dominance of fiction is waning. Figures 1-1 and 1-2 display the percentage of fourth and eighth grade students with teachers saying they emphasize fiction and nonfiction to a “great extent.” The relative importance of fiction is clear until 2011, when it begins to slip. In fourth grade, the 25 percentage point gap favoring fiction in 2011 declines to 15 percent in 2013 and to eight percent in 2015. In eighth grade, the 34 percent emphasis favoring fiction declines to 24 percent in 2013 and to 16 percent in 2015. Teachers in 2015 were less likely to embrace the superiority of fiction in reading instruction than in the past, and the change is evident in both fourth and eighth grades after 2011.

The question is whether teachers will readily abandon the lesson plans they've spent years honing through trial and error.

Fourth grade emphasis on particular types of reading
(Percentage of students with teachers reporting “heavy emphasis”)

Figure 1-1

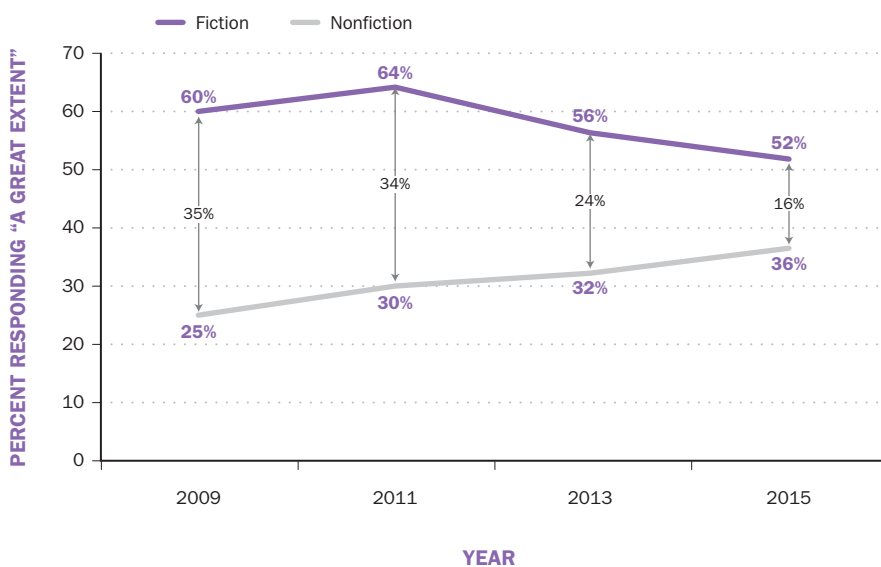


Fourth Grade Math: Less Emphasis on Data Analysis and Geometry

The main NAEP was first administered in 1990 in math and 1992 in reading. Most state mathematics frameworks were modeled after the 1989 standards of the National Council of Teachers of Mathematics (NCTM), with the basic structure of math curriculum stable until adoption of the Common Core. Content was organized by five domains (or strands): number and operations; algebra; measurement; geometry; and data analysis (which includes statistics and probability). The NCTM standards envisioned these five math domains—even algebra—being taught every year from kindergarten through twelfth grade. To this day, NAEP has a similar five-strand structure and awards each strand a different weight.⁴

Eighth grade emphasis on particular types of reading
(Percentage of students with teachers reporting “heavy emphasis”)

Figure 1-2



Common Core takes a more restricted view of mathematics than the five strand approach. That means fewer topics. David Coleman, leader of the Common Core project, has described CCSS as focusing on whole number arithmetic from kindergarten through fourth grade.⁵ Jason Zimba, one of the lead writers of the CCSS math standards states that they “revise the previous ‘strand model’ of mathematics content in order to emphasize arithmetic, algebra, and the connections between them.”⁶

Teachers appear to be responding to the new focus. Fourth grade teachers do not teach as much data and geometry as they once did. Table 1-1 reports the percentage of students with teachers saying that they heavily emphasize geometry and data in the math curriculum. The proportions stayed fairly steady until 2011, when they began to decline. Neither domain received as much attention in 2015 as in 2011 or prior years.

The role of geometry and data in fourth grade math
(Percentage of students with teachers reporting “heavy emphasis”)

Table
1-1

	2009	2011	2013	2015
Geometry	36	36	31	29
Data	31	30	24	23

Source: NAEP Data Explorer

Enrollment in eighth grade math courses (percent of students)

Table
1-2

	2000	2005	2009	2011	2013	2015
Advanced	27	42	44	47	48	43
General 8	38	25	25	25	26	32
Difference	-11	17	19	22	22	11

Source: NAEP Data Explorer.

Note: “Advanced” includes: Algebra I, geometry, and either the first or second year of a two-year algebra sequence. “General 8” consist of either eighth grade or basic math.

Eighth Grade Math: A Shift in Course Taking

The first decade of the 21st century witnessed a dramatic change in the math courses taken by eighth graders. An “algebra for all in eighth grade” movement was bolstered by the rising belief that middle school students should take more rigorous math classes. Enrollment in advanced courses—mainly Algebra I, but also other forms of algebra and geometry—steadily increased. Enrollment in general eighth grade math courses fell. As shown in Table 1-2, general math was the dominant course in 2000, with a 38 percent to-27 percent enrollment advantage over advanced math. In the following decade, enrollment in advanced classes boomed. Advanced math courses held a 47 percent to 25 percent advantage in 2011.

But something happened around 2011. From 2011 to 2013, the relative growth of advanced courses stopped dead in its tracks. Then, from 2013 to 2015, enrollment in advanced math declined from 48 percent to 43 percent. Enrollment increased from 26 to 32 percent in general math.

Common Core supports this recent trend by delineating a single eighth grade math course for all students. Options for accelerating mathematically precocious middle school students by compacting three years into two are discussed in an appendix to CCSS, but the standards themselves lay out a single eighth grade math course that all eighth graders are expected to take. As noted in the study of tracking and Advanced Placement in this report, opponents of tracking have lauded, as a step towards equity, placing all eighth graders in the same math course.

A change occurred in four key practices related to curriculum and instruction.

Is Common Core Related to Changes in Practice?

These national trends indicate that, concurrent with CCSS's implementation, a change occurred in four key practices related to curriculum and instruction. By itself that does not constitute convincing evidence that CCSS is connected to the trends, but additional evidence strengthens the case. Let's start with the obvious: CCSS endorses all of these particular changes in practice. Moreover, three of the changes—the shift away from fiction at both fourth and eighth grades and the movement away from advanced math classes in eighth grade—have provoked widespread public debate. In those debates, policymakers stated that the changes were compelled by Common Core.

The California State Board of Education, for example, rescinded its policy promoting universal eighth grade algebra when it adopted the CCSS math standards.⁷ On the 2013 NAEP, 73 percent of California eighth graders were enrolled in advanced math courses, a number that plummeted

to 45 percent in 2015. In 2013, only seven percent of students were enrolled in a general eighth grade math class. In 2015, the share of students in general math jumped to 32 percent. EdSource, a California publication devoted to education topics, dubbed the trend, “Retreat from Universal 8th Grade Algebra.”⁹

Table 1-3 displays the changes in practice in light of the BCR's 2013 implementation model. To calculate the data, first, state-level changes in the four practices were standardized by conversion to z-scores, with a mean of 0.00 and a standard deviation of 1.00. The base year for each period of change reflects the point in time when, as indicated in the charts above, a reversal in the prevailing national trend occurred. For eighth grade math courses, changes occurring from 2013 to 2015 are reported. For the other three practices, 2011 to 2015 is the time period under scrutiny.

Because a z-score of 0.00 represents change in the average state, the signs of the values in the table are important. Positive scores indicate movement in accord with CCSS, and negative scores indicate change in practice contrary to CCSS recommendations. The z-score of -0.31 in the first column, for example, indicates that fourth grade teachers in non-CCSS states were, relative to the average state, less likely to de-emphasize fiction in favor of nonfiction. They went against Common Core. States with medium (0.03) and strong (0.11) implementation of CCSS, on the other hand, were slightly more likely than the average state to place a greater emphasis on nonfiction.

The two math practices clearly differentiate the states. The fourth grade nonadopters ($z = -0.76$) are reluctant to de-emphasize geometry and data, whereas the strong implementers are more likely to do so (0.32). The eighth grade nonadopters are more likely

Change in school and classroom practices, by implementation status

(State standardized changes, z-scores. Implementation status from IMP13 model)

Table
1-3

Implementation status	4 th grade reading: fiction v. nonfiction (2011–2015)	4 th grade math: geometry and data (2011–2015)	8 th grade reading: fiction v. nonfiction (2011–2015)	8 th grade math course: advanced math v. general math (2013–2015)
Nonadopters (n = 7)	-0.31	-0.76	-0.28	-0.61
Medium implementers (n=32)	0.03	0.09	0.15	0.03
Strong implementers (n=11)	0.11	0.32	-0.25	0.33

Source: calculations by author using data from NAEP Data Explorer.⁸

Note: Implementation status n's are for reading practices. For math, Minnesota switches from strong implementer (n=10) to nonadopter (n=8).

to continue offering advanced math classes (-0.61) while the strong implementers of CCSS are increasingly favoring a single, general math class (0.33).

Take note of differences among the table's rows. Changes in these four key practices differentiate the nonadopters from states embracing CCSS, but they are not as successful in drawing a contrast between strong and medium implementers. The eighth grade reading practice (emphasis on fiction vs. nonfiction) is particularly weak in this regard, with strong implementers looking more like nonadopters than medium implementers. In a way, this makes sense. When the brouhaha over nonfiction erupted in 2012, defenders of the Common Core argued that the recommendation applied to reading materials taught in all middle and high school classes—history, science, and art—and not only to those taught by language arts teachers. Perhaps ELA eighth grade teachers in strong CCSS states do not feel the same urgency to use nonfiction texts as their fourth grade counterparts who, after all, typically teach all subjects in a self-contained class. The decision to give nonfiction texts greater prominence in the fourth grade curriculum affects the fourth grade teacher alone.

Are Changes in NAEP Scores Associated with CCSS Implementation?

Tables 1-4 and 1-5 present changes in NAEP fourth grade reading scores with states categorized by implementation status. Table 1-4 sorts the states using the IMP11 model and Table 1-5 uses IMP13. The last column in the tables presents data covering the six year period, 2009–2015. Regardless of implementation status, states made only slight progress in reading. The medium implementers made the largest gains (2.15) in IMP11 and the nonadopters made the

Changes in NAEP fourth grade reading (in scale score points)
(Using 2011 Implementation Index)

Table
1-4

Implementation rating	2009–2011	2011–2013	2013–2015	ALL 2009–2015
Strong (n=19)	0.22	0.64	0.45	1.31
Medium (n=27)	0.17	0.81	1.17	2.15
Nonadopters (n=4)	-0.78	0.53	2.07	1.82

Changes in NAEP fourth grade reading (in scale score points)
(Using 2013 Implementation Index)

Table
1-5

Implementation rating	2009–2011	2011–2013	2013–2015	ALL 2009–2015
Strong (n=11)	0.19	1.08	0.43	1.70
Medium (n=32)	0.20	0.62	0.75	1.56
Nonadopters (n=7)	-0.78	0.53	2.77	2.52

largest gains in IMP13 (2.52). The gains for all three implementation categories fall within a single NAEP scale score point of each other. The standard deviation (sd) of the 2015 NAEP fourth grade reading scores is 37 points. One point is less than 0.03 sd units. That is a miniscule difference, especially if taking six years to emerge. The NAEP fourth grade reading assessment has experienced other six-year intervals of underperformance, most notably from 1992 to 1998, when scores were flat.

It is interesting that CCSS's implementers and nonadopters have experienced different periods of superior performance. In the early years of CCSS implementation (2009–2011 and 2011–2013), both strong and medium implementers made larger gains than the nonadopters in fourth and eighth grades. In 2013–2015, the

Curriculum and instruction are changing at the ground level of schooling.

Changes in NAEP eighth grade math (in scale score points)
(Using 2011 Implementation Index)

Table 1-6

Implementation rating	2009–2011	2011–2013	2013–2015	ALL 2009–2015
Strong (n=19)	1.65	0.23	-2.37	-0.49
Medium (n=26)	0.39	0.61	-2.22	-1.22
Nonadopters (n=5)	1.30	-0.69	-0.98	-0.37

implementers in the 2013 model (-0.08). As in reading, different periods of superiority are evident. Medium or strong implementers notched larger gains from 2011–2013; nonadopters outpaced medium implementers in 2009–2011. The 2013–2015 interval may be a turning point, with nonadopters’ scores declining, but declining less than the scores of CCSS states. Future NAEP scores will tell the tale.

Changes in NAEP eighth grade math (in scale score points)
(Using 2013 Implementation Index)

Table 1-7

Implementation rating	2009–2011	2011–2013	2013–2015	ALL 2009–2015
Strong (n=10)	1.86	0.47	-2.41	-0.08
Medium (n=32)	0.62	0.45	-2.29	-1.22
Nonadopters (n=8)	1.30	-0.69	-1.30	-0.69

Summary and Conclusion

This section of the Brown Center Report investigated the Common Core. Recent trends in four practices were charted using NAEP data—the teaching of nonfiction in both fourth and eighth grades, the amount of emphasis given to data and geometry in fourth grade math, and course enrollments in eighth grade math. The CCSS takes a position on all four activities. The data collected from teachers and principals on NAEP indicate that these four practices have recently begun trending in the direction recommended by the CCSS: teachers are giving greater emphasis to nonfiction texts at both fourth and eighth grades; data and geometry are receiving less attention in fourth grade math; and at eighth grade, enrollments are falling in advanced math courses and rising in general math courses. Curriculum and instruction are changing at the ground level of schooling.

nonadopters made larger gains, two points or more in the 2013 model. NAEP changes that occur in a single two-year interval can easily evaporate, so whether this potential trend continues in 2017 NAEP data will be important. The discussion below returns to this question.

Tables 1-6 and 1-7 examine NAEP scores in eighth grade math. Again, the states are grouped by implementation status as defined by the 2011 and 2013 implementation models. The three implementation categories all experienced small declines in NAEP performance from 2009–2015. To put this event in context, the smallest gain in any previous six year period on the eighth grade NAEP test in math has been four points. A loss has never happened before. Nonadopters register the smallest loss in the 2011 model (-0.37) and strong

States vary on how they look individually on the CCSS reforms. Previous installments of the Brown Center Report introduced models for evaluating the implementation of CCSS. The models sort states into three groups: nonadopters of CCSS, medium implementers, and strong implementers. As one would expect, the nonadopters appear more resistant to the CCSS curricular and instructional changes compared to states categorized as having

Regardless of CCSS implementation status, states made only slight progress in reading.

either medium or strong CCSS implementation. The strong implementers were the states most likely to embrace CCSS reforms, with the exception of a greater emphasis on nonfiction texts in eighth grade. The medium implementers were the states most accepting of that reform.

The implementation models were used to analyze NAEP scores from 2009–2015. The models show that CCSS implementation is associated with a change of no more than a single point (plus or minus) in NAEP for both fourth grade reading and eighth grade math scores. Here's how that finding was calculated. Dividing the CCSS adopters into two groups allows for two separate comparisons with the non-adopters. For the eight comparisons produced for 2009–2015, the largest advantage for CCSS adopters is +.61 (Table 1-7, comparing strong implementers with nonadopters on eighth grade math scores). The largest advantage for nonadopters is +0.96 (Table 1-5, comparing nonadopters with medium implementers on fourth grade reading scores). A change of one NAEP scale score point is trivial—and especially so over six years.

This study's analytical approach has weaknesses. It cannot determine causality. States engage in a lot of policymaking when it comes to education, so declaring that CCSS or any other singular policy caused a change in NAEP scores is unwarranted. Grouping the states by strength of implementation helps mitigate the effect of other policies (assuming that non-CCSS policies are distributed randomly over groups). But the nonadopter group is too small—only four states in 2011 reading—to have confidence in the assumption. Using changes in NAEP scores over two year intervals helps control for demographic changes—demographic influences should be baked into baseline scores and not change much

over two years—but such influences could bias calculations for the six-year interval of 2009–2015.¹⁰ The approach is also statistically unsophisticated. Future evaluations of CCSS will probably take a similar approach, modeling variation among states (and maybe districts), then calculating test score changes over time. But they also may collect richer data sets and employ methods that go beyond the current study's simple comparisons of group means.

An intriguing finding from the study pertains to implementation. Adopters and nonadopters of CCSS appear to have experienced different periods of superiority on NAEP. The adopters registered larger NAEP gains early in the implementation process, 2009–2013, whereas the nonadopters look better in the most recent interval, 2013–2015 (in the case of math, by registering smaller losses). Many advocates of CCSS have a theory of implementation that believes these standards are so new, so revolutionary, so different from what teachers have experienced previously that Common Core won't bear fruit for many years. Professional development, new textbooks, teaching that targets “deeper learning,” and all the other accoutrements of CCSS take time to unfold. Just wait, good things will happen, proponents say. Michael Kirst, chair of the California State School Board, probably gave the most honest statement of this theory by saying it will take at least another five years (meaning 2020), a full decade after California's 2010 adoption of CCSS, before the state knows if Common Core “works.”¹¹

The data above support a competing hypothesis. Maybe CCSS has already had its best years and additional gains will be difficult to attain. Major top-down reforms can have their strongest effects when first adopted, whether it's the NSF-funded

Maybe CCSS has already had its best years and additional gains will be difficult to attain.

science and math curricula of the 1960s, including New Math, or the more recent No Child Left Behind Act. Policy elites rally around a new policy, advocates trumpet the benefits that will occur, a public relations campaign is launched to garner support, and local educators respond enthusiastically to the new reforms. New Math started with a bang, but as criticism grew and teachers' support dissipated, the materials fell out of use.¹²

The 2015 NAEP scores were a political disaster for Common Core. Eighth grade

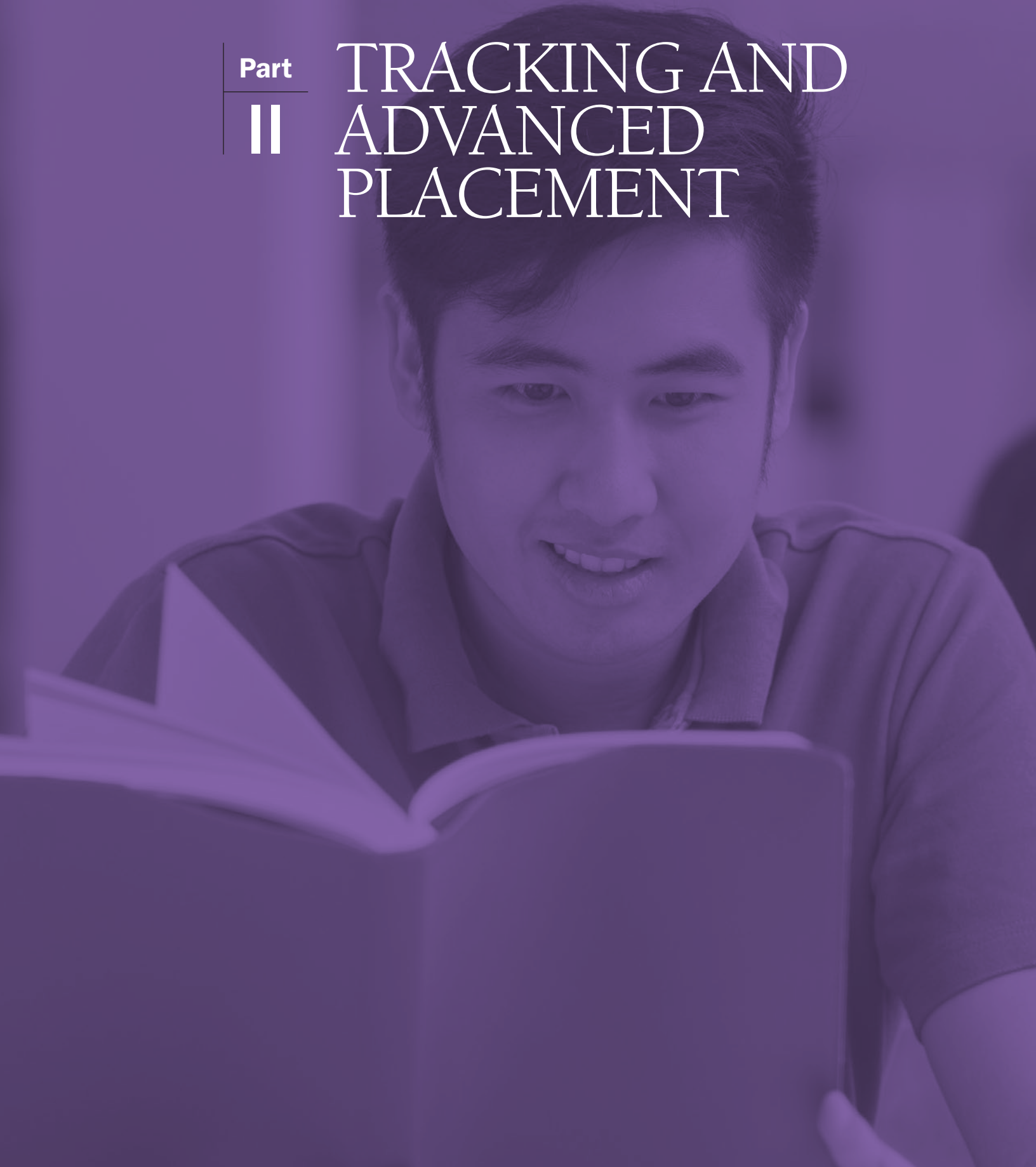
math scores, for example, fell for the first time in NAEP's 25 year history (down three points). Some observers were quick to point a finger at CCSS. That's probably unfair. The analysis above indicates that, yes, nonadopters performed better than CCSS states, but only by declining less, not through improved performance. None of the states are setting the world on fire. Whatever is depressing NAEP scores appears to be more general than the impact of one set of standards or another.

Whatever is depressing NAEP scores appears to be more general than the impact of one set of standards or another.

Part

II

TRACKING AND ADVANCED PLACEMENT



THIS SECTION PRESENTS A TIME-LAGGED ANALYSIS OF THE relationship between tracking—the practice of assigning students to different academic classes based on prior achievement¹³—in eighth grade and two later outcomes related to the Advanced Placement program (AP): participation rates and successful performance on AP tests in high school. The theory motivating the analysis is that academically advanced students may gain long term benefits from accelerated coursework in middle school. Just as star high school athletes do not walk onto a basketball court or football field for the first time as seniors in high school, successful AP calculus students do not encounter advanced mathematics for the first time in 12th grade.

Preparation matters. In communities across the country, pipelines are in place to nurture and develop promising young athletes.¹⁴ Not so with academic stars. Why not? In a word, because singling out advanced students for special coursework involves *tracking*. Accelerated or honors courses, offering above grade level curricula to students who are ready for it, typically start in middle school. They allow high-achieving youngsters to move at a faster pace than their grade level peers.

But tracking is controversial. By definition, it involves differentiating students in terms of their skills and knowledge.

Black, Hispanic, and socioeconomically disadvantaged students are historically underrepresented in accelerated tracks. As such, the charge that tracking discriminates against these students has shaped the frequency of its use across different communities. Tracking is more prevalent in suburban middle class communities and in schools serving white and Asian students and less prevalent in urban schools and schools serving predominantly black, Hispanic, or disadvantaged populations.

Whether middle school tracking is associated with AP outcomes is a timely question. Recent research on tracking that

employs techniques to minimize selection bias and other shortcomings of previous research, has documented examples of tracking being used to promote equity. AP classes, along with the International Baccalaureate program, represent the pinnacle of advanced coursework in U.S. high schools. They are the end of the pipeline preparing academically gifted students for college. Boosting access to AP classes for groups historically underrepresented in AP is a key element of the contemporary equity agenda for high schools. In opposition to these trends, tracking's critics remain steadfast. The advent of the Common Core State Standards (CCSS) may furnish critics with a politically powerful shield for dismantling tracking in middle schools (see the study of Common Core in this issue).

Background

In the 1970s, the charge that tracking produces discriminatory social effects rose to public awareness just as tracking itself was changing. Since the early 20th century, curriculum differentiation occurred by assigning students to tracks that encompassed all academic subjects. The names of tracks vaguely denoted post-secondary destinations, with “college prep,” “vocational,” and “general” being the most common labels. Students were assigned to tracks based on IQ tests measuring general aptitude or achievement tests measuring prior learning. By the 1970s, tracking had changed. Omnibus tracking was replaced by subject-specific assignment to courses (i.e., students simultaneously could be placed in remedial reading and a higher level math class), IQ testing fell into disfavor, and parents increasingly could override schools’ initial placement and demand a different track if they wanted more or less challenge for their children than schools recommended.

The changes did not reduce the attacks on tracking. In 1985, Jeannie Oakes’ “Keeping Track” was published. Oakes acknowledged that tracking had changed but dismissed the modifications as trivial. Schools, Oakes charged, were still systematically denying kids opportunity in ways that correlated with race and class. Oakes built her critique on the theories of Marxian analysts Samuel Bowles and Herbert Gintis, whose 1976 book, “Schooling in Capitalist America,” argued that schools are structured with the intention to reproduce social inequalities. Despite its ideological underpinnings, the tracking critique drew surprising support across the political spectrum. In “What Do Our 17-Year-Olds Know?”, two former members of the Reagan and Bush administrations respectively, Checker Finn and Diane Ravitch single out tracking as a cause of students’ poor performance on National Assessment of Educational Progress (NAEP) tests of history and literature.¹⁵

The anti-tracking movement gained steam in the 1990s. It had little effect on high schools, but middle schools were another story. Across the country, middle schools began paring back tracking, especially in English-language arts, science, and history. By the end of the decade, a majority of middle school students attended heterogeneously grouped classes in those subjects.¹⁶ Math classes remained tracked, but with fewer levels—typically just one level offering algebra and one level offering pre-algebra or a general eighth grade math course. The frequency of tracking in academic subjects remains similar today.

Recent Research on Tracking and Equity

A challenge to research on the effects of tracking has been adequately controlling for selection effects. In this case, the term

The anti-tracking movement gained steam in the 1990s. It had little effect on high schools, but middle schools were another story.

Recent research indicates that high-achieving students may benefit from tracking.

“selection effect” refers to the nonrandom assignment of students to tracks. High- and low-track students are assigned to their respective tracks because of different amounts of prior learning and the anticipation of different amounts of future learning. To discover that high-track students learn more than low-track students may simply be an extension of how the students were “selected” into respective tracks in the first place and may have nothing to do with tracking itself. In addition, schools do not make policy choices randomly, and they may have decided to track or to heterogeneously group students for reasons related to achievement.

Experiments in which students are randomly assigned to tracked and untracked settings are rare. In 2005, an experiment in Kenya could be conducted because schools were granted extra funds to hire first grade teachers.¹⁷ More than a hundred schools (121) had only one first grade teacher, and the new money allowed the addition of a second teacher. The schools were randomly assigned to either a tracked or untracked condition. In the tracked schools, one of the classes was made up of higher achievers, the other of lower achievers. Students were placed in either the higher- or lower-achieving class based on whether they scored above or below the median for all students. Students in the untracked schools were assigned to the two classes randomly, creating classes heterogeneous in ability.

The experiment ran for 18 months. Both high- and low-achievers in the tracked schools gained more on achievement tests compared to students in the untracked schools. The benefit for students in higher-achieving classes was 0.19 standard deviations and for those in the lower-achieving classes, 0.16 standard deviations.

Conditions that allow for experiments are quite unique, so analysts have also used

quasi-experimental designs to evaluate tracking. Takako Nomi investigated a 1997 policy in Chicago that abolished remedial math classes in ninth grade and created mixed-ability algebra classes in their place. Employing an interrupted time-series design and difference-in-differences analysis, Nomi found that high achievers paid a price for abandoning tracking in favor of heterogeneously grouped classes. An analysis of class composition using instrumental variables indicated that peer effects were driving much of the effect. A one standard deviation decline in peer skills was associated with about a one-quarter standard deviation decline in high achievers’ test scores.¹⁸

David N. Figlio and Marianne E. Page (2000) also used an instrumental variable strategy to isolate the effects of tracking. They found that wealthier families consider whether a school tracks when making enrollment decisions. After controlling for those parental decisions, Figlio and Page found that disadvantaged students benefited from tracking, contradicting the notion that abolishing tracking promotes equity. As they put it, “...tracking programs are associated with test score gains for students in the bottom third of the initial test score distribution. We conclude that the move to end tracking may harm the very students it is intended to help.”

Chao Fu and Nirav Mehta (2015) looked at tracking using data from the Early Childhood Longitudinal Study, a large national database. In contrast to Figlio and Page, they found a trade-off, with tracking benefitting high-ability students and hurting low-ability students. Defining low- and high-ability students in the same manner as the study in Kenya (above and below the median of achievement), Fu and Mehta’s model predicts that de-tracking would raise the test scores of low-achieving students by

0.04 standard deviations and depress high-achievers' scores by 0.05 standard deviations.

David Card and Laura Giuliano (2014) studied the effects of gifted classes in a large Eastern school district. The district had mandated that schools with even a single gifted student (most of whom were identified by IQ tests) must provide separate gifted classes in fourth and fifth grades, with open seats in these classes filled by high achievers—the school's highest performers on the annual state assessment. The policy dramatically increased the proportion of disadvantaged students in the gifted classes to about 40 percent districtwide. The researchers found significant positive effects for high achievers in the program, in particular for low-income black and Hispanic students. Card and Giuliano concluded, "Our findings suggest that a comprehensive tracking program that establishes a separate classroom in every school for the top-performing students could significantly boost the performance of the most talented students in even the poorest neighborhoods, at little or no cost to other students or the District's budget."¹⁹

In sum, recent research indicates that high-achieving students may benefit from tracking and suffer losses from heterogeneous grouping. The studies have primarily assessed achievement effects from one to two years of attending high tracks. The following study takes a longer perspective and examines outcomes at the end of high school that may be associated with tracking in eighth grade.

Data

The analysis below examines data from the national cohort of students who were eighth graders in 2009 and graduated from high school in 2013.²⁰ Data on eighth grade

tracking come from the National Assessment of Educational Progress.²¹ The percentage of students attending schools with tracked eighth grade math classes, aggregated to the state level, serves as a proxy for middle school tracking practices. Data on AP participation and performance come from the "Tenth Annual AP Report to the Nation."²² As with the NAEP data, state-level data are used in the analysis. AP participation refers to the percentage of each state's public high school graduates who took at least one AP exam during high school. AP performance represents the percentage of each state's AP test takers who scored a three or better on at least one test. That is the typical threshold that colleges and universities require for granting college credit.

The data possess several limitations. Tracking practices are modeled using information from only one subject. Math is the most commonly tracked subject in middle schools, but using data from other subjects might yield different results. Taking an AP test is not the same as taking an AP course. Some students take AP courses but do not sit for the AP exam. Students are counted as AP participants if they took an AP exam at any point in their high school careers. Students who took multiple AP tests only count as one test taker in the data, and the count of students scoring three or higher (3+) are those who did so on any single AP test, regardless of their scores on other AP tests.

The initial research question this study examines is: Were state tracking practices for eighth graders in 2009 related to AP outcomes in 2013? A question pertinent to equity will also be explored: Do the results vary by race? AP outcomes for black, Hispanic, and white students are scrutinized.

The following study examines outcomes at the end of high school that may be associated with tracking in eighth grade.

AP outcomes (2013), tracking in eighth grade (2009), and demographic variables (2009)
(Percentages)*

**Table
2-1**

State	AP participation of HS grads (2013)	AP test-takers scoring 3+ (2013)	Students tracked in 8 th grade (2009)	Students scoring NAEP advanced (2009)	Child poverty rate (2009)
Alabama	25	43	65	4	25
Alaska	23	65	85	6	13
Arizona	25	59	80	6	23
Arkansas	46	35	50	4	27
California	41	66	88	5	20
Colorado	39	62	91	10	17
Connecticut	39	74	90	10	12
Delaware	31	55	64	6	16
District of Columbia	56	25	63	2	29
Florida	53	51	90	6	21
Georgia	40	54	67	5	22
Hawaii	29	41	67	4	14
Idaho	20	66	94	8	18
Illinois	32	66	80	7	19
Indiana	35	46	84	7	20
Iowa	18	62	82	7	16
Kansas	17	61	75	8	18
Kentucky	32	51	74	5	26
Louisiana	15	35	54	4	24
Maine	36	63	66	8	17
Maryland	47	63	94	12	12
Massachusetts	39	72	79	17	13
Michigan	26	65	69	7	23
Minnesota	32	64	87	13	14
Mississippi	13	35	52	2	31
Missouri	16	60	74	7	21
Montana	21	63	79	10	21
Nebraska	17	58	88	8	15

State	AP participation of HS grads (2013)	AP test-takers scoring 3+ (2013)	Students tracked in 8 th grade (2009)	Students scoring NAEP advanced (2009)	Child poverty rate (2009)
Nevada	32	53	97	5	18
New Hampshire	24	76	86	11	11
New Jersey	32	74	74	14	13
New Mexico	28	43	76	3	25
New York	38	67	63	8	20
North Carolina	31	62	83	9	23
North Dakota	15	61	63	7	13
Ohio	23	65	74	8	22
Oklahoma	22	50	65	3	22
Oregon	24	63	85	8	19
Pennsylvania	24	66	84	10	17
Rhode Island	26	55	66	6	17
South Carolina	29	60	84	7	24
South Dakota	20	61	68	7	19
Tennessee	19	53	78	4	24
Texas	36	52	57	8	24
Utah	36	70	89	7	12
Vermont	31	70	60	13	13
Virginia	44	64	73	8	14
Washington	34	61	82	11	16
West Virginia	22	42	65	2	24
Wisconsin	32	70	72	8	17
Wyoming	17	59	89	7	13
Average	29	58	76	7	19
Standard deviation	10	11.2	11.7	3.1	5
Range	13–56	25–76	50–97	2–17	11–31

*All data are rounded to the nearest percent. Original values available upon request.

Analysis

Table 2-1 displays the study's data, with summary statistics reported in the bottom rows. The state average for AP participation in 2013 was 29 percent, meaning that for the typical state almost three out of 10

graduates in the class of 2013 had taken an AP exam at some point during their high school years. Participation rates ranged from a low of 13 percent in Mississippi to a high of 56 percent in the District of Columbia.

Relationship of eighth grade tracking (2009) to AP participation and scores (2013)
(Correlation coefficients)

Table 2-2

	AP participation	Scoring 3+ on AP	Adjusted participation	Adjusted 3+ scoring
All	0.09	0.52**	N/A	N/A
Black	0.06	0.41**	0.05	0.51**
Hispanic	-0.21	0.31*	-0.23	0.43**
White	0.00	0.41**	N/A	N/A

**p<.01, *p<.05

Adjusted = Dropped states with fewer than 50 AP tests takers. Blacks (n=43): AK, ID, MT, NH, ND, SD, VT, WY were dropped. Hispanics (n=47): MT, ND, SD, VT were dropped.

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.

In the average state, more than half (58 percent) of students who had taken an AP exam earned a score of three or higher. The lowest 3+ rate was registered by the District of Columbia (25 percent) and the highest by New Hampshire (76 percent), suggesting a possible trade-off between heightened access to AP and selectivity. As just mentioned, D.C.'s participation rate was the highest in the country; New Hampshire ranked 35th.

The contrast is merely suggestive. The data do not allow for one to tease out whether access and selectivity are inversely related. Trade-offs made by educators at the school or district levels may be masked by aggregating data to the state level. Further research is needed using school or district data, collected, in other words, at the policy-making level where AP offerings are decided.

The popularity of tracking in eighth grade math is evident. The average state tracked about three-quarters of its math students, with Arkansas the least tracked state (50 percent) and Nevada the most tracked (97 percent). The percentage of eighth graders scoring at the "advanced" performance level on the 2009 NAEP math test is included as a control variable. Notice

how stringent the NAEP advanced level is. The average state has only about 7 percent of eighth graders scoring at this level. Prior achievement is an important covariate in any model predicting academic outcomes, whether the outcomes of interest are measured at the individual, school, or state level. Considering the current study's focus on high achievers, a state's percentage of students reaching the NAEP advanced level is an appropriate control. States that had a lot of high-achieving eighth graders in 2009 probably also had a lot of high-achieving high school graduates in 2013—and that will surely influence the AP outcome variables. The final column shows the percentage of children in poverty for each state.

Table 2-2 reports correlation coefficients for the relationship of eighth grade tracking to AP outcomes. Correlations are also reported for AP outcomes disaggregated by race and ethnicity. Eighth grade tracking shows no statistically significant relationship with AP participation. The percentage of a state's graduating class that has taken an AP test is unrelated statistically with the amount of tracking going on four years earlier. Tracking is significantly

Tracking is significantly correlated with performance on AP tests.

The positive relationship holds for the performance of black, Hispanic, and white subgroups.

Regression output: Modeling percentage of 2013 high school graduates scoring 3+ on AP

Table 2-3

	Coefficient	Standard error	T-statistic
Intercept	0.431	0.111	3.881
Tracking—2009	0.002	0.001	2.192*
Scoring advanced on NAEP—2009	0.017	0.004	4.245**
Poverty—2009	-0.668	0.273	2.451*

*p<.05, **p<.01

Model fit: R² = 0.670, F = 31.75, p<.001

correlated with performance on AP tests, and the positive relationship holds for the performance of black, Hispanic, and white subgroups.

States with larger percentages of tracked eighth graders produce larger percentages of high-scoring AP test takers. States where tracking is less prevalent tend to have a smaller proportion of high scorers. Highly tracked states with an above average share of 3+ AP scorers include: California (88 percent tracked), Colorado (91 percent), Connecticut (90 percent), Maryland (94 percent), Minnesota (87 percent), and Utah (89 percent). States with sparser eighth grade tracking and a below average proportion of high-scoring AP students include: Delaware (64 percent tracked), District of Columbia (63 percent), Louisiana (54 percent), Mississippi (52 percent), and Texas (57 percent).

The significantly positive correlations for black and Hispanic high performers on AP are important for equity considerations. Two sets of figures are presented. The adjusted correlations were calculated after dropping states with fewer than 50 AP participants. The number of black AP test takers fell below that criterion in eight states;

for Hispanics, the shortfall occurred in four states. All states had at least 50 white AP tests takers, which is why adjusted figures for whites are not presented. States with small numbers of participants may produce unstable AP scores. AP has dramatically increased the participation of black and Hispanic students in the past decade—and continues to push for greater participation—so the adjusted figures are probably better indicators of future statistical relationships.

Let's consider the pipeline hypothesis, the idea that eighth grade tracking offers high achieving students an opportunity for acceleration that can pay off in high school. The current study cannot test the causal claims of the hypothesis, but the findings do support further research on the topic. States with a larger percentage of kids scoring 3 or better on AP tests in 2013 had a larger percentage of kids in tracked classes four years earlier. That association occurs without any apparent increase in selectivity. The relationship of tracking with AP participation is indistinguishable from zero. Moreover, the finding holds for black, Hispanic, and white subgroups. If eighth grade tracking operates in a manner discriminatory to blacks and

Hispanics, it is not apparent here. The sign of the correlation for Hispanic participation in AP tests is negative, however; and even though the value doesn't reach statistical significance, it should be investigated further with more precise data and hierarchical models that can tease out state, district, and school effects.

Regression analysis is useful for parsing out the influence that confounding variables may exercise in making two variables appear correlated when they in fact are not. Table 2-3 exhibits regression output controlling for two potential confounders. Tracking in eighth grade maintains a significantly positive relationship with later AP performance even while controlling for states' advanced achievement on NAEP and level of child poverty. As expected, both control variables are also statistically significantly associated with AP performance. Neither regression nor correlation coefficients are sufficient to determine causality.

To put the tracking coefficient in simpler terms, an increase of 10 percentage points in 8th grade tracking is associated with a two percentage point increase in high performing AP students. That effect is equivalent to about 0.18 standard deviations. The increase associated with boosting tracking by ten percentage points is over 1,300 additional high scoring AP students in New York and more than 2,000 in Texas. Nationally, a ten percentage point increase in eighth grade tracking is associated with an additional 20,000 students scoring 3 or higher on AP exams.

Conclusion

This section of the Brown Center Report examined the relationship of eighth grade tracking in 2009 with two AP outcomes in 2013: participation and high performance on AP tests. State level data were analyzed.

No association was found between the percentage of a state's students who were tracked in eighth grade mathematics and—four years later—the percentage of graduating seniors who had taken an AP test. A positive relationship was found between tracking and superior performance on AP tests, the percentage of test takers scoring a 3 or better on AP tests. The positive relationship was statistically significant for white, black, and Hispanic students.

The analysis cannot prove or disprove that tracking caused the heightened success on AP tests. The findings do support future research on the hypothesis that tracking benefits high achieving students—in particular, high achieving students of color—by offering accelerated coursework that they would not otherwise get in untracked schools. That hypothesis is supported by several recent studies, as described above, including that of David Card and Laura Giuliano (2014).

The hypothesis that middle school tracking is associated with AP outcomes rests on the notion of an academic pipeline—that superior academic performance must be nurtured and developed over time. Think of how the following three phenomena coalesce to shape opportunity. First, students are assigned to tracks primarily based on achievement test scores. Because of the test score gaps between white and Asian students, on the one hand, and black and Hispanic students, on the other hand, honors classes or tracks designed to accelerate students often are demographically unrepresentative of their schools. That fact has invited severe criticism. Second, in accordance with political opposition, schools in communities serving large numbers of black and Hispanic students tend to shun tracking. Accelerated classes are less likely to exist for students of color. Third,

much of the research on tracking has found that students in high tracks benefit academically from separate, accelerated coursework. Researchers believe that high-track students receive a boost from exposure to academically-oriented peers, teachers trained in acceleration, and a challenging curriculum.

These three phenomena combine to limit opportunity for black and Hispanic youngsters. If tracking and accelerated coursework in eighth grade represents the beginning of a pipeline for promising young stars in mathematics or literature, that opportunity is more open to white and Asian students in suburban schools than to disadvantaged youngsters in schools serving students of color.

AP courses represent the end of the pipeline for academically gifted students. If we are serious about expanding opportunity, and serious about increasing the numbers of students of color who not only take AP courses but also score extraordinarily well on AP tests, policymakers need to take another look at strategies for nurturing academic talent in middle schools. Long condemned by political opponents, tracking has been overlooked as a potential tool for promoting equity.

Part

III

PRINCIPALS AS INSTRUCTIONAL LEADERS: AN INTERNATIONAL PERSPECTIVE



HISTORICALLY, ONE OF THE SCHOOL PRINCIPAL'S KEY instructional duties has been observing teachers as they teach and then providing feedback on the lesson. Ellwood Cubberly offers a vignette in the third edition of *Public School Administration* (1929) in which a young teacher's math lesson is critiqued as follows: "entirely wrong procedure for type of problems used," and "no attempt at problem solving instruction." The teacher is praised for "managerial ability" but the principal's notes reveal that his debriefing of the lesson included telling the novice instructor, "Being a new teacher to our school, she evidently did not know how we taught Arithmetic."²⁴

Feedback from principals was taken seriously in the 1920s, but, in reality, observations were infrequent, and teachers made all of the important day-to-day instructional decisions. Contemporary scholars believe times have changed, and principals now have a greater say in how instruction is conducted. The change may be partially due to accountability systems that require schools to demonstrate growth on annual state assessments. According to Dan Domenech, executive director of the American Association of School Administrators, increasing the stakes associated with state-measured school performance has meant principals assuming a greater role in shaping classroom instruction.²⁵

The call for principals to act as "instructional leaders" predates No Child Left Behind (NCLB) by at least two decades; it first appeared as a prominent policy recommendation in the "effective schools" research.²⁶ Has the value of this recommendation been confirmed by sound evaluations? It has not. A 2003 meta-analysis of studies on the relationship of principals' leadership activities to student achievement calculated an average correlation coefficient of .25.²⁷ Of the factors differentiating successful and unsuccessful principals, involvement in the design of curriculum, instruction, and assessment practices only ranked 20th out of 21 characteristics ($r = .16$).

Economists have applied the statistical techniques of their field to the question as well. A study analyzing data from *High School and Beyond*, a large national database collected in the 1980s, found selection of teachers, along with setting academically ambitious goals, as means by which principals positively influence achievement.²⁸ A recent study examining data from Texas schools found that highly effective principals produce, for the average student in their schools, approximately two to seven months of additional learning in mathematics compared to students in schools with an average principal.²⁹ The effect appeared to be driven by how principals mold a faculty through the selection of teachers.

As suggested by the econometric literature, a strain of the research on instructional leadership is concerned with specifying the exact behaviors that can produce a positive impact on student learning. The challenge is compounded by the fact that instructional leadership can be defined many ways. Grissom, Loeb, and Master followed 100 principals of urban schools for three years.³⁰ They discovered that principals spent about one-eighth of their time on instructional activities (12.6 percent). Overall, they found no relationship between instructional leadership and student achievement. Individual behaviors mattered, however. Two ways that principals interact with teachers—evaluating and coaching—were found to be positively associated with achievement gains. Perhaps the most intriguing finding was that informal classroom walkthroughs were negatively associated with achievement gains, especially in high schools. In the study, walkthroughs were the most common tool for principals to supervise instruction (consuming 5.4 percent of time), followed by formal evaluations (2.4 percent).³¹

The effort to identify positive instructional leadership behaviors has garnered the attention of international researchers. A 2003 review of international research on the topic located 125 empirical studies, concluding that principals affect student achievement indirectly, through their dealings with teachers and by shaping school culture.³²

The current study investigates principal leadership from an international perspective. It is presented in three parts. The first section sets the stage by looking at the most recent data on instructional leadership. How does the U.S. compare to other nations? The second section digs into the archives of international data to explore how principals' leadership activities changed during two intervals—2003 to 2007 and 1995 to 1999—and how those changes were related to student achievement. The final section discusses the implications of the findings for future research.

Instructional Leadership on TIMSS

The Trends in International Math and Science Study (TIMSS) is an international assessment of students in grades four and eight. In 2011, 63 nations took part. The test was originally given in 1995 and has been repeated every four years since then. Along with the assessment, surveys of teachers and school principals are conducted to collect contextual information on schools. In 2011, principals were asked how much time they spend on various leadership activities. Let's look at how they responded to questions involving instructional leadership.³³

Table 3-1 shows the responses of principals from several countries, along with international averages (in the bottom row). The data are for schools containing a fourth grade, which typically means a school serving elementary grade students (kindergarten

The current study investigates principal leadership from an international perspective.

Principals in Finland and Japan appear particularly “hands off” when it comes to instructional leadership.

Fourth grade, principals’ time spent on leadership activities, 2011
(Percent of students whose principals spend “a lot of time”)

Table 3-1

Country	Promoting the school’s educational vision or goals	Developing the school’s curricular and educational goals	Monitoring teachers’ implementation of the school’s educational goals in their teaching	Advising teachers who have questions or problems with their teaching
England	61	62	56	17
Finland	36	34	18	16
Iran	77	88	79	61
Hong Kong	52	68	58	16
Japan	40	28	47	27
Korea	88	82	81	72
Norway	27	19	17	16
Qatar	70	81	81	69
Russian Federation	80	81	81	34
Saudi Arabia	48	61	77	52
Sweden	52	40	17	27
Thailand	68	74	76	74
United States	72	68	71	42
International Average	59	60	53	39

Source: Selected countries and activities from Exhibit 6.5, TIMSS 2011 International Results in Mathematics, p. 258.

through fifth or sixth grade). The leadership activities are arrayed left to right from the least direct (goal setting) to the most direct (monitoring and advising) in terms of supervising teachers. The data report the percentage of students attending schools whose principal spends “a lot of time” on each activity.

Several interesting patterns are apparent. More than 50 percent of students internationally have a principal who devotes a lot of time to developing and promoting their schools’ educational goals and monitoring teachers’ implementation of those goals in teaching. Less time is given to giving advice to teachers about questions or problems with teaching (39 percent). The U.S. comes in well above the world averages on three of the four activities. The proportion of

American students with principals spending a lot of time offering instructional advice to teachers is about average (42 percent).

Four countries that are well known for consistently scoring at the top of international assessments—Finland, Hong Kong, Japan, and Korea—vary on these dimensions of principal leadership. (In the discussion below, subnational entities such as Hong Kong are referred to as “countries” or “nations.”) Indeed, principals in Finland and Japan appear particularly “hands off” when it comes to instructional leadership, registering well below the international norms. Finland’s principals are known to defer to teachers on instructional decisions.³⁴ Only 18 percent of students in Finland attend schools in which principals monitor teachers’ implementation of school

Eighth grade, principals' time spent on leadership activities, 2011*(Percent of students whose principals spend "a lot of time")*

Table

3-2

Country	Promoting the school's educational vision or goals	Developing the school's curricular and educational goals	Monitoring teachers' implementation of the school's educational goals in their teaching	Advising teacher who have questions or problems with their teaching
England	64	67	55	20
Finland	34	25	22	17
Hong Kong	41	47	48	21
Iran	84	91	81	48
Japan	31	21	32	18
Korea	88	78	77	61
Norway	29	20	20	20
Qatar	72	78	79	66
Russian Federation	80	82	68	27
Saudi Arabia	53	59	81	56
Sweden	45	44	20	21
Thailand	72	78	69	61
United States	65	64	64	38
International average	64	62	62	44

Source: Selected countries and activities from Exhibit 6.6, TIMSS 2011 International Results in Mathematics, p. 260.

goals in instruction, compared to 53 percent internationally. In Japan, lesson study is a popular activity, in which teams of teachers meet to plan and review instruction. Only 27 percent of Japanese principals advise teachers who have questions or problems with their teaching, compared to 39 percent internationally.³⁵

Principals in Korea, on the other hand, appear to be active instructional leaders. The same is true for Thailand, where 74 percent of students have principals who spend a lot of time giving instructional guidance to teachers. Asian countries appear heterogeneous in approaches to instructional leadership by school principals. Countries in the Middle East—Iran, Qatar, and Saudi Arabia are included here—look similar to each other and are more likely to have principals who monitor instruction and offer

advice to teachers on teaching. But there are still some differences within the region. Saudi principals are less likely to promote or to set educational goals than their colleagues in Qatar or Iran.

Table 3-2 displays data for principals of schools with an eighth grade. In many countries, the eighth grade is housed in lower secondary schools (e.g., middle schools in the U.S.) with a departmentalized structure. On all four leadership activities, the international averages are greater than reported at fourth grade. Principals in schools with an eighth grade tend to engage more actively with instruction than principals of schools with primary grades. The U.S. is an interesting exception to that pattern, with data at the eighth grade that are statistically indistinguishable from the fourth grade figures.

Is Instructional Leadership Related to Student Achievement?

Does it matter if principals exercise instructional leadership? Unfortunately, the 2011 TIMSS questions on the topic had never been used before; however, two other time periods in the TIMSS archives did pose the same battery of questions on instructional leadership: 2003–2007 and 1995–1999. The current study analyzes data using a simple difference in difference approach to examine whether changes in principal behaviors are correlated with changes in TIMSS math scores.

Modeling data drawn from multiple cross sectional samples gives the analysis a quasi-longitudinal perspective on change, diminishing potential bias from unobserved variables (provided they remain constant). The widespread belief that cultural values influence performance on international tests illustrates the value of the approach. National culture doesn't change much in four years. Whatever influence culture has on a country's test score will be present at both point A and point B and will bias cross-sectional analyses at either point in time, but that influence is subtracted out when national change in test score is calculated from A to B.

Aggregating data to the national level can also dampen selection effects that might bias findings from smaller observational units. Assume, for example, that in a particular high scoring TIMSS country, many local authorities go out of their way to hire principals who demonstrate strong instructional leadership behaviors. A cross-sectional analysis might conclude that leadership and achievement are highly correlated. Modeling the data longitudinally with two cross-sections diminishes that selection effect (as long as it remains constant). As Jan-Eric

Gustaffson explains, "There should be no mechanisms generating selection bias at the country level, and the fact that change over fixed countries is analyzed turns many of those factors that vary over countries into constants so that they cannot correlate with the independent variables under study."³⁶ The technique has been used to investigate the effects of student age, class size, school choice, instructional time, and student engagement.

Table 3-3 presents correlation coefficients on the relationship of TIMSS math score with the amount of time principals spend on instructional leadership. For an explanation of correlation coefficients, please see the highlighted text box in part two of this report. The first two columns in Table 3-3—2003 and 2007—report the cross-sectional relationships. The final column reports the correlation of change in TIMSS score to change in instructional leadership for 2003 to 2007.

Principals were asked to estimate the percentage of time they spend on instructional leadership. The underlying data are not shown, but U.S. principals of schools with a fourth grade reported spending 26 percent of their time on instructional leadership in 2003 and 2007; for schools with an eighth grade, the figure was 24 percent, and it also remained unchanged from 2003 to 2007. The 2007 international averages on the question were 21 percent for fourth grade and 20 percent for eighth grade, both statistically significantly lower than the American figures. It is interesting that the time estimate for U.S. principals in TIMSS is much larger than the estimate in the Grissom, Loeb, and Master study described above, which was about 12.6 percent. Principals may be inclined to overestimate the amount of time they spend on instructional leadership when asked on

**Relationship of instructional leadership with TIMSS math score
(schools with fourth and eighth grades, 2003–2007)**
(Correlation coefficients)

Table
3-3

Grade	2003	2007	Change 2003–2007
Fourth grade N=21	0.39	0.42	0.30
Eighth grade N=33	0.30	0.41*	0.01

* $p < 0.05$

Note: In 2003 and 2007, principals reported percentage of time spent on activities, including instructional leadership. Questionnaire gave “developing curriculum and pedagogy” as an example of instructional leadership.

Source: Exhibit 8.5, 2007 TIMSS International Mathematics Report, pp. 334–337.

a questionnaire. Grissom, Loeb, and Master followed principals around on campus to record their daily activities. That study also focused solely on principals in urban schools, so it is possible that a national sample of principals, as in TIMSS, would reveal a different pattern of behaviors.

Note that in fourth and eighth grades, the cross-sectional correlation coefficients are stronger in both 2003 and 2007 than in the longitudinal correlations for 2003–2007. The relationships at fourth grade fail to reach statistical significance, although they come close: 0.39 ($p=0.08$) in 2003 and 0.42 ($p=0.06$) in 2007. The correlation of change in leadership and test scores (0.30) is not close ($p=0.19$). The 2007 correlation in eighth grade (0.41) is statistically significant, but the relationship completely washes out when modeled over time (0.01). The bottom line to Table 3-3 is that there is no compelling evidence from the 2003–2007 data that changes in instructional leadership of school principals is related to changes in TIMSS math scores.

Table 3-4 displays data collected in 1995 and 1999 from principals of schools with eighth graders. Fourth grade was not part of TIMSS in 1999. The principal

questionnaire was the same in 1995 and 1999. This time, principals were asked how many hours per month they devote to several activities, including instructional leadership. In 1995, U.S. principals estimated that they spent about 36 hours monthly on instructional leadership, a bit more than (but not statistically significantly different from) the international average of 32 hours. National changes in instructional leadership from 1995 to 1999 were unrelated to changes in TIMSS scores (0.11).

An activity that is correlated with achievement gains is communication. Note how negative correlations in the cross-sectional data (-0.27 in 1995 and -0.32 in 1999) appear positively related to achievement gains (0.47). This reversal is almost surely due to selection effects. In the questionnaire item, communicating is defined to include talking with parents, as well as counseling and disciplining students, activities that may dominate a principal's time on unruly campuses. It also includes responding to requests from local, regional, or national authorities, a burden that would certainly increase at low achieving schools facing sanctions under an accountability system. The relationship is interesting but

U.S. principals estimated that they spent about 36 hours monthly on instructional leadership.

Principals are most likely to influence instruction by developing and setting educational goals for their schools.

Relationship of principals' activities with TIMSS math score (schools with eighth grade, 1995–1999)
(Correlation coefficients)

Table
3-4

Activity	1995	1999	Change 1995–1999
Instructional leadership	0.15	0.33	0.11
Communication	-0.27	-0.32	0.47*
Administrative	0.35	0.26	0.18
Teaching	-0.36	-0.28	-0.38

*p<0.05

Note: In 1995 and 1999, principals were asked to report number of hours spent on activities. Instructional leadership activities were defined as: discussing educational objectives with teachers, initiating curriculum revision and/or planning, training teachers, and professional development activities.

Source: Exhibit 7.3, TIMSS 1999 Eighth Grade Mathematics, pp. 236.

difficult to interpret (and beyond the scope of the present study). Future research would benefit from investigating these activities separately.

Summary and Conclusion

This section of the Brown Center Report offered an international perspective on the role of the school principal as an instructional leader. Principals' responses to TIMSS surveys were examined at the national level, along with test scores in mathematics. The analysis calculated correlation coefficients to estimate whether test scores and the time principals spend on instructional leadership are associated. The analysis does not allow for causal conclusions, but it is useful for generating ideas and hypotheses for further research.

Instructional leadership varies from country to country, although a few commonalities were detected. Principals are most likely to influence instruction by developing and setting educational goals for their schools. They also monitor teachers' implementation of goals, but are less likely to give advice to teachers who have questions or problems with their teaching. In some countries, giving pedagogical guidance may be considered an infringement

on teacher autonomy, or perhaps teachers simply consult with peers or other sources for advice. Principals in three consistently high achieving countries—Finland, Hong Kong, and Japan—are especially reluctant to give advice. Principals in Korea, on the other hand, another perennially high achieving country, are more activist in this regard.

U.S. principals of schools with a fourth grade (typically an elementary school) are about average in terms of giving instructional advice, but register above the international average on activities related to school goals. American principals of schools with an eighth grade (typically a middle school) appear about average on all four surveyed activities, with one interesting side note. Whereas international averages suggest principals of schools housing an eighth grade are more likely to spend time on instructional leadership than principals of schools with a fourth grade, the U.S. data indicates the opposite for American schools. Elementary school principals appear more involved with instruction than their middle school counterparts.

Archival TIMSS data were analyzed from two periods: 2003–2007 and 1995–1999. No clear evidence emerged

that principals' instructional leadership is correlated with achievement. A statistically significant positive relationship was discovered for eighth grade achievement in 2007, but the relationship faded away in the longitudinal analysis of data from 2003–2007. Culture looms as an omitted variable in the cross-sectional analysis, suggesting that the cross-sectional correlation may be spurious. In an influential study of classroom instruction in three countries, Stigler and Hebert argued that teachers follow “cultural scripts” in their classroom instruction, teaching essentially in the same manner that they themselves were taught.³⁷ That same notion of cultural scripts probably extends to the way principals view principal-teacher relations, making the longitudinal statistic, with its ability to control for cultural influences, a more reliable measure.

This study offers three takeaways for future research. First, the term “instructional leadership” is problematic. Without the description of specific behaviors, it can mean different things to different people. The definition may also change over time. It's significant that the Grissom, Loeb, and Master study was published in 2013 and pinpointed leadership practices that are both positively and negatively related to student achievement. As a scholarly enterprise, measuring precise behaviors in rigorously designed studies is in its infancy.

Second, the context of instructional leadership is important. The current study examined data from principals of elementary and middle schools. Other conditions that could affect instructional leadership at an individual school include: the length of time a principal has worked with staff, the history of principal-teacher relations at the school, the degree of change that principals seek in teachers' current instruction, the involvement of parents in curriculum and instruction at the school, the instructional resources available (e.g., funds for new materials and professional development), and the priority given to instructional leadership by the principals' district, regional, or state authorities. Some of these contextual factors are empirical and can be quantified with administrative data, but many are not and will require more sophisticated forms of data collection.

Finally, the current status of principal leadership research is much like the effective teacher research. As pointed out in the brief review of literature above, great strides have been made in the ability to identify successful principals, those who make a difference in boosting student achievement at the schools they lead. Research has been less successful, however, in describing why particular principals are successful leaders—why they succeed while others, regrettably, struggle with leading their schools.

No clear evidence emerged that principals' instructional leadership is correlated with achievement.

NOTES

- 1 The survey asked state education agencies if they had adopted the CCSS and 1) provided, guided, or funded professional development on CCSS, 2) provided curriculum or instructional materials for the CCSS; and 3) worked with a consortium to develop assessments aligned with CCSS.
- 2 Michael Lipsky (1980). *Street Level Bureaucracy: Dilemmas of the Individual in Public Services* (Russell Sage Foundation).
- 3 Catherine Gewertz, "Scale Tips Toward Nonfiction Under Common Core," *Education Week*. (November 13, 2012).
- 4 On the 4th grade NAEP assessment, data analysis and geometry are targeted to comprise 25% of items, and in 2015, they actually represented 28% of items. National Center for Education Statistics, National Assessment of Educational Progress (NAEP), Distribution of Mathematics Questions: 2015, <https://nces.ed.gov/nationsreportcard/mathematics/distributequest.aspx>
- 5 "What We're Watching: David Coleman on the Common Core Standards," *Education Next*. (July 11, 2011). Available at: <http://educationnext.org/what-are-watching-davidcoleman-on-the-common-core-standards/>
- 6 Jason Zimba, "The Common Core and the Potential for Mathematicians to Improve the Teaching of School Mathematics," *Notices of the AMS*, vol. 63, no. 2 (February, 2016.), pp. 154–158.
- 7 Sharon Noguchi, "California abandons algebra requirement for eighth-graders," *San Jose Mercury News*. (February 2, 2013).
- 8 State implementation ratings are fixed in IMP11. Ratings are dynamic in IMP13, changed for each NEAP interval to reflect policy changes. Based on policy changes from 2013 to 2015, three states that rescinded CCSS were reclassified as nonadopters (Indiana, Oklahoma, South Carolina) and one state, Georgia, that revised its standards was reclassified from strong to medium. Minnesota is classified as a nonadopter in math and a medium (IMP11) and strong (IMP13) implementer in reading. Alaska, a non-adopter, has no data on eighth grade course enrollments and is not included for those calculations.
- 9 John Fensterwald, "Districts must ensure equity in rules for accelerating students in math," EdSource. (December 14, 2015).
- 10 Free and reduced lunch statistics have historically been used as measures of socioeconomic status, but changes in that program's rules for eligibility cast doubt on their continued reliability.
- 11 "The real test of Common Core is about to begin," *The Sacramento Bee*. (January 17, 2015).
- 12 Tom Loveless (2001). "A Tale of Two Math Reforms: The Politics of the New Math and the NCTM Standards," in *The Great Curriculum Debate: How Should We Teach Reading and Math?* (Brookings Institution Press), pp. 184–209.
- 13 Tracking refers to how students are organized between classes. It is different from ability grouping, which involves grouping students into small groups within classes.
- 14 Colleges have been criticized for being overly zealous in recruiting talent. The University of Washington and Louisiana State University have promised scholarships in exchange for signed letters of commitment from eighth grade football players. In 2010, David Sills, a 13 year old quarterback in Elkton, MD committed to attend USC when he was in seventh grade. Adam Himmelsbach and Pete Thamel, "Middle School Is Basketball's Fiercest Recruiting Battleground," *The New York Times*. (June 25, 2012).
- 15 Samuel Bowles and Herbert Gintis, *Schooling in Capitalist America* (Basic Books, 1976); Jeannie Oakes, *Keeping Track* (Yale University Press, 1985); Diane Ravitch and Chester E. Finn, Jr., *What Do Our 17-Year-Olds Know?* (Harper & Row, 1987).
- 16 National trends are presented in Loveless, T. (2013). "The Resurgence of Ability Grouping and Persistence of Tracking," *The 2013 Brown Center Report* (Brookings Institution Press), pp. 12–20. For a study focusing on tracking and disadvantaged students, see Loveless, T. (in press). "Tracking in Middle School: A Surprising Ally in Pursuit of Equity?" in *Education for Upward Mobility* (Petrilli, Michael J., ed.). Lanham (MD): Rowman & Littlefield, 2016.
- 17 Esther Duflo, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review* 101 (August 2011): pp. 1739–1774.
- 18 Takako Nomi (2012). "The Unintended Consequences of an Algebra-for-All-Policy on High-Skill Students: Effects on Instructional Organization and Students' Academic Outcomes," *Educational Evaluation and Policy Analysis*
- 19 David Card and Laura Giuliano (2014). "Does Gifted Education Work? For Which Students?" NBER Working Paper Series, Working Paper 20453 (September, 2014).
- 20 At the time of the study, the 2013 graduating class was the most recent year for which a complete panel of state level AP data could be matched to NAEP data collected four years earlier.
- 21 All NAEP data were collected from the NAEP Data Explorer. Tracking based on responses from question #C072801 on the school questionnaire, "student assigned to math by ability."
- 22 College Board (2014). Tenth Annual AP Report to the Nation. AP race and ethnicity data retrieved from individual state reports in *State Supplements*, available at: <http://apreport.collegeboard.org/>. Child poverty rates collected from Kids Count Data Center (kidscount.org).
- 23 A regression model using 2013 poverty data produced a slightly smaller t-statistic for poverty and slightly larger t-statistic for tracking.
- 24 Robert J. Marzano, Tony Frontier, and David Livingston (2011). *Effective Supervision: Supporting the Art and Science of Teaching*. (Association for Supervision and Curriculum Development)
- 25 Ed Finkel (2012). "Principals as Instructional Leaders: But Can They Do It All? And at What Cost?" *District Administration*, June 2012.
- 26 Hallinger, Phillip and Murphy, Jerome (1985). Assessing the Instructional Leadership Behavior of Principals," *Elementary School Journal*, vol. 86, 2, pp. 217–248.
- 27 Tim Waters, Robert J. Marzano, and Brian McNulty (2003). "Balanced Leadership: What 30 Years of Research Tells Us About the Effect of Leadership on Student Achievement," Mid-Continent Regional Educational Lab.
- 28 Dominic J. Brewer (1993). "Principals and Student Outcomes: Evidence from U.S. High Schools." *Economics of Education Review*, vol. 12, no. 4, pp. 281–292.
- 29 "School Leaders Matter," Gregory F Branch, Eric A. Hanushek, and Steven G. Rivkin, *Education Next*, vol. 13, no. 1 (Winter 2013).
- 30 Grissom, J. A., Loeb, S., & Master, B. (2013). Effective instructional time use for school leaders: Longitudinal evidence from observations of principals. *Educational Researcher*, 42(8), pp. 433–444.
- 31 See discussion of this study and classroom walkthroughs on Daniel Willingham's blog: <http://www.danielwillingham.com/daniel-willingham-science-and-education-blog/the-classroom-walkthrough-and-student-achievement>
- 32 Bob Witziers, Roel J. Bosker, and Meta L. Kruger (2003). "Educational Leadership and Student Achievement: The Elusive Search for an Association," *Educational Administration Quarterly*, 39, 3, pp. 398–425.
- 33 Principals were also asked about administrative duties, school safety, and student discipline (not shown in the tables).
- 34 Center on International Education Benchmarking "Finland: Teacher and Principal Quality," available at: <http://www.ncee.org/programs-affiliates/center-on-international-education-benchmarking/top-performing-countries/finland-overview/finland-teacher-and-principal-quality/>
- 35 Elizabeth Green (2014), *Building a Better Teacher*. (New York: W. W. Norton & Company).
- 36 Jan-Eric Gustafsson (2007). "Understanding Causal Influences," in *Lessons Learned: What International Assessments Tell Us About Math Achievement*, Brookings Institution Press, pp. 37–65. Quote is from page 60.
- 37 James Stigler and James Hebert (1999), *The Teaching Gap* (The Free Press).

THE BROOKINGS INSTITUTION

STROBE TALBOTT
President

DARRELL WEST
Vice President and Director
Governance Studies Program

BROWN CENTER STAFF

DIANA QUINTERO CASTELLANOS
Research Assistant

STEPHANIE CELLINI
Nonresident Senior Fellow

WILLIAM A. GALSTON
Interim Director and the Ezra K. Zilkha
Chair in Governance Studies

MICHAEL HANSEN
Senior Fellow and Deputy Director

HELEN LADD
Nonresident Senior Fellow

TOM LOVELESS
Nonresident Senior Fellow

ADELA SOLIZ
Fellow

BETH STONE
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