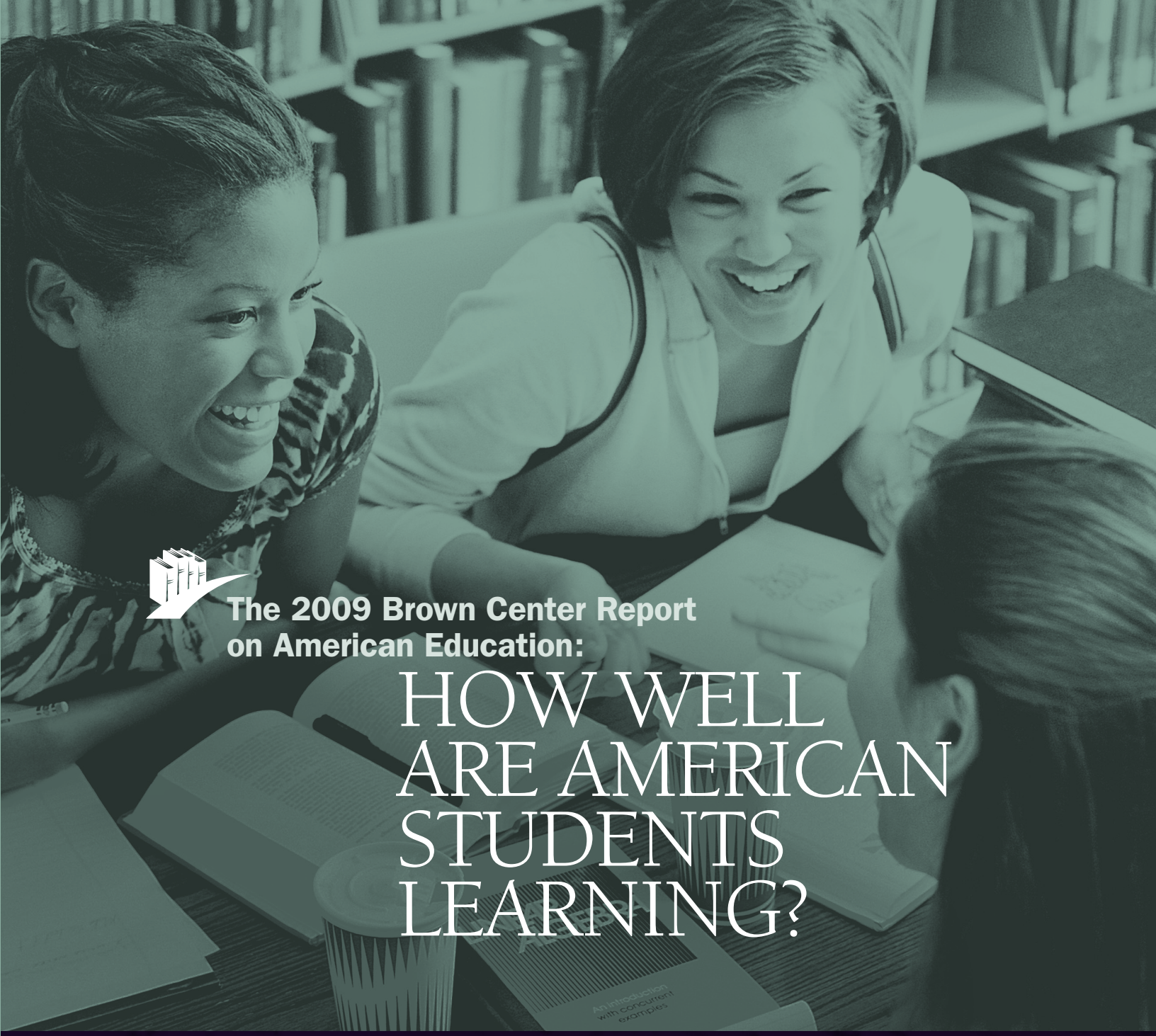


The 2009 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on NAEP trends, the persistence of school test scores, and conversion charter schools

BROOKINGS

ABOUT THE BROOKINGS INSTITUTION

The Brookings Institution is a private nonprofit organization devoted to research, education, and publication on important issues of domestic and foreign policy. Its principal purpose is to bring knowledge to bear on current and emerging policy problems. The Institution maintains a position of neutrality on issues of public policy. Interpretations or conclusions in Brookings publications should be understood to be solely those of the authors.

ABOUT THE BROWN CENTER ON EDUCATION POLICY

Established in 1992, the Brown Center on Education Policy conducts research on topics in American education, with a special focus on efforts to improve academic achievement in elementary and secondary schools. For more information, see our website, www.brookings.edu/brown.aspx.

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

The 2009 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With sections on NAEP trends, the persistence of school test scores, and conversion charter schools

January 2010
Volume II, Number 4

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

TABLE OF CONTENTS

Research assistance by:

MICHELLE CROFT
KATHARYN FIELD-MATEER
Brown Center on Education Policy

THE 2009 BROWN CENTER REPORT ON AMERICAN EDUCATION

This year's Brown Center Report contains studies taking a long view. Part I examines national test data going back to 1971 from the National Assessment of Educational Progress (NAEP). The study in Part II compares the 1989 test scores of more than 1,000 schools to the same schools' scores in 2009. Part III compares the test scores of conversion charter schools from 1986, when they operated as traditional public schools, to those from 2008, when they operated as charter schools. The studies tackle perennial questions that, as often happens in education, manifest themselves as controversial topics on the contemporary scene: how to interpret trends in test scores, the distribution of achievement, school turnarounds, and charter schools.

Part I rejects the conventional reaction to the 2009 NAEP scores. Scores in fourth-grade math were unchanged from 2007 to 2009. Eighth-grade scores were up a little. Press articles featured expressions of disappointment and concern, primarily from protagonists who used the flat scores to support policy arguments. Part I places the 2009 scores in the context of the 19-year history of the main NAEP, and after comparing the latest scores with results from other equally trustworthy tests of U.S. math achievement, concludes that the hand-wringing is unwarranted.

So when is a purported NAEP trend really a trend? Part I continues by examining achievement gaps, not between two racial, ethnic, or socioeconomic groups, but between the nation's highest- and lowest-achieving students. It focuses on the distribution of academic achievement instead of the direction of average achievement. The study is a follow-up

to a 2009 Fordham Institute paper documenting that the gap between high- and low-achieving students has been shrinking in recent years. The data in Part I show that the trend, which began sometime around 1998 or 1999, is historically unprecedented and extends across subjects (reading and math), grades (fourth and eighth), and tests (long-term trend and main NAEP). It is also more pronounced in public schools than in private schools. The two analyses in Part I highlight the contrast between a trend indicated by data collected from several independent sources over an extended period of time and speculative assertions arising from "instant analysis" of a single set of test scores.

Part II asks a simple question: do schools ever change? The sample consists of 1,156 schools in California that offered an eighth grade in 1989 and 2009. Test scores from 1989 are compared to scores from 2009. The scores are remarkably stable. Of schools in the bottom quartile in 1989—the state's lowest performers—nearly two-thirds (63.4 percent) scored in the bottom quartile again in 2009. The odds of a bottom quartile school's rising to the top quartile were about one in seventy (1.4 percent). The reverse was true as well, with similar percentages of top quartile schools staying among the top performers (63.0 percent) or falling to the bottom quartile (2.4 percent). Changes in a school's socioeconomic status had only a marginal statistical relationship with test score changes.

The persistence of test scores has major implications for today's push to turn around failing schools. It can be done, but the odds are daunting. California certainly cannot be accused of inactivity in education reform from 1989 to 2009. Few states tried as many diverse, ambitious reforms that targeted every aspect of the school system—finance, governance,

curriculum, instruction, and assessment. Not only have these efforts failed to elevate California from its low national ranking on key performance measures, but they have also had little effect on the relative ranking of schools within the state.

The study suggests that people who say we know how to make failing schools into successful ones but merely lack the will to do so are selling snake oil. In fact, successful turnaround stories are marked by idiosyncratic circumstances. The science of turnarounds is weak and devoid of practical, effective strategies for educators to employ. Examples of large-scale, system-wide turnarounds are nonexistent. A lot of work needs to be done before the odds of turning around failing schools begin to tip in a favorable direction.

Part III looks at charter schools. Conversion charters are favored by the Obama administration as a restructuring strategy. Most charter schools are start-ups, begun from scratch by their founders. Conversion charters are schools that are traditional public schools and convert to charter school status. They typically continue to rely on their home districts for several functions (e.g., maintenance of buildings, managing pension obligations, transportation services) but are freed from regulations pertaining to curriculum and instruction. The idea is that schools can be more productive if they are allowed to tailor core educational operations to the needs of their students.

California has the largest number of conversions, and the study was able to collect data on two cohorts: 49 schools from 2004 and 60 schools from 2008. For both cohorts, test score data were also available from 1986, allowing a comparison of scores before and after the schools

converted. The analysis is exploratory and mainly descriptive. No causal conclusions can be derived from the data.

What do we know about conversions? Test scores look similar before and after conversion. The 2004 cohort evidences a 2 to 3 percentile point advantage as charters, but the 2008 cohort's scores declined slightly, less than 2 points, from 1986 to 2008. On several key characteristics, conversions look more like traditional public schools than start-up charters. Compared with start-ups, conversions are more concentrated in urban areas, have larger student enrollments, and serve greater numbers of Hispanic and black students. Teachers at conversions are more experienced and more likely to hold teaching certificates, particularly in bilingual education. It is clear that future evaluations of charter schools must differentiate between start-ups and conversions because of the significant institutional differences between the two types of charters.

To sum up, the studies in this year's Brown Center Report focus on long-term changes. Part I analyzes NAEP data. Parts II and III examine California test scores from the 1980s and compare them to scores from recent years. Because of its long history of testing, California is currently one of the few states able to provide assessment data for such long-term comparisons. That will change as other states continue to test students annually. Creating rich archives of student performance data bodes well for school reform. Improving schools requires patience and persistence, what education professors Richard Elmore and Milbrey McLaughlin[1] call "steady work." It also requires good information to verify whether reforms have paid off, or, like many efforts in education, produced hopeful signs that soon vanish. The future looks bright if analysts' capacity to peer into the past continues to improve.

Part I WHAT DO THE 2009 NAEP SCORES TELL US?

THE LATEST SCORES FROM THE NATIONAL ASSESSMENT OF Educational Progress (NAEP) were released in October 2009. Commonly called "the nation's report card," the NAEP assesses the reading and math achievement of fourth and eighth graders every two years on the main NAEP tests. Occasionally twelfth graders are tested. Other subjects are assessed less frequently and by no set schedule. The main NAEP is one of four assessments regularly administered to a randomly selected, representative sample of American students. The long-term trend NAEP (LTT NAEP)—a separate test with an age-based sample of students—is another, and two international tests—the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA)—round out the group.

Only math scores were released in October (reading scores are scheduled for release in Spring 2010). As usual, the scores drew a lot of press coverage. The scores showed a potential slowing in what has been a nearly two-decade upswing in NAEP math scores. Has that upward trend stopped? Is a new trend starting? This section of the Brown Center Report will take a look at the 2009 scores in the context of NAEP's history and lay out some rules for deciding when a pattern in scores really does constitute a trend. Then those rules will be applied to illustrate a trend that has been developing in NAEP scores over the past decade but not widely discussed: a narrowing of the difference between students scoring at the 90th percentile and those scoring at the 10th.

Main NAEP Math Scores, 2009

The 2009 scores are shown in Figure 1-1. Fourth graders scored 240 scale score points, unchanged from 2007. Eighth graders gained 2 points, from 281 to 283, a small but statistically significant increase. Press coverage featured comments expressing concern. U.S. Secretary of Education

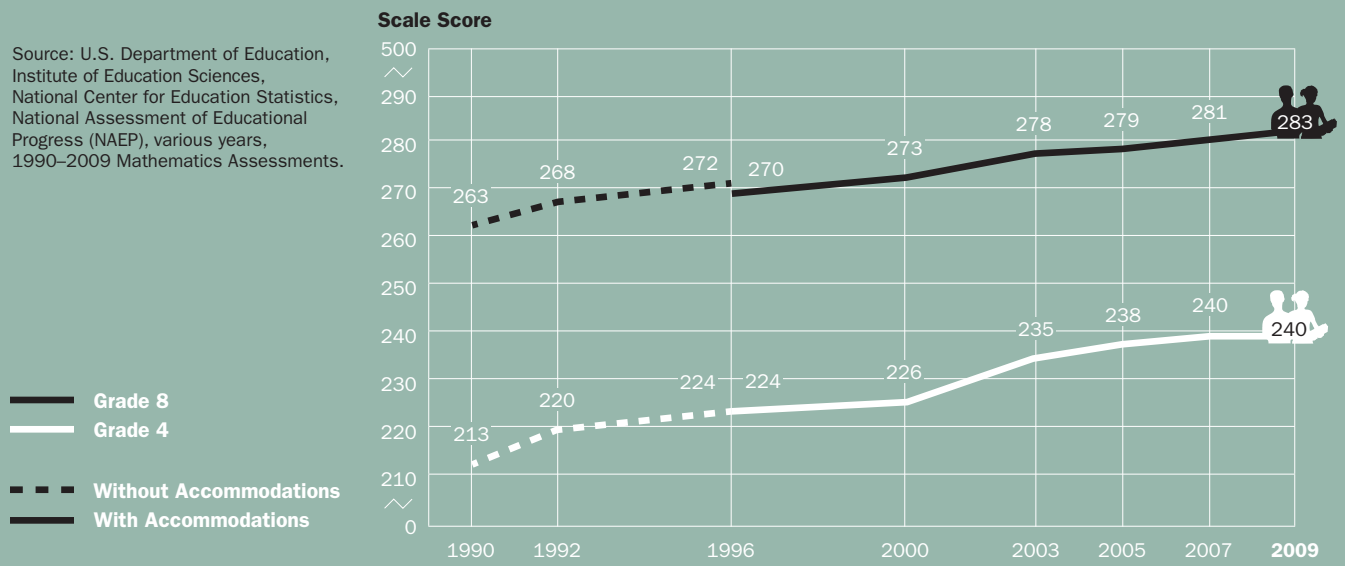
Test scores should be evaluated in the context of time and over several administrations of the test.

Arne Duncan said the scores demonstrated the need for education reforms that will "accelerate achievement." David Driscoll, chair of the National Assessment Governing Board, the policy-making body for NAEP, argued that the scores indicated elementary math teachers need better training. Mark Schneider, a political scientist at the American Institutes of Research, and Diane Ravitch, the renowned education historian, concluded that because a larger increase in NAEP scores took place in the years immediately preceding 2003 than from 2003 to 2009, the scores were bad news for No Child Left Behind. The *Wall Street Journal*'s article on the scores was headlined, "U.S. Math Scores Hit a Wall." *Education Week*'s lead sentence for its story stated that the scores would bolster support for the national standards movement.[2]

These comments should be taken with a grain of salt. Their most serious fault is reading too much into a single set of NAEP scores. Test scores should be evaluated in the context of time and over several administrations of the test. Examine the fourth-grade scores in Figure 1-1. From 1990 to 2007, fourth graders' scores rose from 213 to 240, a gain of 27 scale score points. Analysts usually consider 10 to 11 points as equal to one year's worth of learning. Using that metric, the gain represents more than two and a half grade levels of mathematics, a truly incredible—some would say unbelievable—increase over 17 years.[3] The gain from 1990 to 2009 is the same, 27 points, also an incredible increase, but over 19 years. The comments above imply that great significance can be read into the difference between an incredible gain over 17 years versus an equally incredible gain over 19 years—all because scores for the last two years in the interval were flat.

Main NAEP, 1990–2009, Mathematics

Fig 1-1



Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), various years, 1990–2009 Mathematics Assessments.

The gain at eighth grade was consistent with the average gain registered since 1990 for eighth graders, about one scale score point per year. The comments above ignored the gain at eighth grade and focused on the fourth-grade scores. Is there cause for concern? Has something bad happened in fourth-grade math?

Now, it may be that a new trend of sideways or even declining main NAEP scores is beginning in fourth-grade math. That would be an important development, but no one knows if it is really happening. No one will know until scores from 2011 and 2013 and later years are released. It may also be true, as noted in the comments cited above, that we are hitting a wall; NCLB is bad law; and we need new and different education reform. Such explanations are premised on the belief that we must have been doing something terribly wrong between 2007 and 2009 for math scores at fourth grade to show no gain.

There is another plausible explanation.

An Alternative Explanation

Tables 1-1 and 1-2 show the gains registered from 1990 to 2007 on three tests: the main NAEP, the long-term trend NAEP (LTT NAEP), and TIMSS (TIMSS was first given in 1995). The three tests employ similar sampling strategies and possess technical qualities that make all of them, not just the main NAEP, reliable indicators of U.S. national achievement.[4] The crucial difference is in the skills and knowledge they assess. The mathematics on the main NAEP is not the same as that on the LTT NAEP, and both NAEPs assess different content than TIMSS. PISA also differs in content from these three tests, but because it tests an age-based sample of older students (15-year-olds) and has a much shorter history, having started in 2000, it is not included in the remainder of the discussion.

4th Grade/Age 9 Math Achievement

Table 1-1

Test	Years/Sample	Point Gain	Gain in SD Units
Main NAEP	1990–2009 (4th grade)	27	0.84
LTT NAEP	1990–2007 (age 9)	13	0.39
TIMSS	1995–2007 (4th grade)	11	0.11

Note: SD gains computed from SD at baseline: 1990 Main NAEP (32 points); 1990 LTT NAEP (33 points); TIMSS 1995 (100 points).

Source: Author's calculations from data provided by NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/) and Mullis, et al., (2008), TIMSS 2007 International Mathematics Report (http://timssandpirls.bc.edu)

8th Grade/Age 13 Math Achievement

Table 1-2

Test	Years/Sample	Point Gain	Gain in SD Units
Main NAEP	1990–2009 (8th grade)	20	0.56
LTT NAEP	1990–2007 (age 13)	11	0.35
TIMSS	1995–2007 (8th grade)	16	0.16

Note: SD gains computed from SD at baseline: 1990 Main NAEP (36 points); 1990 LTT NAEP (31 points); TIMSS 1995 (100 points).

Source: Author's calculations from data provided by NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/) and Mullis, et al., (2008), TIMSS 2007 International Mathematics Report (http://timssandpirls.bc.edu)

Table 1-1 reports the gains for fourth grade (9-year-olds on the LTT NAEP) and Table 1-2 for eighth grade (13-year-olds on the LTT NAEP). The age groups assessed on the LTT are not a perfect match with the grades on the main NAEP and TIMSS, but they are close enough to constitute a similar sample of students. Because the scores are reported on different scales, all gains have been converted into standard deviation (SD) units to make them comparable.

Fourth-grade math on the main NAEP exhibits the largest gains. The gains on that test (0.84) are more than twice as large as the gains of 9-year-olds on the LTT NAEP (0.39) and more than seven times as large as the 0.11 gain on TIMSS. At the eighth grade, the main NAEP is also reporting the largest

The mathematics on the main NAEP is not the same as that on the LTT NAEP, and both NAEPs assess different content than TIMSS.

Perhaps the skyrocketing gains had to stall on this particular test, and elementary teachers did not suddenly become horrible math instructors in 2007.

gains of the three tests (0.56), followed by the LTT NAEP (0.35) and TIMSS (0.16).

To appreciate how out of step the main NAEP's fourth-grade results are in math, consider the following scenario. The LTT NAEP is next given in 2012. Let's imagine that 9-year-olds make the largest gain they have ever made in math from one administration of the test to the next, an increase of 9 points (last registered in 1999–2004). Assume also that fourth graders on the main NAEP show no gain in 2011 (the next administration of the main NAEP) and no gain again in 2013. Despite the dismal prospect of flat main NAEP scores for six consecutive years—imagine the hand-wringing!—and this rosy scenario for the LTT NAEP, the main NAEP gains since 1990 will still exceed the gains registered on the LTT NAEP.

TIMSS is next given in 2011. How large a gain must the United States make on the fourth-grade TIMSS to equal the gains indicated by the main NAEP? About 73 points, or 0.73 SD units, an extraordinary increase that would put the U.S. score at 602 on the TIMSS scale, statistically indistinguishable from the world's two top scorers in 2007—Hong Kong (607) and Singapore (599). The United States would finally realize the dream of scoring among the top nations in math. Please don't hold your breath for that event to occur.

One reasonable explanation for the flat 2009 main NAEP scores, then, is that the test is simply coming back to Earth, finally reporting progress more in line with other national math tests, and in particular, with the other NAEP test. At the rate fourth graders were progressing from 1990 to 2007, they would have been performing at the eighth-grade level on the main NAEP by 2022 and at a high school senior level—that is, ready for college mathematics—in 2053.[5] That projection lacks credibility. So perhaps the flat scores are due to the main NAEP test itself, not to something going wrong. Perhaps the skyrocketing gains had to stall on this particular test, and elementary teachers did not suddenly become horrible math instructors in 2007. If you go by the main NAEP, don't forget, they had been miracle workers the previous 17 years.

Why the Different Results?

Why is the main NAEP showing much larger gains than the LTT NAEP and TIMSS? The likely reason is different content. The main NAEP is dominated by whole number problem solving (see the 2004 Brown Center Report for an analysis). Compared with the LTT NAEP and TIMSS, the main NAEP also has more items requiring students to complete number patterns (called "algebra" by NAEP), in which students tell what number comes next in a sequence; data display, in which students read data from graphs and tables; and recognition of simple geometric figures. The LTT NAEP and TIMSS have more items involving fractions, decimals, and percentages; more items assessing whether students can compute accurately; and fewer (if any) pattern items. The main NAEP allows students to use calculators on a portion of the test. The LTT NAEP and TIMSS do not.

The topics emphasized on the main NAEP are a more prominent part of the curriculum now than in 1990; that is, they are taught more frequently in today's classrooms and featured in contemporary fourth-grade textbooks. A key reason for the creation of the main NAEP was to reflect changes in curriculum. It does just that. A good example are NAEP's pattern items, mentioned above and considered by NAEP to assess fourth-grade algebra. Students had little familiarity with these items in 1990. NAEP

periodically asks fourth-grade teachers how much emphasis they place on algebra. In 1990, 81 percent answered "little" or "none." In later years, teachers were asked a slightly different question: how much emphasis they place on algebra and functions. The percentage responding "little" or "none" plummeted to 16 percent in 2003 and 7 percent in 2009. The share of fourth-grade teachers answering that they place a heavy emphasis on algebra was 2 percent in 1990, 26 percent in 2003 (on algebra and functions), and 42 percent in 2009 (again, on algebra and functions). Including functions in the question may have boosted positive responses a bit in the later years. But the point remains that NAEP test score gains have undoubtedly been inflated by fourth graders' exposure to NAEP-like content in classrooms.

Relevance forces a trade-off. The main NAEP is better at measuring students' progress in learning new content rather than traditional skills and knowledge. The main NAEP's gains are not fake. Students have made progress in learning mathematics since 1990, but that progress may be limited to the mathematical topics that the main NAEP assesses. Some of the traditional mathematics that students need to know to be adequately prepared for algebra—that is, formal algebra as presented in an advanced mathematics course—may be neglected. With this in mind, the National Mathematics Advisory Panel recommended that the main NAEP increase the proportion of the test devoted to fractions, reduce the prevalence of pattern items, and not allow calculators on items designed to assess fluency with computation skills.[6]

So what should the average citizen believe when NAEP scores are released? First, when you hear comments like those after the latest release, ask yourself, especially in evaluating expressions of disappointment or glee, what kind of scores the commentators were expecting. If they were expecting fourth graders to be ready for college-level math in 2053—and believe any scores falling short of that trend line indicate something going wrong—then disappointment is inevitable. Second, be very skeptical of statements implying causality. NAEP data are poorly equipped to demonstrate causality.[7] Such statements are usually made by people who have a political ax to grind. NCLB is a perfect example. Every time NAEP scores blip upward, supporters of the law say it shows the law is working. With every dip in scores, NCLB's critics proclaim that the law is failing. Both conclusions are faulty. A good analyst first examines data to determine how the world works. That exercise is quite distinct from speculating about *why* the world works in the manner that it does. Instant analysis that does not consider whether a new batch of scores is just a random fluctuation or part of a larger trend—then speculates as to the cause of scores going up, down, or sideways—is merely piling guesswork on top of guesswork.

Let's set aside causality for now and discuss a strategy for tackling the empirical question. When can one have confidence that a trend in NAEP really exists? A simple rule to follow is: when the same pattern persists over time, of course, with longer periods of time better than shorter periods. In addition, the case for a trend strengthens when it also appears across NAEP tests (LTT and main), ages (elementary and middle school), and subjects (math and reading). Focusing on the release of a single set of scores—on one of the NAEP tests in a single subject and a single grade level—leaves out an enormous amount of data, enhancing the danger that fluctuations will be mistaken for trends.

Students have made progress in learning mathematics since 1990, but that progress may be limited to the mathematical topics that the main NAEP assesses.

A Recent Trend in NAEP: The Contracting Achievement Gap

The rules just laid out can be applied to explore fluctuations in the gap between 90th and 10th percentile students on NAEP (see Table 1-3 on the following page). Scores at the 90th percentile indicate how well the nation's highest achievers, the top 10 percent of students, are performing. The 10th percentile scores indicate how well the lowest-achieving 10 percent of students are performing. When the gap between the two groups expands, it means that the difference in their achievement is growing. When it contracts, it means that the difference is lessening. Note that this is a relative measure. The gap may contract or expand as the scores of both groups rise or fall together.

A previous study by this author documented significant contraction of the 90–10 gap on the main NAEP from 2000 to 2007. Both the 90th and 10th percentiles evidenced rising test scores during the period. The contraction was due to 10th percentile scores improving more than 90th percentile scores. Growth at the 10th percentile accelerated sharply while the languid progress at the 90th percentile continued. In the 1990s, the gap had remained fairly stable as both groups made similar progress. A notable exception was found in states that adopted accountability systems in the late 1990s. In those states, a contraction of the 90–10 gap occurred after accountability was implemented.[8]

The current study builds on the earlier one by examining the complete history of NAEP data instead of only scores from the 1990s and 2000s. It also includes data from the LTT NAEP, allowing for a more precise estimate of when the gap contraction began. And the study compares the phenomenon of gap contraction in public and private schools, a comparison many analysts of NAEP scores overlook when attributing fluctuations in test scores to federal or state education policies. Private schools are unaffected by most public education policies, providing something akin to a control group for detecting correlations between particular policies and test scores. When commentators use national NAEP scores to argue that public policies are succeeding or failing, they are including the test scores of private school students (about 10 percent of the national sample), who are unaffected by the policies in question.

Why is the gap between the 90th and 10th percentiles important? Mainly because education policy in recent years has attempted to boost the scores of low-performing students. High-achieving students have been ignored. Contrary to conventional wisdom, NCLB was not directly designed to reduce race gaps.[9] Only indirectly. It was designed—that is, operationally defined by the statute—to reduce the number of students scoring below the "proficient" cut point on state tests. The same was true for state accountability systems that preceded NCLB. They were all designed to raise the scores of low achievers. A large percentage of black, Hispanic, and poor students score below the proficiency threshold, so reducing the number of students below proficiency would do a lot to raise black and Hispanic scores. The provisions of the law that require progress by subgroups (i.e., blacks, Hispanics, disadvantaged students) also encourage the closing of racial/ethnic gaps. But the objectives of the law can be achieved without closing racial/ethnic gaps. Most states have more white than black or Hispanic students scoring below proficiency. Raising their scores is rewarded by accountability systems, too, and does not close race gaps. More importantly, no incentive exists for schools to boost the test scores of black,

Hispanic, and poor students once they clear the proficiency bar. Boosting the learning of mid- to high achievers who are black, Hispanic, or poor goes unrewarded.

Data and Discussion

Table 1-3 displays data from the LTT NAEP. The table reports the changes in the 90–10 gap each time the LTT NAEP has been given. The rows are the intervals between two consecutive administrations of the test. The columns are the age and subject combinations of the test—9-year-olds in math, 9-year-olds in reading, 13-year-olds in math, and 13-year-olds in reading. The final column reports the total change in points for the row. It is for summative purposes only—the row totals cannot be compared because the numbers of years in the intervals vary and in some years not all four age-subject combinations were tested. Altogether, the table comprises 40 interior cells holding every change in NAEP data since the test's inception.

The first thing that leaps out is how much the 90–10 gap has contracted in recent years. From 2004 to 2008, the gap contracted for all four age-subject combinations. For 9-year-olds in math, the gap shrank by 1 point. (This reflects the width of the gap decreasing from 87 scale score points in 2004 to 86 points in 2008.) The contraction was 7 points for 9-year-olds in reading, 2 points for 13-year-olds in math, and 4 points for 13-year-olds in reading—a total of 14 points for the 2004–2008 interval. The gap also contracted during the 1999–2004 interval, by a total of 12 scale score points. Math scores of 13-year-olds bucked the trend by expanding 1 point from 1999 to 2004.

The last time the 90–10 gap contracted by as much as recent years was in the three earliest intervals, 1971–1975, 1975–1980, and 1978–1982. The 1990s are marked by minor, offsetting changes in the gap. The biggest expansions occurred from 1988 to 1990 and from 1980 to 1984. Reading was the only subject tested during those intervals. The 90–10 gap generally contracted in math during the 1980s but expanded in reading, especially among 9-year-olds.

Table 1-4 reports the 90–10 gap changes on the main NAEP. The top row is incomplete until the 2009 scores are reported for reading. The same recent narrowing of the gap is evident, more so in 1998–2002 (eighth-grade reading), 2000–2002 (fourth-grade reading), and 2000–2003 (math in both grades) than in later intervals. The 2002–2003 scores in reading are an exception as the 90–10 gap expanded. The earliest main NAEP intervals, 1990–1992 and 1992–1994, also show expansion of the

From 2004 to 2008, the 90–10 gap contracted for all four age-subject combinations.

Gap Changes on the Long-Term Trend NAEP
90th and 10th percentile, 1971–2008

Table **1-3**

Years	9-year-olds Math	9-year-olds Reading	13-year-olds Math	13-year-olds Reading	TOTAL
2004–2008	-1	-7	-2	-4	-14
1999–2004	-3	-6	+1	-4	-12
1996–1999	+1	+1	+2	-1	+3
1994–1996	+2	-4	-1	-2	-5
1992–1994	-1	0	+3	+1	+3
1990–1992	+1	-12	-1	+9	-3
1986–1990	-2	—	+1	—	-1
1988–1990	—	+10	—	+3	+13
1984–1988	—	0	—	-3	-3
1982–1986	-3	—	-7	—	-10
1980–1984	—	+9	—	+3	+12
1978–1982	-3	—	-14	—	-17
1975–1980	—	-2	—	-2	-4
1971–1975	—	-9	—	-1	-10

Note: All cells report change in scale score points. Rules on accommodations changed in 1996 for math and 1998 for reading. For intervals with 1996 as an end point in math, gaps are computed from scores in which accommodations were not permitted. For math intervals with 1996 as a starting point, gaps were computed from scores in which accommodations were permitted. In reading, the same rules apply but with 1998 as the key year.

Source: Author's calculations from data provided by NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/)

Gap Changes on the Main NAEP
90th and 10th percentile, 1990–2009

Table 1-4

Years	4th-Grade Math	4th-Grade Reading	8th-Grade Math	8th-Grade Reading	TOTAL
2007–2009	0	TBA	+1	TBA	
2005–2007	0	-2	-1	-1	-4
2003–2005	0	-3	0	0	-3
2002–2003	—	+2	—	+4	+6
2000–2003	-8	—	-4	—	-12
2000–2002	—	-10	—	—	-10
1998–2002	—	—	—	-5	-5
1998–2000	—	+4	—	—	+4
1996–2000	+1	—	+2	—	+3
1994–1998	—	-8	—	-6	-14
1992–1996	-2	—	-1	—	-3
1992–1994	—	+13	—	+2	+15
1990–1992	0	—	+2	—	+2

Note: All cells report change in scale score points. Rules on accommodations changed in 1996 for math and 1998 for reading. For intervals with 1996 as an end point in math, gaps are computed from scores in which accommodations were not permitted. For math intervals with 1996 as a starting point, gaps were computed from scores in which accommodations were permitted. In reading, the same rules apply but with 1998 as the key year.

Source: Author's calculations from data provided by NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/)

gap, especially in fourth-grade reading (13 points) in 1992–1994.

The two NAEP tests are showing evidence of a common trend: the narrowing of the achievement gap between the 90th and 10th percentiles. The trend extends across NAEP combinations of subject with age or grade. The narrowing appears to have begun about 1999 on the LTT NAEP and 1998 on the main NAEP. As a way of seeing this, one can add up and compare the total gap changes before and after these dates. On the LTT NAEP, the gaps contracted by a total of 26 points from 1999 to 2007 (only eight years), surpassing the 22-point contraction from 1971 to 1999 (28 years). On the main NAEP, the contraction was 24 points on all intervals beginning in 1998 or later compared with an expansion of 3 points during the previous intervals.

Accountability systems began to dominate the policy arena in the late 1990s. As noted above, however, NAEP data do not allow for testing causal relationships. Being cross-sectional—that is, the data are collected from different cohorts of students at a single point in time—the best one can conclude from NAEP scores is that correlations exist between them and particular policies or practices at the time of testing. That said, the accountability movement must be considered a prime candidate for influencing the trends reported here.

The No Child Left Behind Act federalized accountability in January 2002, but the policy had already been embraced by a number of states. The 1999 edition of *Quality Counts* was on the theme of accountability, highlighting the widespread adoption of accountability systems at the state and district levels, in particular, as a strategy for turning around low-performing schools. Nineteen states tested students, published the results for individual schools, and offered incentives for improving test scores. In a few years, accountability spread across the country. In the 2003 edition of *Quality Counts*, policy analyst Thomas Timar summarized, "By the end of 2002, 43 states issued school report cards, 30 rated schools on the basis of performance, 28 provided some form of technical assistance to low-performing schools, 18 rewarded schools for increased performance, and 20 imposed sanctions on schools that failed to improve."[10]

Ironically, the widespread adoption of a policy makes it more difficult to measure its potential effects. As settings with alternative policies disappear, researchers lose the ability to compare the effects associated with different interventions. Research predating NCLB compared test score changes in states with accountability systems to those in states without them. It found a narrowing

of achievement gaps on NAEP related to race, ethnicity, and socioeconomic status.[11] But NCLB nationalized an incentive structure targeting low-performing students. How can we investigate the association between accountability and trends in the 90–10 gap today?

Private school students are the only group of U.S. students remaining relatively untouched by NCLB. Private schools are immune from the sanctions of NCLB. If the same gap-closing trend is evident among private schools, the phenomenon may reflect a societal-wide emphasis on raising the achievement of low achievers rather than public policies designed to do so.

Table 1-5 compares public and private school gap changes on the main NAEP. The LTT NAEP does not separately report public and private school scores. In the table, cells are shaded in each row to indicate which sector favored a narrowing of the 90–10 gap during that particular testing period. Again, the trend seems to have shifted course in approximately 1998. Simply adding up the gap changes is revealing. For intervals beginning since 1998 (i.e., starting with 1998–2000), the 90–10 gaps in the public sector have narrowed by 24 points. In private schools, they have narrowed by only 1 point. For the intervals before 1998, the gaps actually expanded by 7 points in the public sector while narrowing by 8 points in the private sector. On the main NAEP, a contraction of the 90–10 achievement gap is associated with public schools from 1998 to 2009 and private schools from 1990 to 1998. The math scores from 2009 run contrary to the prevailing trend, but also note that in many of the intervals in the table, a single set of scores runs against the prevailing trend.

Summary of Public and Private School Student Gap Changes 90th and 10th Percentiles, 1990–2009, Main NAEP

Table 1-5

Years	Public	Private
2007–2009 (partial)	+3	0
2003–2007*	-10	-4
2002–2003	+8	+5
2000–2003	-13	0
2000–2002	-11	-4
1998–2002	-5	+2
1998–2000	+4	0
1996–2000	+4	+5
1994–1998	-15	-2
1992–1996	-3	-13
1992–1994	+18	-3
1990–1992	+3	+5

Note: Does not include 2009 reading scores at 4th or 8th grade.
Sector favoring narrower gaps shaded in each row.

Source: Author's calculations from data provided by NAEP Data Explorer (http://nces.ed.gov/nationsreportcard/naepdata/)

*Scores for private school students did not meet reporting standards in 2005

Summary and Conclusion

This section of the Brown Center Report explored NAEP data. The recent release of math scores on the main NAEP triggered an outpouring of concern. The results suggested a potential slowing in the skyrocketing scores of fourth graders, an upward trend over the past two decades that, if it does not slow, projects that fourth graders in 2053 will know about the same amount of mathematics as high school seniors knew in 1990. The gains on the long-term trend NAEP and TIMSS are much smaller.

Rules are proposed for telling when a pattern in NAEP scores constitutes a real trend: when the pattern is evident in scores over multiple administrations of both NAEP tests, in reading and math, and at NAEP's two age and grade levels. The rules are applied to document a recent trend in NAEP data, the narrowing of the gap between the

… a contraction of the 90–10 achievement gap is associated with public schools from 1998 to 2009 and private schools from 1990 to 1998.

Acknowledging that NAEP data cannot demonstrate causality, the analysis nominates accountability systems as a potentially leading factor in shrinking the 90–10 gap.

nation's highest (90th percentile) and lowest (10th percentile) achievers. Acknowledging that NAEP data cannot demonstrate causality, the analysis nominates accountability systems as a potentially leading factor in shrinking the 90–10 gap. Scores from public and private school students are compared, utilizing private schools' immunity from accountability policies to form a comparison group. The contraction of the 90–10 gap is apparent in the public sector after 1998 but not in the private sector, a contrast consistent with the hypothesis that public school policies are associated with the trend.

Obviously, the association reported here is not ironclad. The phenomenon of regression to the mean may be in play, in which, over time, scores near the top of a test's distribution decline and scores near the bottom increase regardless of true changes in student performance. Cross-sectional test scores, especially those derived from a scale with a high ceiling like NAEP's, are not especially susceptible to regression to the mean, but the possibility remains that gap contraction may be a statistical artifact, not a true indicator of change in student learning. Note that the 90–10 gap on the LTT NAEP contracts throughout its 38-year history, albeit more dramatically after 1999. Tens of thousands of education policies have been enacted by state and local governments in the past several decades. Perhaps another policy or combination of policies has driven the trend. Accountability may have nothing to do with it.

Let's conclude with an important question regarding equity that lurks behind these data. Historian Larry Cuban and others have observed how the educational needs of high and low achievers are often in tension when considering the best policies to pursue equity in education.[12] In one sense, the trends since 1998 represent the best possible scenario. NAEP scores for both groups have increased. High achievers' scores have not declined, and yet the gap between the nation's best and worst students has narrowed. No one has lost out. Scores at the bottom of the distribution have simply gone up more than scores at the top. All boats are rising, but the boats at the bottom are rising a little faster than the other boats.

From another perspective, high-achieving students have lost out, and the nation has missed an opportunity to boost the achievement of its best students. If incentives can be put in place to spur student achievement to greater heights, why wouldn't we want to raise the academic learning of the nation's top students? No one loses if high achievers make gains similar to those of low achievers.

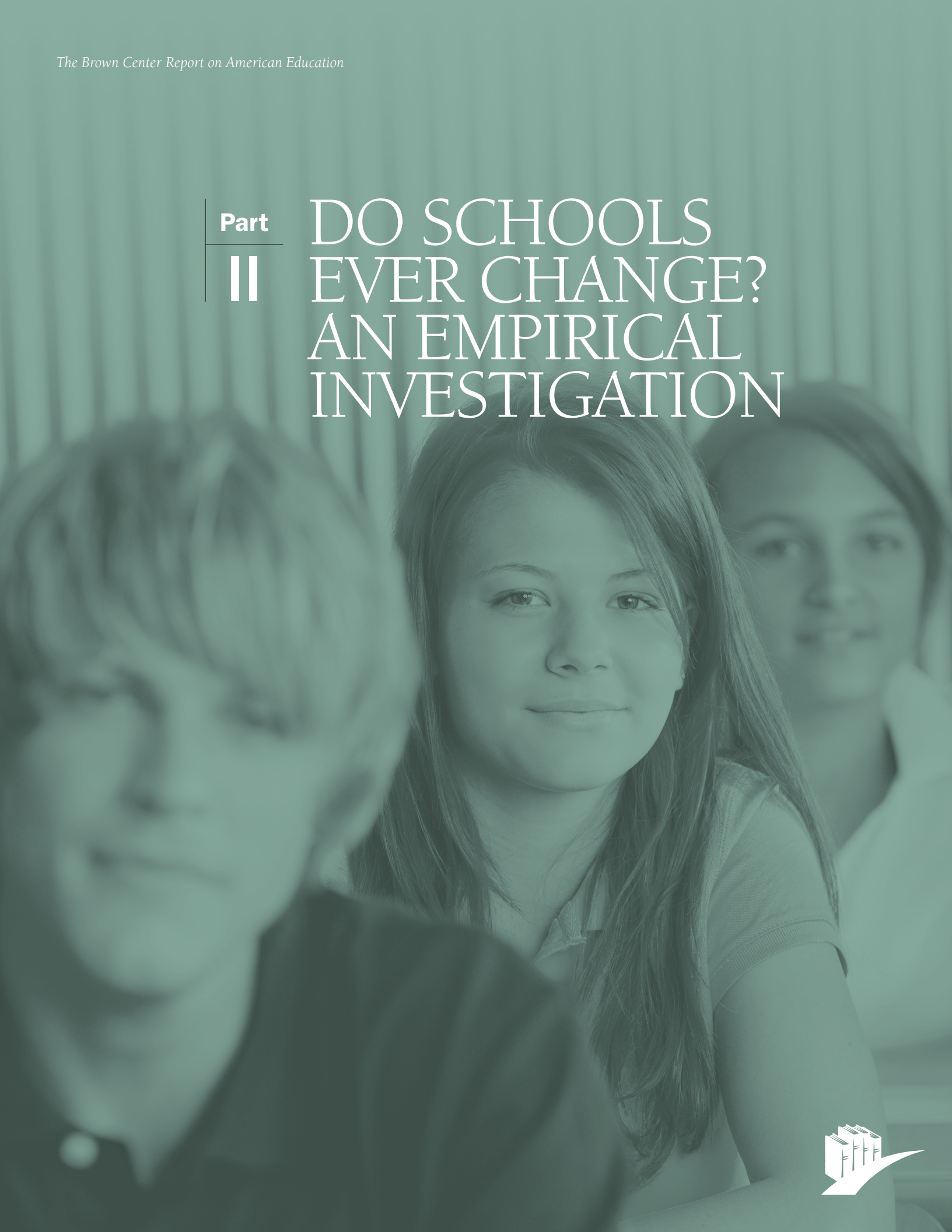
The question is what type of equity the nation wishes to seek. One version, which would be accomplished at least in theory if the current trend persisted into the distant future, holds that distinctions between high and low achievers should vanish. All students should perform the same, ideally at a very high level, but essentially the same. The 90–10 gaps would be eradicated.

A second version of equity sees differences in performance as a virtue. Consider reading and math. Some students are strong in one subject but not the other; some students excel at both subjects, and, unfortunately, some students struggle at both. Hard work, intelligence, and persistence surely contribute to a student's academic success. But we do not live in a perfect society. It is unacceptable that characteristics unrelated to achievement—race, ethnicity, gender—are statistically related to test scores. In this view of equity, achievement differences are acceptable, even celebrated, but only if they are correlated with attributes of human character

and behavior and not with demographic characteristics present at birth.

NAEP scores cannot settle which of these views to embrace. They cannot even pinpoint the exact policies or practices that may be contributing to the attainment of one ideal or the other. What they can do is map long-term trends in American student achievement, the purpose for which NAEP was originally intended. To do that accurately requires patience when test scores are released, the willingness to consider test score changes in a broad, historical context, and a healthy dose of skepticism when commentators try to attribute changes in NAEP scores to a particular policy or practice.

Part II DO SCHOOLS EVER CHANGE? AN EMPIRICAL INVESTIGATION



THE OBAMA ADMINISTRATION HAS MADE FIXING PERSISTENTLY failing schools a primary focus of its education policy. "Turnarounds" are a hot topic. States applying for Race to the Top grants in 2010 will receive more favorable treatment if plans for turning around failing schools are included in the application. Much of the rhetoric on turnarounds is pie in the sky—more wishful thinking than a realistic assessment of what school reform can actually accomplish. That said, schools do indeed change. Turnarounds are not unicorns. But how likely are they to be seen?

This section of the Brown Center Report presents an investigation into the probability of turning around failing schools. It compares the 1989 and 2009 test scores of 1,156 California schools, all of the schools that contained an eighth grade in 1989 and were still operating in 2009. The turnaround question is one of several addressed in the analysis. How much did any school's performance change over this twenty-year period? How many schools that were languishing at the bottom in 1989 joined the state's top-performing schools by 2009? How many made even a little bit of progress? Conversely, how many schools fell from the ranks of high-performing schools to the lower end of the continuum?

Background

Let's cover a few details about the study before turning to the data. California has a long history of testing. Using standardized tests to annually measure student progress began in 1962. Complaints arose that nationally normed commercial tests, although informative for revealing where the state stood relative to the rest of the country, did not adequately reflect California's curriculum and took too long to administer. Beginning in 1973, California tested students in Grades 3, 6, 8, and 12 on the California Assessment Program (CAP), the state's own test that employed the then novel technique of matrix sampling. Eighth graders were

How many schools that were languishing at the bottom in 1989 joined the state's top-performing schools by 2009?

School Composite Test Scores, 1989 and 2009, by quartile

Table 2-1

		1989 Composite Score				
		Quartile 1	Quartile 2	Quartile 3	Quartile 4	TOTAL
2009 Composite Score	Quartile 4	4	22	81	182	**289**
	Quartile 3	23	78	108	80	**289**
	Quartile 2	79	121	69	20	**289**
	Quartile 1	184	69	29	7	**289**
	TOTAL	**290**	**290**	**287**	**289**	**1156**

tested in the spring, and results for every school were published to great fanfare in the fall.[13] In the 1980s, the scores were an important element of the state's accountability system, the Program Quality Review, which also included periodic visits from a team of reviewers who made recommendations for improvement.[14] CAP testing lasted until its funding was vetoed by Governor George Deukmejian in 1990.

For the current study, 1989 eighth-grade CAP scores were available for reading, math, history/social science, and science. The 2009 scores are in English/language arts (which is paired with the 1989 reading score), math, history, and science.[15] For both 1989 and 2009, we created a composite score for each school—simply the average of the four subject scores—and computed percentiles for the composite scores. As a check on whether the results might be influenced by weighting, a composite score using differential weighting of subjects (counting reading, for example, as more important) was also calculated to mirror the formula of California's Academic Performance Index. The results were not significantly affected. Using scale scores or percentile scores also did not produce appreciably different results.

The analysis uses percentiles, which place performance on a common scale. Imagine a list of schools ranked from 1 to 99 by their scores on a test. Percentiles describe where in that order a particular school falls, with the 99th rank assigned to the highest-performing school and 1 assigned to the lowest-performing school. Percentiles readily demarcate quartiles of performance, the cut points being at the 75th, 50th, and 25th percentiles.

Limitations of the study should be noted. Schools that closed or did not have test scores available in either year are omitted. That means schools that opened since 1989—and California opened hundreds of them—are not part of the study. In addition, this is an empirical investigation summarizing an historical pattern so it assumes that the past has something germane to say about the future. That any twenty-year period of the past can be instructive on the question posed by the study's title, whether schools "ever" change, is admittedly debatable. Perhaps circumstances in the next twenty years will be so different from the 1989–2009 period that any inference about the future is invalid. And perhaps California schools are not representative of schools nationally and a study of schools in another state would generate a different set of findings.

Results

Table 2-1 displays the cross tabs of quartiles of academic performance for 1989 and 2009. Quartile 1 consists of schools scoring in the bottom 25 percent; Quartile 4 are those scoring in the top 25 percent. Columns represent the quartiles for 1989, and rows are the quartiles for 2009.

Let's examine the first column, showing schools that scored in the bottom quartile in 1989 (hereafter called "low-performing" schools). How were they doing twenty years later? Of these 290 low-performing schools, 184 (or 63.4 percent) scored in the lowest quartile again in 2009. Approximately 27.2 percent (seventy-nine schools) moved up to

the second quartile; 7.9 percent (twenty-three schools) improved to the third quartile, and 1.4 percent (four schools) moved all the way up to the fourth quartile (hereafter called "high-performing" schools).

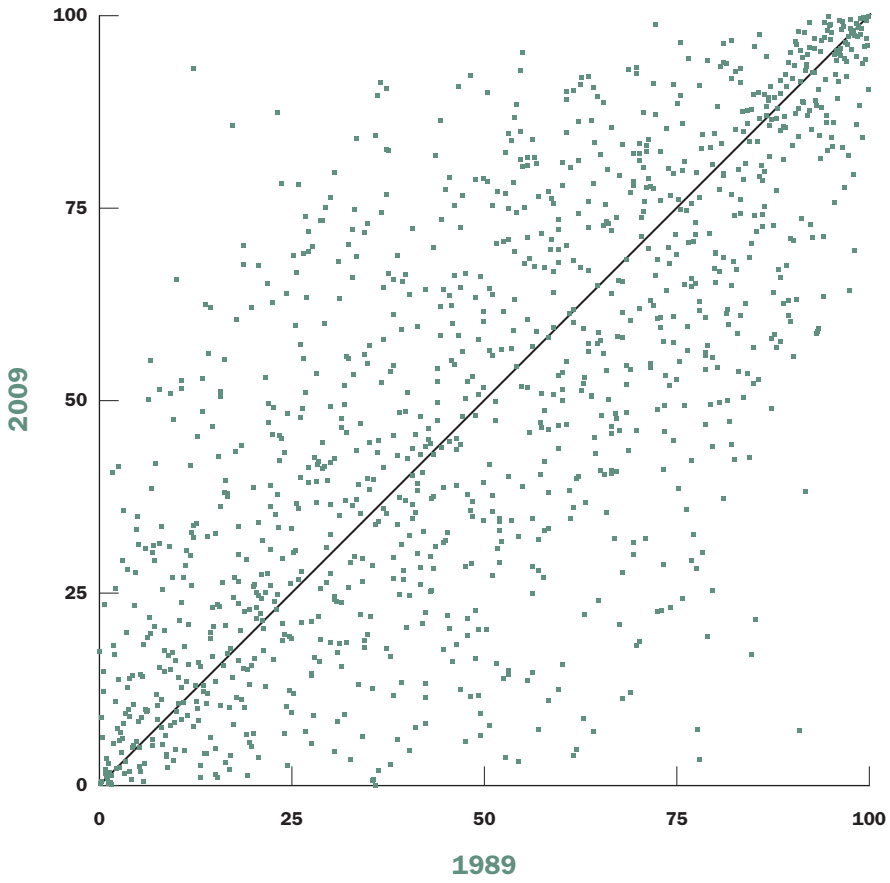
The statistics are eye-popping and, in a way, depressing. School achievement appears astonishingly persistent. Nearly two-thirds of low-performing schools in 1989 are still low performers two decades later. But there is a ray of hope here, too. About one-third of these schools evidence improvement. Nevertheless, it is highly unlikely that a low-performing school becomes a high-performing school. The chances (four out of 290) are less than one out of seventy.

A mirror image of this pattern can be found in the Quartile 4 schools. Falling from the highest quartile is as difficult as rising from the lowest. Almost two-thirds of Quartile 4 schools in 1989 were still there in 2009 (63.0 percent, 182 schools). About 27.7 percent (eighty schools) fell to the third quartile, and 6.9 percent (twenty schools) declined to the second quartile. Only 2.4 percent (seven schools) of 1989's highest-achieving schools scored in the lowest quartile in 2009.

Perhaps the persistence of school achievement is most acute in the two-tails of the distribution—that is, among the lowest- and highest-performing schools. After all, schools in the bottom quartile cannot move lower (called a "floor effect"), and schools in the top quartile cannot move higher (a "ceiling effect"). Let's scrutinize the two middle quartiles from 1989 to see if those schools are more likely to change. Of the 290 schools in Quartile 2 in 1989, 121 (41.7 percent) remained in Quartile 2 in 2009. Approximately equal numbers moved up to Quartile 3 (seventy-eight) as moved down to Quartile 1 (sixty-nine). Combined, these 147 movers to an adjacent quartile comprise 50.7 percent of schools, meaning that about 92.4 percent of the Quartile 2 schools scored within one quartile in 2009. The remaining 7.6 percent, twenty-two schools (or about one out of thirteen), joined the highest-achieving schools in Quartile 4.

School Achievement, 1989 and 2009

Fig 2-1



The Quartile 3 schools of 1989 show a similar pattern, with 89.9 percent scoring within one quartile in 2009. A few more moved up (eighty-one schools, or 28.2 percent) than down (sixty-nine schools, or 24.0 percent). To answer the question posed in the previous paragraph, the middle-quartile schools do evidence more mobility than the schools in the top

… it is highly unlikely that a low-performing school becomes a high-performing school. The chances (four out of 290) are less than one out of seventy.

Probability of School Test Score Changes, 1989 to 2009

Table 2-2

Change (percentile points)	Frequency (percent)	Proportion Exceeding
4.54	25.0	30 out of 40 schools
10.00	43.5	23 out of 40 schools
12.22	50.0	20 out of 40 schools
20.00	68.5	13 out of 40 schools
23.54	75.0	10 out of 40 schools
30.00	84.4	6 out of 40 schools
40.00	92.2	3 out of 40 schools

Note: Change is an approximated upper limit. For example, 25% of schools changed 4.54 percentile points or less from 1989 to 2009. That means that 75%, or 30 out of 40 schools, had larger changes.

and bottom quartiles. But not a lot more. The movement that occurs tends to be to an adjacent quartile and can be either up or down, that is, a manifestation of either improving or deteriorating test scores.

Figure 2-1 displays the relationship of 1989 and 2009 composite scores in a scatter plot. Clearly, the two scores are related, but the relationship is not fixed. If the relationship were a perfect correlation, with 1989 scores exactly predicting 2009 scores, the dots would all fall on the 45-degree line. Some schools do indeed change a great deal, as evidenced by their deviation from the line. The Pearson correlation coefficient for 1989 and 2009 achievement is 0.73. Correlation coefficients are a measure of association between two variables and range from -1.00 to 1.00. A correlation coefficient of 1.00 indicates a perfectly positive correlation and -1.00 a perfectly negative one. A value of 0.00 indicates no statistical relationship. The value of 0.73 indicates a moderately strong relationship.

Table 2-2 provides statistics on the likelihood of score changes of different magnitudes. The median change from 1989 to 2009 was 12.22 percentile points, meaning that half of the schools changed about 12 percentile points or less and half changed more. About 68.5 percent changed no more than 20 percentile points, 84.4 percent changed no more than 30 percentile points, and 92.2 percent changed no more than 40 percentile points.

What was the probability that a school functioning at the 10th percentile or below in 1989 improved to the state's average (50th percentile) in 2009? Today, officials attempting to boost the performance of schools on NCLB's intervention list face a similar challenge. Many of these institutions score in the bottom 10 percent, which corresponds to about 9,700 schools nationally. How likely is a gain of 40 percentile points?[16] About 7.8 percent, or three of every forty schools, in our sample evidenced a change that large, a change being either a gain or a loss. But only schools falling in the middle of the distribution can change that much in either direction. A school at the 10th percentile cannot decline more than 9 percentile points (the floor effect again), so a reduced estimate is in order. In fact, of the 115 schools scoring at the 10th percentile or below in 1989, only four of them (or 3.5 percent) scored at the state average or above in 2009.

Discussion

The data in this study empirically support an intuition probably held by many observers of school reform: turning around a failing school is extremely difficult—but not impossible. Reformers are by nature hopeful, and they will no doubt consider the long odds presented here as irrelevant to their own probability of success. They will reason: no one is doing what our school is doing, with the talent that we have assembled, working as hard as we are working. That kind of dedication is admirable, and nothing in this study should be construed to diminish it.

Nevertheless, there are some hard lessons here. If all California schools had improved academically from 1989 to 2009, the study's findings might be easily dismissed. Percentiles are a relative measure. A school that stays at the 10th percentile for twenty years can be making substantial progress if the state is gaining academically. Being at the tail end of the line isn't so bad if the entire line is moving forward. Unfortunately, that is not the case. The state scored 6 points below the national NAEP average for eighth-grade math in 1990 and 12 points below the national average in 2009.

Even if California lagged behind the rest of the nation, one would think that the numerous reforms that have been attempted since 1989 would mix up school rankings within the state. Californians tried just about everything: traditional and reform mathematics; whole language and phonics-based reading instruction; accountability systems running from "soft" and "professional" to "hard" and "punitive"; tests dominated by multiple-choice items and one notable "state of the art" test (the California Learning Assessment System) that offered constructed-response items; professional development galore; a Rube Goldberg system of financing schools featuring equalized per-pupil spending supplemented by dozens of categorical programs; state takeovers of several schools; a vibrant charter school movement; home-grown philanthropists and Silicon Valley high-tech companies pouring money into their favorite schools; booming advanced placement enrollments in high schools; detracking in middle schools; and the largest bilingual program in the nation that, by a 1998 voters' initiative, was sharply curtailed.

That's a lot of activity that should have shuffled the deck. But twenty years later, the deck looks remarkably unshuffled. Factor in the natural volatility in school test scores—the amount they vary from year to year due to measurement error—and the stability of the scores in this study is truly amazing.

What about other efforts to turn around organizations? How successful are they in shuffling the deck over a twenty-year period? Professional sports teams make an interesting comparison group. We identified the lowest quartile of performers in 1989 in three professional sports: baseball, basketball, and football. The selection was made based on win-loss records for the year, with seven teams from each sport (a total of twenty-one teams) constituting the sample. How did they do twenty years later? In 2009, five of these teams performed in the bottom quartile: a single team in football (the Tampa Bay Buccaneers), two in basketball (the Los Angeles Clippers and Sacramento Kings), and two in baseball (the Pittsburgh Pirates and Cleveland Indians). Granted, the sample size is small; however, the actual rate of Quartile 1 repeaters (23.8 percent) is close to the 25 percent that one would expect if the 2009 performance rankings were simply assigned to teams randomly. Recall that the rate of repeaters for Quartile 1 schools was 63.4 percent.

Sports franchises and schools differ in a multitude of ways, of course. The professional sports leagues aggressively attempt to break the link between past and future performance, to reshuffle the deck so that losing teams have an enhanced chance of becoming winning teams. Teams draft new players in reverse order of their previous year's record so that the worst teams have the best chance of selecting the superstars of the future. Salary caps protect small franchises from being raided by bigger, wealthier organizations. Teams share revenue from television and other media. These policies are compensatory. They help low-achieving teams.

… one would think that the numerous reforms that have been attempted since 1989 would mix up school rankings within the state.

Correlation Matrix of 1989 and 2009 Variables

Table 2-3

	1989 SES	2009 SES	1989 Achievement	2009 Achievement
1989 SES	—	0.75	0.79	0.67
2009 SES	0.75	—	0.74	0.77
1989 Achievement	0.79	0.74	—	0.75
2009 Achievement	0.67	0.77	0.75	—

Compensatory interventions exist in education, too, but they appear to have a more difficult time boosting low-performing schools. The most profound way that schools differ from other organizations is that schools are inextricably bound to the communities from which they draw students. Since the Coleman Report in 1966, researchers have documented the correlation of students' achievement and socioeconomic status (SES).[17] Table 2-3 displays the correlation coefficients for the achievement and SES variables in the current study. The 1989 measure of school SES correlates with the 2009 measure at 0.75, the same as the correlation coefficient for 1989 and 2009 test scores. The relative ranking of schools by SES is as stable as the ranking by test scores.

Is the persistence of test scores driven by the persistence of SES? No, the relationship weakens when changes in both variables are examined. The correlation of change in SES and change in achievement is only 0.34, meaning that less than 12 percent (the correlation squared) of the variance in achievement change can be explained by change in SES. We sorted the schools into quartiles of SES change and discovered that the top 25 percent of schools, those with the most positive SES changes, gained about 8.5 percentile points in achievement from 1989 to 2009. The bottom quartile of schools, those with the largest declines in SES, lost about 10.5 percentile points. Changes in the SES of students were related to changes in school performance, but they were by no means determinative.

What causes, or at least reinforces, the persistence of school test scores over the decades? The response to that question can only be speculative. Achievement seems to be part of the institutional DNA of schools, handed down from decade to decade, the past influencing the future. Some of it may be due to how school populations change, with teachers and administrators—and kids and their parents—slowly transitioning in and out of schools. The newcomers learn about the culture of a school from those who have been there and are preparing to leave. If failing schools are ever to be turned around, much more must be learned about how schools age as institutions—how they got to where they are and the factors influencing where they are going. More research is needed analyzing longitudinal data and tracking the institutional trajectories of schools over extended periods of time.

Future research may also be able to find out how particular policies affect the fate of schools. This study has documented not only the persistence of school test scores, but also the formidable odds against turning around failing schools. Hopefully, the next generation of research will shed light on ways of making that goal more achievable.

Part III WHAT DO WE KNOW ABOUT CONVERSION CHARTER SCHOOLS?

MOST CHARTER SCHOOLS ARE START-UPS, CREATED ORIGINALLY as charter schools. Conversion charter schools, in contrast, were once conventional public schools and then converted to charter status. Conversions make up only about 10 percent of charter schools nationally.[18] Despite their small numbers, conversions attract a lot of interest from school reformers as a tool for turning around failing schools. The idea is this: after several years of failing to improve by conventional means, a school is shut down, the staff relieved or reassigned, and the school reopened as a charter—with new teachers and administrators.

Converting a failing school to a charter school is one of the remedies of No Child Left Behind. The Obama administration has also embraced charters, most notably in the Race to the Top program, which encourages states to lift statutory limits (or "caps") on the number of charters allowed to open. Currently, fewer than 1 percent of the schools in restructuring under NCLB have been converted to charters. It is not yet a popular option, and not much empirical evidence is available on the strategy's impact.[19]

Studies analyzing the effectiveness of charter schools have produced mixed results. Both statistically sophisticated meta-analyses and narrative reviews of the research have concluded that although the quality of charters varies dramatically—there are excellent and awful charter schools—the difference between charters and traditional public schools is probably small.[20] Only a few studies have disaggregated charters by type and specifically examined conversions.

The following analysis examines conversions in California. About 16 percent of charters in California are conversions, the most of any state. The discussion draws upon data from two recent Brown Center studies that compare conversions before and after they became charter schools. The first study (coauthored by Tom Loveless, Andrew P. Kelly, and Alice M. Henriques) examines reading and math scores, along with data on other school characteristics, from two eras: 1986 to 1989 and 2001 to 2004.[21] The second study, conducted by Brown Center staff, serves as a follow-up, comparing test scores from 1986 and 2008.

Two cohorts of schools are analyzed. To be included in the first cohort, called the 2004 cohort, a school had to have third-grade test scores available in reading and math for 1986 and 2004, operate in 1986 as a traditional public school, and operate in 2004 as a charter school. The second cohort, called the 2008 cohort, had the same requirements except that 2008 served as the most recent year of data collection. Conversion charters are a fluid group, with schools opening and closing, reconverting to public schools, new schools converting, schools adding grades, and so on. Only about half of the schools in the 2004 cohort are also members of the 2008 cohort. The studies should be regarded as exploratory, not as an evaluation of conversion charters' effectiveness. The analysis of achievement focuses on school test scores (as opposed to student-level scores) and cannot control for selection effects that would affect the scores—selection of students into schools, schools into conversion status, and teachers and administrators into one type of school or the other.

These forms of selection are non-random and well-known in the research literature.[22] For example, families who choose to send their children to charter schools may do so because the children are struggling academically in school. That biases charter school test scores downward. Parents who go to the trouble of transferring children to charters, however, may be highly motivated toward education, biasing scores upward. Schools that elect to convert to charter status may have unusually talented teachers who are willing to innovate, also biasing test scores upward. But converting a failing school to a charter is also a turnaround strategy, biasing scores downward.

Moreover, readers should keep in mind that although conversions provide the closest possible match to schools that may be chartered through reconstitution in future years, they differ in one critical respect. Almost all of today's conversions became charters through their own initiative. They kept most of their original staff. Schools compelled to convert and to completely change personnel may perform differently.

Achievement

Previous studies of academic achievement in California's conversions have reported positive findings, albeit modest in size and qualified by sampling constraints. A RAND study found that elementary grade conversions outperform both start-ups and traditional schools in math and produce similar results in reading.[23] The analysis did not include cyber charters, focusing exclusively on brick-and-mortar schools. At the secondary level, start-ups outperformed both conversions and traditional public schools. A study in the 2003 Brown Center Report analyzed three years of achievement data. The study found that both conversions and start-ups produced lower test scores than traditional public schools, but all three groups of schools produced similar gains. Once demographic controls were introduced, conversions outperformed start-ups in producing gains.

Table 3-1 presents achievement data from the current analysis. Percentile scores reflect a school's ranking relative to the state as a whole, with the California average pegged at the 50th percentile (see

Achievement in Conversion Charters: Two Cohorts

Table 3-1

	2004 Cohort (N = 49)			2008 Cohort (N = 60)		
SUBJECT	**1986**	**2004**	**CHANGE**	**1986**	**2008**	**CHANGE**
Reading	**41.2**	**43.4**	**+ 2.2**	**53.6**	**53.9**	**+ 0.3**
Math	**40.8**	**43.6**	**+ 2.8**	**55.7**	**54.4**	**- 1.3**
SES	**41.7**	**45.2**	**+ 3.5**	**47.4**	**59.0**	**+11.6**

What else do we know about conversions? They differ from start-up charter schools in several respects.

Part II of this Brown Center Report for an explanation of percentiles). Functioning as traditional public schools in 1986, the 2004 cohort scored a little below the state average—41.2 in reading and 40.8 in math. As charter schools in 2004, the schools scored 2.2 points higher in reading and 2.8 points higher in math. The 2008 cohort scored a little above average in both 1986 and 2004. The reading score of 53.6 in 1986 inched ahead to 53.9 in 2008. The math score fell just over a point, from 55.7 to 54.4.

The biggest surprise is in socioeconomic status (SES). The state computes a school's SES index based on a survey of parents' education.[24] The index was scaled differently in 1986 from more recent years, but using percentiles in the analysis mitigates the discrepancy. We calculated percentile ranks for all of the state's schools on this statistic and then computed the average for the two cohorts of conversion charters. This indicator is probably more reliable than the percentage of students qualifying for free and reduced lunch, the conventional proxy for SES, because charter schools often do not participate in that federal program.

The SES index for the 2004 cohort behaves like its test scores, rising from 41.7 in 1986 to 45.2 in 2004. Students in conversion schools were more likely to come from disadvantaged backgrounds than the typical California student, and this was true both before and after the schools converted to charter schools. But the 2008 cohort looks different; SES rose more than 11 percentile points, from 47.4 in 1986 to 59.0 in 2008. The conversion students of 2008 came from significantly more advantaged households than students attending the same schools in 1986.

Why are the two cohorts different? Several conversions left the ranks of charter schools from 2004 to 2008 and returned to their local school districts. Many of these schools, in particular, eleven Los Angeles schools in what was known as "the Crenshaw Dorsey cluster," were located in poor communities. Schools that joined the ranks of conversions during this time were from average to above-average SES communities. As a result, the state's conversion schools are now attended by students whose SES is slightly above average.

How Conversions and Start-Up Charters Compare

What else do we know about conversions? They differ from start-up charter schools in several respects. In 1992, after California followed Minnesota to become the second state with a charter school law, a large number of traditional public schools considered converting to charter status. Despite the hurdles, in the first few years about half of California's charters were conversions.[25] Under state law, a public school may apply to its district for conversion only if a majority of its full-time, tenured teachers sign the application. The conversion must win the approval of the district school board; rejections can be appealed to the state. Staff members maintain collective bargaining rights after conversion. Also, the schools must give enrollment priority to students residing within the old geographical attendance areas. Conversions usually continue operating at the same facility with most of the same faculty and students.[26] In contrast, start-ups do not have attendance boundaries, operate in all kinds of facilities, and hire mostly less experienced, nonunion teachers.

Table 3-2 highlights some of the key differences between start-ups and conversions. State averages are also provided. Note that the statistics are based on the 2004 cohort. A confession: Statistics on school demographic and staffing characteristics usually change so slowly that there was

School Characteristics (2004 Cohort)

Table 3-2

3rd grade	State Average (N=5153)	Start-Ups (N=57)	Conversions (N = 49)
STUDENT DEMOGRAPHICS			
White	35.7%	54.1%	30.2%
Hispanic	43.1%	22.8%	44.4%
Black	8.0%	13.6%	20.5%
Asian	11.0%	4.0%	4.1%
Other	2.2%	5.5%	0.8%
COMMUNITY			
Urban	43.2%	40.4%	61.2%
Suburban	46.7%	40.3%	28.6%
Rural	10.1%	19.3%	10.2%
ENROLLMENT			
Median enrollment	572	245	641
Median students per grade (calculated)	86.9	28.9	98.3
Student/teacher ratio	19.2	18.8	18.1

no apparent need to update them for this report. So we did not collect fresh data on the 2008 cohort. That was a mistake. Indeed, considering the changes noted above in the population of conversion charters, these statistics undoubtedly have changed. Nevertheless, they are worth reporting—both to inventory what is currently known about conversions and to pinpoint areas in which future research is needed. Most of the differences that existed are so large that they probably still hold. It's the size of the differences that may now be different.

Conversions in the 2004 cohort looked more like traditional public schools than start-up charters. Compared with start-ups, they served a larger percentage of Hispanic and black students and fewer white students. Conversions tended to be located in urban areas (61.2 percent), while start-ups were divided equally between the suburbs and urban areas. Conversions were much larger (641 students compared with 245 for start-ups) and had more students per grade level (a way to control for schools having different numbers of grades). One characteristic on which conversions were closer to start-ups than the state average was student-teacher ratio. The state average for third grade was 19.2 students; it was 18.8 for start-ups and 18.1 for conversions.

When it comes to teacher characteristics, conversions also looked more like traditional public schools compared with start-ups (see Table 3-3). They had more teachers certified in elementary education (91.4 percent versus 81.2 percent) and bilingual education (27.6 percent versus 6.7 percent). Favoring bilingual certification mirrors the conversion schools' larger Hispanic clientele. Teachers at conversion schools also had more years of teaching experience (10.8) than those at start-ups (7.9), but less than the average teacher in the state (12.7). Not shown in the table is that teachers in the 2004 cohort averaged 14.5 years of experience in 1989 (the state average was 14.0). Conversions hired less experienced teachers after attaining charter status. All of this must be put in perspective. Although both conversions and start-ups have less experienced teachers than the average California school, they still have very experienced teaching staffs.

Summary and Conclusion

Converting failing schools to charters has been proposed as an effective way to reform schools. But we do not know much about the success or failure of conversions, despite their existence for more than 15 years. This study examined data on California conversions, the state with the most, and many of the oldest, conversions. The schools' reading and math scores have not changed a lot from 1986, when they operated as traditional public schools, to more recent years when they operated as charters. That is no reflection on the schools' quality. Charters are difficult to evaluate and require more complicated analyses than the current study

Teacher Characteristics (2004 Cohort)

Table 3-3

3rd grade	State Average (N=5152)	Start-Ups (N=57)	Conversions (N = 49)
Certified in elementary education	92.8%	81.2%	91.4%
Certified in bilingual education	13.8%	6.7%	27.6%
Number of years teaching	12.7	7.9	10.8
Education—master's	27.6%	26.4%	24.8%
White	72.0%	84.0%	61.1%
Hispanic	16.3%	8.0%	17.0%
Black	4.5%	4.2%	15.4%
Asian	5.0%	1.4%	4.6%
Female	85.6%	82.3%	84.7%
Male	14.4%	17.7%	15.3%

allows. But it is fair to say that the two cohorts of conversions in the current study evidence no significant institutional change in achievement over two decades (echoing the finding in Part II of little change in California's traditional public schools during the same time period).

What kinds of studies are needed? As noted above, the most daunting methodological challenge in charter school research is controlling for selection bias stemming from unobserved variables. Most students are not randomly assigned to schools, and schools are not assigned to charter or traditional public school status. Recent studies that have employed randomized designs have been generally favorable toward charters. They exploit the fact that charters hold lotteries for open seats, creating randomly selected experimental (the lottery winners) and control groups of students (the latter being lottery losers who return to traditional public schools).[27] These studies have their own limitations, but they represent a significant step forward in evaluating charter schools.[28] Randomized studies of conversions are needed.

We do know a few things about conversions. Compared with start-up charters, they are two to three times as large, more likely to be located in urban communities, and they serve a larger proportion of black and Hispanic youngsters. Conversions feature more experienced teachers who are more likely to hold formal teaching credentials, especially in bilingual education. In these respects, conversions look more like traditional public schools than start-ups. Note that the composition of the teaching staff is one element that differentiates compelled conversions from traditional public schools that convert willingly.

Very careful research is also needed on why many conversions revert to traditional public schools. As noted above, in California about half of the early charters in the 1990s were conversions. Now the figure is only 16 percent. Some of the largest charter management organizations have been reluctant to take on failing schools as turnaround projects.[29] They prefer starting schools from scratch rather than inheriting struggling schools, even those starting over after reconstitution. Conversions must negotiate with their former districts over the use of district facilities, provision of services, and union rules. Moreover, flexibility in lengthening the school day or year—an innovation many successful charters have embraced—can be constrained by the collective bargaining agreements that conversions must follow.

Converting failing schools to charter schools has generated tremendous interest in recent years. That interest rests on the hope of reformers that chartering offers a way to radically change the operations of a school, to redirect its institutional energies toward success rather than failure. Based on what is currently known about conversion schools, that is only a hope, not an intervention documented as having a high probability of success. More must be learned about conversion charters if they are to realize their promise as a tool of school reform.

NOTES

1 Richard F. Elmore and Milbrey Wallin McLaughlin, *Steady Work: Policy, Practice, and the Reform of American Education* (Washington, DC: The RAND Corporation, 1988).

2 Arne Duncan and David Driscoll quoted in Libby Quaid, "Math Tests: Fourth-Grader Progress Stalls," *Associated Press Online*, October 14, 2009. Mark Schneider, NAEP Math Results Hold Bad News for NCLB, American Enterprise Institute Blog, October 14, 2009. Available at http://blog.american.com/author=50. Diane Ravitch, "Time To Kill 'No Child Left Behind,'" *Education Week*, June 10, 2009. Robert Tomsho, "U.S. Math Scores Hit a Wall," *Wall Street Journal*, October 15, 2009, p. A3. Sean Cavanagh, "NAEP Scores Put Spotlight on Standards," *Education Week*, October 19, 2009.

3 The extrapolation is imprecise. It is also theoretical. Projecting gains several years above grade level is dicey on a test that contains few, if any, above-grade-level items.

4 Fordham Institute's review of math tests ranked TIMSS ahead of NAEP in mathematics content. W. Stephen Wilson's math review in *Stars by Which to Navigate*, (Washington, DC: Thomas B. Fordham Institute, 2009). Available at http://edexcellence.net/doc/20091008_NationalStandards.pdf

5 The projections for 2022 and 2053 are based on fourth- and eighth-grade scores in 1990 and the rates of gains for fourth graders from 1990 to 2007.

6 National Mathematics Advisory Panel, *Report of the Task Group on Assessment*, 2008. Available at www.ed.gov/MathPanel

7 Using a discontinuity design bolsters the ability to uncover meaningful correlations from cross-sectional data. See Thomas Dee and Brian Jacob, *The Impact of No Child Left Behind on Student Achievement. NBER Working Paper # 15531* (Cambridge, MA: National Bureau of Economic Research, 2009).

8 Tom Loveless, Part I of *High-Achieving Students in the Era of No Child Left Behind* (Washington, DC: Thomas B. Fordham Institute, 2008).

9 Sam Dillon, "'No Child' Law Is Not Closing a Racial Gap," *New York Times*, April 28, 2009, p. A1.

10 Education Week, *Quality Counts: Rewarding Results, Punishing Failure* (January 11, 1999); Education Week, *Quality Counts: If I Can't Learn from You* (2003).

11 Martin Carnoy and Susanna Loeb, "Does External Accountability Affect Student Outcomes? A Cross-State Analysis," *Educational Evaluation and Policy Analysis* 24, no. 4 (2002), pp. 305–331; Eric A. Hanushek and Margaret E. Raymond, "Does School Accountability Lead to Improved Student Performance?" *Journal of Policy Analysis and Management* 24 (2005), pp. 297–327.

12 Larry Cuban, "Reforming Again and Again and Again," *Educational Researcher*, 19(1), (1990), pp. 3–13.

13 A 1985 headline read, "S.D. 8th graders Follow State Slide in Test Scores," referring to students in San Diego. David G. Savage, *Los Angeles Times* (San Diego County edition), November 9, 1985, Metro p. 1.

14 California State Department of Education, *California Assessment Program Annual Report, 1985–86.* For example of the PQR, see David D. Marsh and Patricia S. Crocker, "School Restructuring: Implementing Middle School Reform," in *Education Policy Implementation,* edited by Allan R. Odden (Albany: State University of New York Press, 1991).

15 For 2009 mathematics scores, we used seventh-grade scores because all students take the same test. Multiple tests are offered in mathematics at eighth grade (algebra, geometry, general math, and integrated math).

16 The 10th percentile is the upper limit of the bottom tenth of schools. The mean percentile of the bottom 10 percent is about 5.0, so the actual gain needed to get to the 50th percentile is more than 40 points.

17 James S. Coleman, *Equality of Educational Opportunity* (Washington, DC: U.S. Printing Office, 1966).

18 NAPCS Dashboard, 2009 (www.publiccharters.org/dashboard/)

19 Caitlin Scott, *Improving Low-Performing Schools: Lessons from Five Years of Studying School Restructuring Under No Child Left Behind* (Washington, DC: Center on Education Policy, 2009).

20 Julian R. Betts and Y. Emily Tang, *Value-Added and Experimental Studies of the Effect of Charter Schools on Student Achievement: A Literature Review*, National Charter School Research Project (Seattle, WA: Center on Reinventing Public Education, 2008; Tom Loveless and Katharyn Field, "Perspectives on Charter Schools," in *Handbook of Research on School Choice*, edited by Mark Berends, et al. (Mahwah, NJ: Lawrence Erlbaum Associates, 2009), pp. 99–114. Also see NAPCS (2009). An exception to the modest effect sizes typical in the literature was found by Caroline M. Hoxby, Sonali Murarka, and Jenny Kang, *How New York City's Charter Schools Affect Achievement, August 2009 Report* (Cambridge, MA: New York City Charter Schools Evaluation Project, September 2009). Students in charter schools gained about 0.8 standard deviations over eight years, closing most of the gap between poor students in NYC schools and peers in Scarsdale.

21 Tom Loveless, Andrew P. Kelly, and Alice M. Henriques, *What Happens When Regular Public Schools Convert to Charter Schools?* Presented at the Entrepreneurship in Education Conference, University of California, Los Angeles, June 9, 2005.

22 For a complete discussion of the methodological challenges of selection in charter school research, see Dale Ballou, Bettie Teasley, and Tim Zeidner, *Charter Schools in Idaho*. Presented at the National Conference on Charter School Research at Vanderbilt University, Nashville, Tennessee, September 29, 2006.

23 Ron Zimmer, et al., *Charter School Operations and Performance: Evidence from California* (Santa Monica: RAND, 2003).

24 The 2005 paper used a different SES indicator: an estimated percentage of students qualifying for free and reduced lunch. See Loveless, et al. *What Happens* for details.

25 Eric Premack, "Charter Schools: California's Education Reform 'Power Tool,'" *Phi Delta Kappan* 78 (September 1996), p. 60.

26 Catherine Maloney and Frank Kemerer, "Charter Schools: Opportunities and Challenges," in *Urban School Reform: Lessons from San Diego*, edited by Frederick Hess (Cambridge: Harvard Education Press, 2005), pp. 243–262.

27 Studies of New York City and Boston charter schools uncovered large positive effects. Thomas Kane et al. *Informing the Debate: Comparing Boston's Charter, Pilot and Traditional Schools* (Boston, MA: The Boston Foundation, 2009). Bifulco and Ladd detected negative effects. Robert Bifulco and Helen Ladd, "The Impact of Charter Schools on Student Achievement: Evidence from North Carolina," *American Education Finance Association* 1 (Winter 2006), pp. 50–90.

28 Among the limitations: 1) Findings can only be generalized to parents motivated enough to apply for the lottery in the first place; 2) Only charters that are oversubscribed hold lotteries, and they may be better than average; 3) More academically oriented lottery losers may go to private schools rather than return to public schools, therefore exiting the control group; 4) Attrition reduces sample sizes and the studies' statistical power to detect effects.

29 Catherine Gewertz, "Duncan's Call for School Turnarounds Sparks Debate," *Education Week*, July 21, 2009.

THE BROOKINGS INSTITUTION

STROBE TALBOTT
President

DARRELL WEST
Vice President and Director
Governance Studies Program

BROWN CENTER STAFF

GROVER "RUSS" WHITEHURST
Senior Fellow and Director

TOM LOVELESS
Senior Fellow

MICHELLE CROFT
Research Analyst

KATHARYN FIELD-MATEER
Former Research Coordinator

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

BROWN CENTER on
Education Policy
at BROOKINGS

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

The Brown Center on Education Policy
Tel: 202–797–6060 • Fax: 202–797–2480
www.brookings.edu/brown.aspx