The 2006 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With special sections on the nation's achievement, the happiness factor in learning, and honesty in state test scores

The Brookings Institution

ABOUT BROOKINGS

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.

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/browncenter.

To order copies of this report, please call 800-537-5487, fax 410-516-6998, e-mail hfcustserv@press.jhu.edu, or visit online at www.brookings.edu.

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

The 2006 Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

With special sections on the nation's achievement, the happiness factor in learning, and honesty in state test scores

October 2006
Volume II, Number 1

by:
TOM LOVELESS
Director, Brown Center on Education Policy

TABLE OF CONTENTS

Research assistance by:

KATHARYN FIELD
Brown Center on Education Policy

THE 2006 BROWN CENTER REPORT ON AMERICAN EDUCATION

This report launches the second volume of the Brown Center Report on American Education. The five issues of volume one were published from 2000 to 2004. Volume one included regular reports on data from the National Assessment of Educational Progress (NAEP) and state assessments, analysis of student achievement in charter schools, a study of trends in homework, evaluations of the federal government's Blue Ribbon Schools Program, an investigation of the academic performance of high schools with powerhouse sports teams, analyses of student achievement in urban school districts and rural schools, a survey of exchange students from countries abroad to see what they think of American schools, and a survey of the mathematics preparation of middle school math teachers. Volume two will explore similar topics related to how well American students are learning.

As in volume one, the reports of volume two will be divided into three parts. Part one reviews the latest data on student learning in the U.S. In this issue, the most recent results from both NAEP tests, the long term trend and the main, receive attention for what they reveal about American students' progress in reading and mathematics. Part two looks at the "happiness factor" in education, analyzing international data to see whether students' self-confidence and enjoyment of math and the relevance of lessons that students experience in classrooms are correlated with higher math achievement. Do nations with happier students score higher on math tests than nations in which students are not quite as happy?

Part three looks at how states have responded to the No Child Left Behind Act. Several analysts have recently concluded that states are "racing to the bottom" by artificially inflating the number of students who demonstrate proficiency on state tests. It is indisputable that states report larger numbers of proficient students than the NAEP test. But the studies have overlooked some key questions. Is NAEP such a good test that it should be used as a benchmark for judging state assessments? Can NCLB be blamed for the discrepancies between reported levels of proficiency on NAEP and state tests? How large were the discrepancies before NCLB?

We hope readers will consider this issue of the Brown Center Report as interesting and provocative as previous ones. On a personal note, this also is the first issue that will be published without Paul DiPerna on hand for the release. After six and a half years in the Brown Center, Paul has moved on to bigger and better things at the Friedman Foundation in Indianapolis. We thank Paul for his tireless work on behalf of the Center and wish him a terrific career in Indianapolis.

Part I THE NATION'S ACHIEVEMENT

WE ARE WADING IN TEST SCORES. All fifty states conduct annual assessments. The National Assessment of Educational Progress (NAEP), often called "the nation's report card," periodically releases scores from two different tests—the long- term trend NAEP and the main NAEP. The fact that both tests carry the NAEP label but may conflict in how they depict student achievement confuses many experts in the field of education, not to mention the general public. The results from the two tests do not always agree. The tests measure substantively different skills and knowledge, are taken by different samples of students, and are calibrated on different scales.

This section of the Brown Center Report sorts through the latest NAEP data in mathematics, reading, and science, identifies dominant trends, and pinpoints areas where the data are muddled.

Mathematics Achievement

Let's begin with mathematics and the scores released on the long-term trend NAEP in 2004. The results are displayed in Table 1-1. Changes in scores since 1990 are presented in scale score points and in years of learning. Scale scores are appropriate for comparing scores on a single test. In addition to measuring different mathematical content, however, the two NAEPs are calibrated on different scales. And they test different groups of students—ages 9, 13, and 17 on the trend NAEP and grades 4, 8, and 12 on the main NAEP. The estimate of changes in years of learning provides a better way of comparing results from the two tests. It offers a real-world, ballpark estimate of how much the average student has learned in terms of a typical year's worth of learning. One tenth of a year (0.1) is equivalent to about one month's worth of learning. This statistic cannot be computed for the oldest cohorts, seventeen year olds on the trend NAEP and twelfth graders on the main NAEP, because

A tale of two NAEPS

Nine year olds showed significant gains in math on the long-term trend NAEP.

Table 1-1

	1990	1999	2004	1990-2004 Change	Change in Years of Learning
Age 9	230	232	241	+11	1.1
Age 13	270	276	281	+11	1.3
Age 17	305	308	307	+2	

Years of learning based on 1990 score differences. Age 9: 1 year equals 1/4th the difference between ages 9 and 13 (10 scale score points). Age 13: 1 year equals 1/4th the difference between ages 13 and 17 (8.75 scale score points). Not applicable for age 17.

Source: http://nces.ed.gov/nationsreportcard/ltt/results2004/nat-math-scalescore.asp
Standard deviations in 1990 were Age 9, 33 points; Age 13, 31 points; Age 17, 31 points

A tale of two NAEPS

But the gains for similarly aged students on the main NAEP were twice as big, an historically unprecedented leap.

Table 1-2

	1990	1992	1996	2000	2003	2005	1990-2005 Change	Change in Years of Learning
Grade 4	213	220	224	226	235	238	+25	2.0
Grade 8	263	268	270	273	278	279	+16	2.1
Grade 12	294	299	304	301	—	—		

Years of learning based on 1990 score differences. Grade 4: 1 year equals 1/4th the difference between 4th and 8th grades (12.5 scale score points). Grade 8: 1 year equals 1/4th the difference between 8th and 12th grades (7.75 scale score points).

Source: NAEP data explorer, http://nces.ed.gov/nationsreportcard/nde/
Standard Deviations in 1990 were: Grade 4, 32 points; Grade 8, 36 points; Grade 12, 36 points

an interval with an older group is needed to compute the statistic.

In 2004, large gains were reported for nine and thirteen year olds in mathematics, up significantly from 1999. Scores for seventeen year olds were not as impressive. From 1990 to 2004, nine and thirteen year olds gained 11 scale score points. The scores of seventeen year olds only increased by 2 points. Since 1990, the two younger groups of students have gained a little more than one year's worth of learning.

The gains of younger students are much larger on the main NAEP. The latest scores, released in 2005, show fourth graders gaining 25 scale score points and eighth graders gaining 16 points from 1990 to 2005 (see Table 1-2). This is equal to about two years of learning, more than twice the gains on the long-term trend NAEP. Fourth graders and nine year olds constitute approximately the same group of youngsters, so the differences between the two NAEPs deserve an explanation. The most likely explanation, investigated in previous Brown Center Reports, rests on differences in mathematical content.[1] The main NAEP measures student learning in mathematics as pushed by reformers in the 1990s. Students are allowed to use calculators on a portion of the test and items involving data analysis, problem solving, and geometry are prevalent. The trend NAEP, on the other hand, contains a larger number of traditional arithmetic items, requires more sophisticated computation skills, and does not allow calculators.

The two NAEPs agree that students are gaining in mathematics but disagree about the size of the increase.

The point is not that the main NAEP has been changed over time to inflate gains. The point is that students have likely become more proficient using calculators and have made greater strides learning data analysis, problem solving, and geometry, as reflected in the gains on the main NAEP, than they have at learning arithmetic and how to compute, the gains reflected on the trend NAEP.

The gains by fourth graders on the main NAEP are historically unprecedented.[2] Notice that the 25 point scale score gain that fourth graders notched on the main NAEP from 1990 to 2005 closed half of the distance that existed between fourth and eighth graders in 1990. If the same rate continues into the future, fourth graders will catch up with the 1990 eighth graders' score in 2020. And sometime around 2039, they will score 295 and pass the score of 1990's twelfth graders. Taken at face value, this means the average fourth graders in 2039 will know as much mathematics as the typical high school senior knew in 1990!

Dominant trend #1 from NAEP, then, is the following: the two NAEPs agree that students are gaining in mathematics but disagree about the size of the increase. The long-term trend NAEP shows modest gains; the main NAEP is reporting gains that are extremely large—perhaps unbelievably so at fourth grade. The discrepancies between the two tests narrowed in 2004. The gains reported in 2004 on the long-term trend may mean that the trend NAEP scores are finally catching up with the huge gains reported on the main NAEP in the 1990s.

Reading Achievement

Gains in NAEP reading scores lag those in mathematics. This is true on both NAEP tests, but especially true on the main NAEP. As displayed in Table 1-3, the long-term trend is showing a little less than one year's gain (0.8) in learning for nine year olds since 1990. Seven out of the 10 points in the gain came in the 1999 to 2004 period. At age 13, scores have barely changed since 1990, up only 2 scale score points—or about 2 months. The scores of seventeen year olds declined 5 scale score points from 1990 to 2004.

The main NAEP confirms that only meager gains have occurred in reading (see Table 1-4). Why has reading achievement remained flat for more than a decade? *Education Week* recently reported an analysis by Marshall Smith that argued increased immigration of non-English speaking youngsters in the 1990s depressed reading scores.[3] Because immigrant children score well below the national average, and their share of the student population increased, immigration did indeed depress national scores. But not by much. Consider that scores of non-Hispanic white fourth graders—a group with very few immigrants—only increased 5 points from 1992 to 2005 and eighth graders by 4 points, still not a large gain. Indeed, if the 2005 NAEP scores are re-computed using the proportion of blacks, whites, and Hispanics in 1992 (in essence, controlling for proportional changes since then), the gain for fourth graders becomes 6 points and for eighth graders slightly more than 4 points. Still nothing to shout about.[4]

Another explanation for flat reading scores concerns the gender gap in reading. Girls read more than boys, and they tend to score higher on reading tests. The *New Republic* and *Newsweek* are among the national publications that recently have featured articles on boys' poor showing on NAEP reading tests.[5] The gender gap in reading is not new. Moreover, the gender gap actually contracted from 1992 to 2005 in fourth and eighth grades as girls' scores

Table 1-3

Reading scores lag behind math on the trend NAEP.

Gains for 7 and 13 year olds are unimpressive, and scores for 17 year olds have actually declined.

	1990	1999	2004	1990-2004 Change	Change in Years of Learning
Age 9	209	212	219	+10	0.8
Age 13	257	259	259	+2	0.2
Age 17	290	288	285	-5	

Years of learning based on 1990 score differences. Age 9: 1 year equals 1/4th the difference between ages 9 and 13 (12 scale score points). Age 13: 1 year equals 1/4th the difference between ages 13 and 17 (8.25 scale score points). Not applicable for age 17.

Source: http://nces.ed.gov/nationsreportcard/ltt/results2004/nat-reading-scalescore.asp
Standard deviations in 1990 were Age 9, 45 points; Age 13, 36 points; Age 17, 41 points

Table 1-4

The main NAEP confirms only meager gains in reading.

The reason for no progress remains a mystery.

	1992	1994	1998	2000	2002	2003	2005	1992-2005 Change	Change in Years of Learning
Grade 4	217	214	215	213	219	218	219	+2	0.2
Grade 8	260	260	264	—	264	263	262	+2	0.3
Grade 12	292	287	291	—	287	—	—		

Years of learning based on 1990 score differences. Grade 4: 1 year equals 1/4th the difference between 4th and 8th grades (10.75 scale score points). Grade 8: 1 year equals 1/4th the difference between 8th and 12th grades (8.0 scale score points).

NOTE: Beginning with 2002 scores reported here, results include students who required special accommodations to take the NAEP. Standard Deviations in 1992 were: Grade 4, 36 points; Grade 8, 36 points; Grade 12, 33 points

Source: NAEP data explorer, http://nces.ed.gov/nationsreportcard/nde/

Introduction to NAEP

The National Assessment of Educational Progress (NAEP) is commonly referred to as the Nation's Report Card. Since 1969, it has been the only nationally representative and continuing assessment of what America's students know and can do in academic subject areas.

There are two NAEP test types: (1) the main NAEP gauges national and state achievement while also reflecting current practices in curriculum and assessment, and (2) the long-term trend NAEP allows reliable measurement of change in national achievement over time. These assessments use distinct data collection procedures and separate samples of students.

Since 1990, the math test on the main NAEP has been governed by a framework reflecting recommendations of the National Council of Teachers of Mathematics (NCTM). Beginning with the 2002 assessments, the number of students selected for a NAEP national sample for any particular grade and subject has been 150,000 or more.

Faulty instruction may be the explanation for why American students' reading skills have stagnated, but at this point it is merely one possible explanation.

stayed flat and boys gained a few points. The gender gap does not explain the lack of progress in reading, either.[6]

A third explanation rests on how reading is taught. In 2000, the National Reading Panel (NRP) released a report reviewing the scientific evidence on reading instruction. Among the panel's conclusions were that phonemic awareness is key to teaching young children how to read and that systematic phonics instruction is needed to help youngsters with reading difficulties. In addition, the panel endorsed explicit instruction on fluency, vocabulary, and comprehension.[7] The panel's findings were codified in the Reading First program, which is part of No Child Left Behind (NCLB), and dozens of states followed with legislation of their own supporting Reading First's objectives. For many observers, the federal and state policies emanating from the NRP report struck a decisive blow in the decades-old reading wars between phonics and whole language advocates. Phonics emerged victorious. In fact, whole language had been losing favor among policymakers before the NRP report. More than one hundred bills on the teaching of reading were introduced in state legislatures in the 1990s, and they tended to favor phonics.[8]

On the other hand, we also know that whole language has reappeared under the banner of "balanced literacy" in many major districts, including San Diego and New York.[9] Further, a 2006 survey of education schools found that only 11 of 77 provide "minimal exposure" to the scientific literature on reading, suggesting that teachers aren't learning about phonics in their training programs.[10] Faulty instruction may be the explanation for why American students' reading skills have stagnated, but at this point it is merely one possible explanation.

Dominant trend #2: NAEP tests report little or no progress in reading achievement since 1990. Why reading scores are flat remains a mystery.

Science Achievement

Science is not tested as often as math and reading. In 2007-08, that will change as science testing by the states becomes a requirement of NCLB. On the national level, the main NAEP science test was given in 2005 and will be given again in 2009 and every four years thereafter. The first science test on the long-term trend NAEP was administered in 1969, and it was last given in 1999. The next administration is not currently scheduled.

From 1990-1999, science scores on the long-term trend were flat for nine and thirteen year olds (see Table 1-5). At age 17, scores rose from 290 to 295, a gain that is statistically significant. On the main NAEP, fourth graders made a 4-point gain, and eighth graders' test scores remained unchanged (see Table 1-6). The scores of twelfth graders' declined by 3 points, and in sharp contrast to the gains on the long-term trend, this decline was statistically significant. The long-term trend and the main NAEP diverge sharply in measuring science achievement for the oldest group of students.

Despite the gains accomplished by seventeen year olds in the 1990s, the long-term trend scores for them are down sharply since 1969. That year, seventeen year olds registered a 305 scale score, 10 points higher than in 1999.[11] The science scores on NAEP illustrate another pattern that is evident across both NAEP tests and all three subjects. Over time, the oldest students—twelfth graders on the main NAEP and seventeen year olds on the long-term trend NAEP—perform the worst of the three age /grade groups and have made little progress or even registered declines. Gains that are accomplished by younger students shrink by the time a cohort of students reaches the final year of high school. What is going on?

The prevailing explanation is that high school students do not take the NAEP seriously. By the senior year, students are test weary and results from NAEP mean nothing to individual students. Passing high school exit exams is the concern of low-achieving seniors; doing well on AP tests is the concern of high achievers. Performance on NAEP carries no consequences. Although evidence is inconclusive, an analysis of off-task test behavior on NAEP tests—leaving items blank or using the test sheet for art-work—in 1996 revealed that 25% to 30% of twelfth graders exhibited such behavior (versus 13% of eighth graders and 6% of fourth graders).[12]

The NAEP sampling design selects schools randomly and then samples students randomly within schools. Participation by selected schools and students dropped precipitously in 2002. Only about 55% of the

Table 1-5

On the trend NAEP, science scores were mostly flat in the 1990s.

Scores for 9 and 13 year olds barely budged.

	1990	1996	1999	1990-1999 Change	Change in Years of Learning
Age 9	229	230	229	0	0
Age 13	255	256	256	+1	0.1
Age 17	290	296	295	+5	

Years of learning based on 1990 score differences. Age 9: 1 year equals 1/4th the difference between ages 9 and 13 (6.5 scale score points). Age 13: 1 year equals 1/4th the difference between ages 13 and 17 (8.75 scale score points). Not applicable for age 17.

Source: http://nces.ed.gov/nationsreportcard/ltt/science_1999_results.asp
Standard deviations in 1990 were Age 9, 40 points; Age 13, 38 points; Age 17, 46 points

Table 1-6

The main NAEP confirms this trend.

Scores for each grade have hardly changed.

	1996	2000	2005	1996-2005 Change
Grade 4	147	147	151	+4
Grade 8	149	149	149	0
Grade 12	150	146	147	-3

Source: http://nationsreportcard.gov/science_2005/s0102.asp?printver=
Standard deviations in 1996 were Grade 4, 35 points; Grade 8, 35 points; Grade 12, 33 points

NOTE: Years of learning were not computed because scales are not comparable across the grades.

A content analysis of the two NAEP science tests is an important topic for future research.

combined school/student sample that was originally selected in the twelfth grade agreed to participate, down from about 65% in previous years. The eighth grade rate was 76% and fourth grade, 80%.[13]

A 2004 national commission on the future of the twelfth grade NAEP test recommended several incentives to motivate students and schools, among them: college scholarships for randomly selected students (two per state), redeemable certificates for merchandise given to all participating schools and students, and letters of recognition from the President of the United States to participating students and the principals of their schools. The scholarships and certificates would only be available to students exhibiting minimal off-task behaviors on the NAEP. These strategies need to be experimented with to see if they are effective, but by raising student motivation to do well on NAEP, the nation would receive a more accurate indication of what students know in the final year of high school.

Conclusion

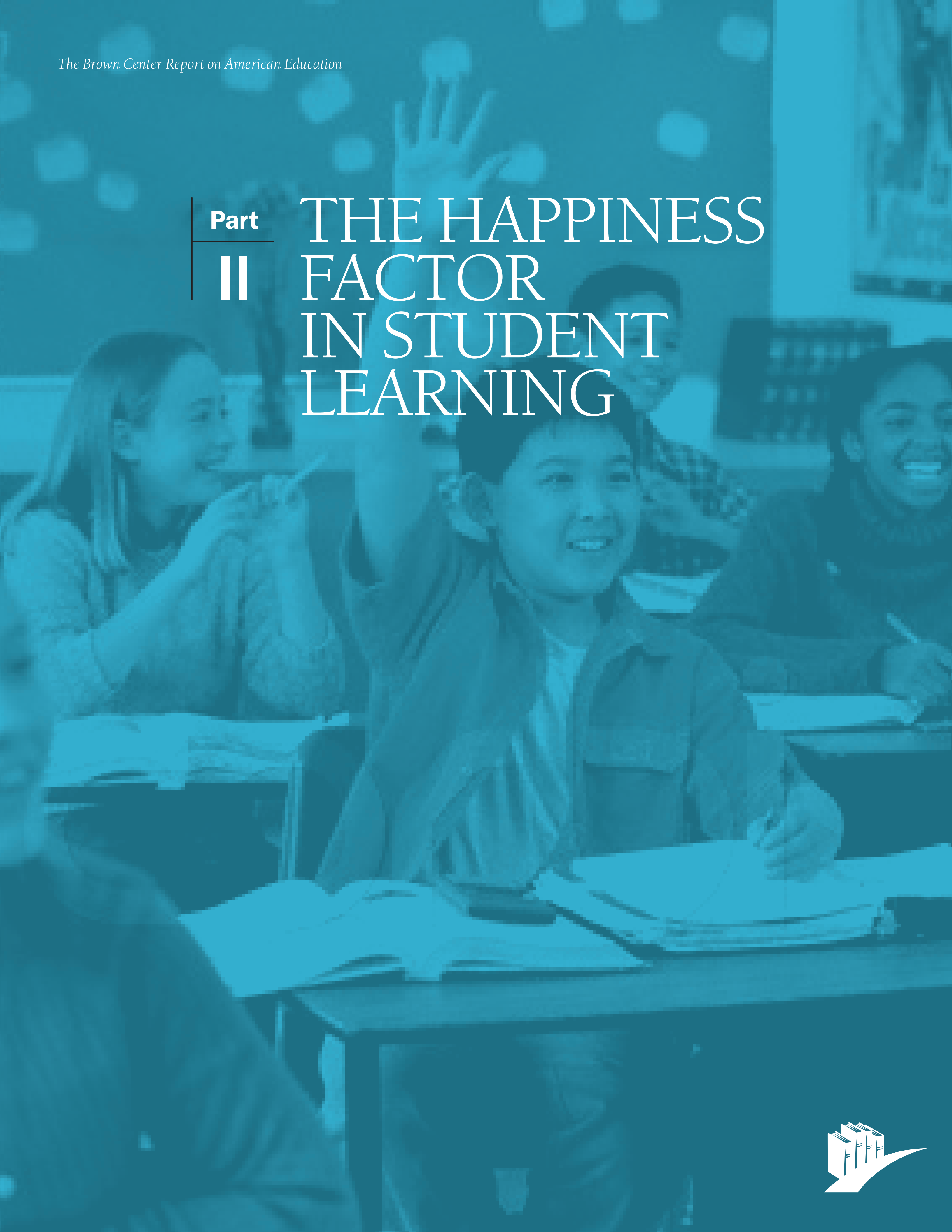
This section of the Brown Center Report has reviewed the latest data on student achievement in math, reading, and science. The main conclusions are:

1. American students are gaining in math achievement, but the two NAEP tests disagree about the magnitude of the gain, especially with younger students. As reflected on the main NAEP, students appear to be making great strides in data analysis, problem solving, and geometry—and in solving problems using a calculator. As reflected on the long-term trend NAEP, gains are smaller in arithmetic, computation skills, and solving problems without a calculator.

2. Reading scores have lagged math scores since 1990. In 2004, the long-term trend NAEP showed a promising increase of 7 points for nine year olds. But all other age and grade groups on both NAEPs have made either meager gains, or, in the case of seventeen year olds and twelfth graders, exhibited test score declines. Prevailing explanations for the poor performance on reading are speculative and lack supporting evidence.

3.Science scores are mostly flat since 1990. Scores on the main and long-term trend NAEPs diverge for the oldest cohort. The long-term trend NAEP reports significant progress for seventeen year olds (a gain of 5 scale score points) but the main NAEP shows a decline. As with the divergences in math, content differences between the two NAEP tests may be the reason for the different results. A content analysis of the two NAEP science tests is an important topic for future research.

4. The poor performance of high school students on NAEP may be an artifact of off-task test behavior and the lack of motivation to do well on the test. A national commission has recommended the introduction of incentives for students and schools to boost participation rates and effort on the test. It is crucial to assess accurately whether twelfth graders are leaving high school with the skills and knowledge they need for future success.



Part II

THE HAPPINESS FACTOR IN STUDENT LEARNING

IT IS ONLY NATURAL THAT ADULTS WANT CHILDREN TO BE HAPPY. Indeed, many of the most popular education reforms of today, once all of the rhetorical flourishes are stripped away, place children's happiness on equal footing with their learning. The pursuit of knowledge may be important, but only if it simultaneously raises student contentment and self-esteem. Bill Gates wants high schools to be more relevant to kids' lives. Oprah Winfrey features a high school backed by the Gates Foundation in which there are no books and no lectures, and students work on projects reflecting their individual interests. The National Council of Teachers of Mathematics urges the development of mathematical power, which, in addition to knowing mathematics, includes "the development of personal self-confidence." A spring 2006 study highlights the national problem of high school dropouts and suggests boring school work drives students out of school. If students were just more confident in their abilities, enjoyed the subject matter more, and were convinced of the relevance of schoolwork to everyday life—so the story goes—American schools would flourish.[14]

Call this the happiness factor in American education. The following study investigates three key components: students' self-confidence, their joy of learning, and the relevance of what is being learned. Most people do not need evidence proving that the happiness factor is essential. Common sense dictates that joyous, confident kids studying a curriculum relevant to daily life will learn more than children without the benefit of such positive experiences. The attractiveness of this idea explains why the urge to make education a more pleasant experience—and schools happier places—has a long history.

Educational progressives made happiness a central theme of the "child-centered" practices advocated in the early twentieth century. Boredom was targeted as particularly evil. Reformers argued that subject matter should correspond to students' interests, not to ancient disciplinary standards or intellectual merits. Book learning, subject matter knowledge, and learning for learning's sake were eschewed in deference to activity-based learning, learning "how to learn," and learning for self-awareness and personal growth. These principles remain paramount among many school reformers today. When Bill Gates tours the "no books, no lectures" school with Oprah Winfrey and declares that projects are "the way to go," he is echoing the sentiments of one of the earliest advocates of progressive education, William Heard Kilpatrick, whose essay, "The Project Method," was published in 1918.[15]

No one advocates that schools should bore children to death or purposely make them unhappy. But what about other countries? Does the happiness factor hold the same sway as it does in the United States? Is the happiness factor related to achievement?

Data produced by TIMSS, an international assessment of student achievement, offer some interesting empirical evidence on these questions. Two conclusions stand out. Countries vary on indices of happiness, and in many other countries, the happiness of children seems to take a back seat to learning. The second conclusion is surprising. National indices of student happiness are inversely related to achievement in mathematics. That is, countries with more confident students who enjoy the subject matter—and with teachers who strive to make mathematics relevant to students' daily lives—do not do as well as countries that rank lower on indices of confidence, enjoyment, and relevance.

The argument here is not that student happiness causes low achievement. Correlations do not prove causality. But school reformers should take note. When thinking about how schools can be improved, the intuitive attractiveness of the idea that making students happier results in better education should be held in abeyance. Happiness is not everything, and by simply producing contented students, good results do not automatically follow. It is interesting that people grasp this notion in other areas of self-improvement—eating healthy foods, getting exercise, saving for retirement—but when it comes to education, for some reason, the limitations of happiness are forgotten.

Table 2-1

Confidence: "I usually do well in mathematics" (8th grade)
(countries ranked by percent agree a lot)

Country	Percent Students "Agree A Lot"	National Score
Jordan	48	424
Egypt	46	406
Israel	43	496 ✔
Ghana	41	276
Bahrain	40	401
Tunisia	39	410
Cyprus	39	459
Palestinian Authority	39	390
United States	39	504 ✔
South Africa	38	264
International Average	27	467
Romania	18	475 ✔
Singapore	18	605 ✔
Latvia	17	508 ✔
Moldova	17	460
Netherlands	16	536 ✔
Malaysia	13	508 ✔
Chinese Taipei	11	585 ✔
Hong Kong	10	586 ✔
Korea	6	589 ✔
Japan	4	570 ✔

(✔ – above international average)

Source: TIMSS 2003 Userguide (see almanacs, bsalm1_m3.pdf, pg.67): http://timss.bc.edu/timss2003i/userguide.html

Reported in the International Mathematics Report: this question is one of four questions constituting the student confidence index.

The following analysis examines TIMSS data on three topics in learning mathematics: students' self-confidence in learning math, how much students enjoy math, and whether teachers strive to make mathematics relevant to students during instruction. Eighth and fourth graders were tested in the 2003 TIMSS, and both students and teachers were surveyed as part of the assessment. Data were collected from 46 nations at eighth grade and 25 nations at fourth grade.[16]

The bottom ten countries in self-confidence include many of the world's highest-achieving nations.

Confidence

One key to understanding the American emphasis on happiness is a deep cultural belief in self-confidence. Some people think that confidence leads to better performance. Others believe that confidence in one's abilities comes from high performance but does not cause it. Recent studies have discovered an interesting paradox involving ethnicity, self-confidence, and academic achievement. White and Asian students tend to score higher on tests of academic achievement but feel badly about their performance. Black and Hispanic youngsters, on the other hand, score lower on achievement tests but feel more confident about their performance.[17] Although a vast research literature has failed to determine whether self-confidence is a product of—rather than a precursor to—academic achievement, instilling in students the belief that they are good at math is an article of faith for many educators.[18] TIMSS data cannot resolve the debate over causality, but they can shed light on correlation. Do those students who believe they are good at math actually know more math?

In the TIMSS data, when one looks at the math scores of students within each country, those who express confidence in their own math abilities do indeed score higher than those lacking in confidence. That is true for 40 of the 46 countries with eighth grade test results.[19] But when comparing national means, the story gets more complicated, and a curious pattern emerges.

Table 2-1 shows the ten nations with the most and least confident students as measured by student responses to the statement, "I usually do well in mathematics." The table shows the percentage of eighth graders who answered that they agreed "a lot" with that statement.

Nations vary a great deal in student self-confidence. In Jordan, 48% of students say they usually do well in math. In Japan, the figure is 4%. The world's most confident eighth graders are found in the Middle East, Africa, and the United States (ranked ninth). More than 40% of the students in Jordan, Egypt, Israel, and Ghana think that they usually do well in math; however, these countries are not particularly high scoring on the TIMSS math exam. Of the ten countries registering the highest levels of student confidence, only two (Israel, 496, and the United States, 504) score higher than the international average of 467.

The bottom ten countries in self-confidence, on the other hand, include many of the world's highest-achieving nations. Five Asian nations—Singapore, Chinese Taipei, Hong Kong, Korea, and Japan—consistently lead the world in math achievement, but students in these countries do not believe that they do very well in math. A few European nations join Asian nations at the bottom of

Table **2-2**

Confidence: "I usually do well in mathematics" (4th grade)
(countries ranked by percent agree a lot)

Country	Percent Students "Agree A Lot"	National Score
Cyprus	**61**	**510** ✔
Tunisia	**58**	**339**
Iran	**56**	**389**
Slovenia	**53**	**479**
Morocco	**49**	**347**
International Average	**37**	**495**
Moldova	**25**	**504** ✔
Netherlands	**24**	**540** ✔
Belgium-Flemish	**23**	**551** ✔
Hong Kong	**16**	**575** ✔
Japan	**10**	**570** ✔

(✔ – above international average)

United States: 46% agree a lot, national mean score 518

Source: TIMSS 2003 Userguide (see almanacs, asalm1_m3.pdf, pg.30): http://timss.bc.edu/timss2003i/userguide.html the confidence rankings. But the trend extends beyond merely a few countries in Asia and Europe. Overall, the correlation coefficient for national math achievement and percentage of highly confident students is -0.64, indicating that across all 46 nations the two phenomena are inversely related.

Table 2-2 displays the data for fourth grade. Only the top and bottom five nations are shown because of fewer nations reporting data on this item at fourth grade. The pattern is the same as that for eighth grade, with high confidence nations performing poorly and low confidence nations performing well. A few new countries join the list. Iranian fourth graders express great confidence in their math abilities, with 56% saying they usually do well in mathematics, but Iran's score (389) is well below the international average. Only 23% of Belgium-Flemish fourth graders, a population that consistently scores well on TIMSS, feel that they usually do well in math. For all nations, the correlation coefficient for fourth grade national achievement and student confidence is -0.58.

So an interesting paradox emerges from the international data on student confidence and achievement. The relationships are the opposite depending on whether within-nation or between-nation data are examined. Within nations, high confidence is positively associated with achievement. More confident students have higher test scores. But between nations, the relationship is reversed. Nations with highly confident students have lower average test scores than nations with less confident students.

Enjoyment

The same paradox appears when exploring the relationship between achievement and enjoying mathematics. Again, within nations the pattern is as one would expect. In 35 of 46 countries, eighth graders who enjoy math score higher than those who do not enjoy the subject. The between-nation pattern is the opposite. Table 2-3 exhibits the national comparisons. Enjoyment is measured by the percentage of students agreeing "a lot" with the statement, "I enjoy mathematics."[20]

All ten of the top nations in enjoyment score below the international average on TIMSS. The ten bottom nations in enjoyment all excel in math achievement. U.S. eighth graders enjoy math a little less than the international average, U.S. fourth graders

Enjoyment: "I enjoy mathematics" (8th grade)
(countries ranked by percent agree a lot)

Table **2-3**

Country	Percent Students "Agree A Lot"	National Score
Botswana	65	366
Egypt	61	406
Morocco	60	387
Iran	58	411
South Africa	56	264
Ghana	53	276
Jordan	50	424
Lebanon	50	433
Tunisia	45	410
Palestinian Authority	45	390
International Average	29	467
Sweden	15	499 ✔
Latvia	14	508 ✔
England	14	498 ✔
Estonia	14	531 ✔
Slovak Republic	13	508 ✔
Chinese Taipei	13	585 ✔
Japan	9	570 ✔
Korea	9	589 ✔
Slovenia	7	493 ✔
Netherlands	6	536 ✔

(✔ – above international average)

United States: 22% agree a lot, national mean score 504

Source: TIMSS 2003 International Mathematics Report, ch. 4, pg. 159

Enjoyment: "I enjoy mathematics" (4th grade)
(countries ranked by percent agree a lot)

Table **2-4**

Country	Percent Students "Agree A Lot"	National Score
Iran	81	389
Armenia	71	456
Morocco	71	347
Tunisia	70	339
Lithuania	58	534 ✔
International Average	50	495
Chinese Taipei	31	564 ✔
Netherlands	30	540 ✔
Hong Kong	30	575 ✔
Japan	29	565 ✔
Belgium-Flemish	27	551 ✔

(✔ – above international average)

United States: 54% agree a lot, national mean score 518

Source: TIMSS 2003 International Mathematics Report, ch. 4, pg. 160

Relevance: Teachers reporting students "relate what is being learned in mathematics to daily lives" (8th grade)
(countries ranked by percent)

Table **2-5**

Country	Percent	National Score
Chile	87	387
Jordan	72	424
Palestinian Authority	72	390
Botswana	71	366
Iran	69	411
United States	66	504 ✔
Slovak Republic	66	508 ✔
Malaysia	64	508 ✔
Egypt	63	406
Ghana	63	276
International Average	50	467
Tunisia	38	410
Belgium–Flemish	32	537 ✔
Singapore	32	605 ✔
Italy	31	484 ✔
Chinese Taipei	27	585 ✔
Netherlands	26	536 ✔
Bulgaria	26	476 ✔
Hong Kong	26	586 ✔
Russian Federation	24	508 ✔
Japan	14	570 ✔

(✔ – above international average)

NOTE: "Percent" refers to percent of students whose teachers reported students doing the activity about half of the lessons or more

Source: TIMSS 2003 International Mathematics Report, ch. 7, pg. 284

a little more than average. The correlation coefficient for national enjoyment and achievement is -0.75. Table 2-4 shows the statistics for fourth grade, and the same negative relationship holds. The correlation coefficient at fourth grade is -0.67. The more math a nation's children know, the less likely they are to enjoy mathematics.

Relevance

To assess the relevance of math lessons, teachers were asked whether they "relate what is being learned in mathematics to students' daily lives" in half of their lessons or more. Making math relevant to students is advocated as an instructional strategy that will make students appreciate the usefulness of mathematics. Table 2-5 displays the data for eighth grade. The question was not asked at fourth grade.[21]

An emphasis on making math relevant to students varies tremendously by country, with Chile at the top (87%) and Japan (14%) at the bottom of the rankings. The United States ranks sixth in making math lessons relevant. The relationship of relevance to national achievement is negative. The more relevant the math, the lower scoring the nation. Among the ten countries ranked at the top on relevance, only three score above average on math achievement. Among the ten countries that have the least relevant math lessons, all but one score above average on math achievement. The correlation coefficient is -0.52. Nations that strive to make math more relevant to their students do not do as well on tests of math achievement as nations for which relevance is less important.[22]

Conclusion

What should we conclude from these rather surprising statistics (see Table 2-6 for a summary)? First, a caveat. Nothing can be concluded about causality from the data, only that national measures of self-confidence, enjoyment of the subject, and relevance of lessons are inversely correlated with student achievement. The evidence presented here does not mean that we should undermine students' confidence, teach math in a way sure to induce revulsion to the subject, or present math in such an abstract manner that it bears no relevance to daily life.

The evidence does suggest, however, that the American infatuation with the happiness factor in education may be misplaced. The international evidence makes at least a prima facie case that self-confidence, liking the subject, and relevance are not essential for mastering mathematics at high levels.

The data also show that the relationship between achievement and happiness is not a simple one. When within-nation data

Are confidence, enjoyment, and relevance related to a nation's achievement?

Table 2-6

	Grade	Correlation Coefficients
Confidence	4	-.58
	8	-.64
Enjoyment	4	-.67
	8	-.75
Relevance	8	-.52

Note: Pearson Correlation Coefficient representing the relationship between each nation's average score and percent answering in each question's "high" category. See box for explanation.

Source: Computed by author using data from the TIMSS 2003 International Mathematics Report

American students are much more confident about their math abilities than Singaporean students.

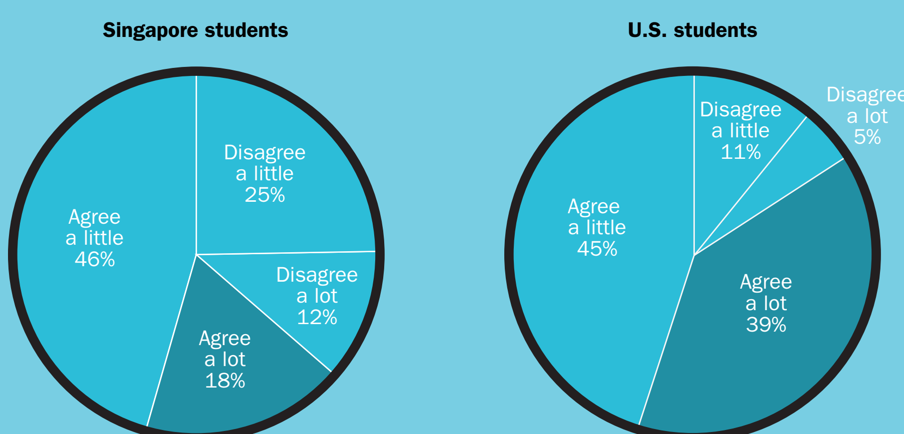
Fig 2-1

Students were asked whether they agreed with the statement, "I usually do well in mathematics."

39% of American students agreed a lot.

NOTE: Data refer only to 8th grade.

Source: TIMSS 2003 Userguide (see almanacs, bsalm1_m3.pdf, pg. 67): http://timss.bc.edu/timss2003i/userguide.html



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.

are examined, math achievement is positively correlated with enjoying the subject and confidence in one's math abilities. But the relationship reverses if between-nation data are the focus. The importance of national culture emerges as perhaps the most important contextual factor. Students' perceptions of their performance are shaped by their relative standing within a comparison group.

Singapore is the top scoring nation on TIMSS. In the United States, 39% of eighth graders agree a lot with the statement, "I usually do well in mathematics." In Singapore, the figure is 18% (see Figure 2-1). In both countries, students who say they do well in math score higher on the TIMSS test than students who disagree with such a rosy self-assessment, with about 100 points separating the most and least confident students (see Figure 2-2). However, doing well in Singapore seems to mean something quite different than in the United States. The most confident student in the U.S. (with an average score of 541) still scores below the least confident student in Singapore (average of 551). Although the U.S. has twice as many eighth graders who possess great confidence in their math abilities, the average score of these supremely confident Americans falls more than 60 points shy of the average score of 605 for all Singaporean eighth graders.

So perspective is crucial. Researchers call this a "frog pond effect." A medium-sized frog can feel awfully small in a big pond or awfully large in a small pond. In a 1987 study, Herbert Marsh compared the self-esteem of students at about the same level of performance—they had similar test scores, about

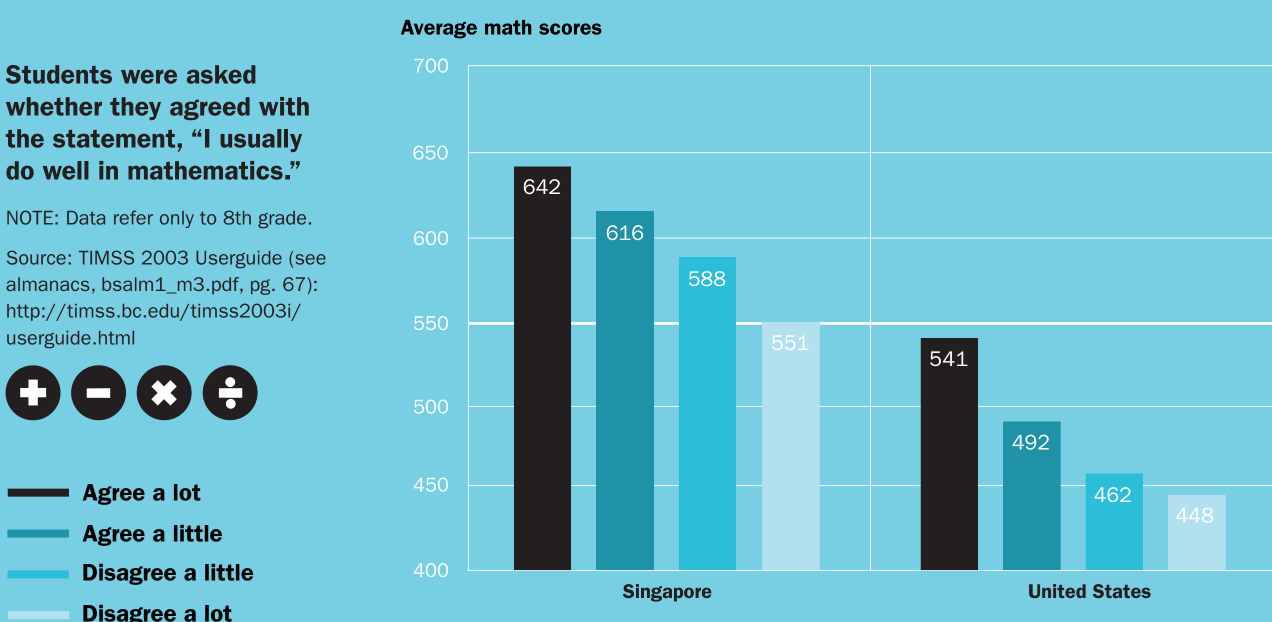
But even the least confident student in Singapore outscores the most confident American student!

Fig 2-2

Students were asked whether they agreed with the statement, "I usually do well in mathematics."

NOTE: Data refer only to 8th grade.

Source: TIMSS 2003 Userguide (see almanacs, bsalm1_m3.pdf, pg. 67): http://timss.bc.edu/timss2003i/userguide.html



Despite the call to make school more relevant, there is little evidence that relevance increases student engagement.

average—but who were attending schools with either predominantly low or high achieving populations of students. Students in the low-achieving schools consistently rated their own performance as higher than the students in the high-achieving schools. The average student in a school filled with low achievers looks around and thinks, "Wow, I'm pretty good." In a school filled with high achievers, the "just average" student may judge his performance as inadequate.[23]

Highly confident American students are doing quite well in comparison to their U.S. counterparts. And they feel good about that. But if they moved to Singapore, those good feelings would surely dissipate.

The international data on the happiness factor offer practical lessons for the nation. Textbooks should be as challenging as the books used in the highest scoring nations in the world. American students need to become more aware of what other students around the world are capable of doing. For this to happen, state assessments and NAEP should reflect world standards, not just national standards. Currently, neither NAEP nor state tests meet this requirement. Truly matching the expectations of the highest achieving students in the world may lead to a more modest assessment of American academic achievement, but it will also offer a realistic picture of where the U.S. stands internationally in education.

A second lesson pertains to student engagement. Despite the call to make school more relevant, there is little evidence that relevance increases student engagement. Students who regularly attend school, complete assigned homework, apply themselves in the classroom, and maintain an academic focus throughout their school careers—in other words, are truly engaged—are more likely to succeed than students who fall short on one of these criteria. These students are also likely to enjoy school and to express confidence in their own abilities. But if enjoyment and self-confidence are divorced from an academic purpose—if they are pursued as ends unto themselves—they appear to lose their educational value. Real student engagement is not about keeping students happy, boosting their self-esteem, or convincing them that what they are learning is relevant; it's about acquiring new knowledge and skills and pursuing the activities that contribute to that attainment.

A final point about the relationship of the happiness factor to learning. The international evidence indicates that American kids score very well on measures of enjoyment and confidence. American teachers rank high on making mathematics relevant. But our test scores are mediocre. If we as a nation want to increase student achievement in mathematics, it will take more than the happiness factor to do it. The happiest, most confident, most soaked-in-relevance American eighth grader cannot compete with the average Singaporean eighth grader in math. Students do not only learn in school, but they also receive signals from peers and families and the broader culture that convey what we as a society believe to be valuable. Right now those signals are not indicating that learning mathematics is very important. And, as a nation, we have the test scores to prove it.

If we as a nation want to increase student achievement in mathematics, it will take more than the happiness factor to do it.

Part III

ARE STATES HONESTLY REPORTING TEST SCORES?

THE NO CHILD LEFT BEHIND ACT (NCLB) REQUIRES STATES TO test students annually in grades 3-8 and to report the percentage of students who have reached proficiency in reading and mathematics (and in science beginning in 2007) by subgroup. States are required to impose sanctions on schools that fail to make adequate yearly progress in boosting the percentage of students attaining proficiency, with 100% proficiency rates mandated by 2014. Notwithstanding the federal mandates, states have the power to establish standards, design tests, and set levels of proficiency for student performance.

No one likes sanctions, including the state officials who must impose them. Given how NCLB is designed, states have a golden opportunity to undermine the system and an incentive to do so. They can water down tests, set cut-points for proficiency at unreasonably low levels, or engage in other shenanigans to make schools appear more successful than they really are. In 2002, when states released the first NCLB lists of schools "in need of improvement," Arkansas and Wyoming reported no failing schools. None. In the whole state. In contrast, Michigan identified 43% of its schools as "in need of improvement."[24]

Several analysts have criticized the states' response to NCLB. On the same day that state NAEP results were released in 2005, the Fordham Foundation issued a press release asserting that states were in "a race to the bottom."[25] The claim was based on the observation that gains made by states on NAEP from 2003 to 2005 fell short of the gains that states reported on their own tests. A 2006 report by Kevin Carey of Education Sector analyzed a more comprehensive set of data that states provide to the federal government. The report's conclusion is revealed in its title, "Hot Air: How States Inflate Their Educational Progress Under NCLB."[26]

States report higher proficiency rates in 8th grade math.

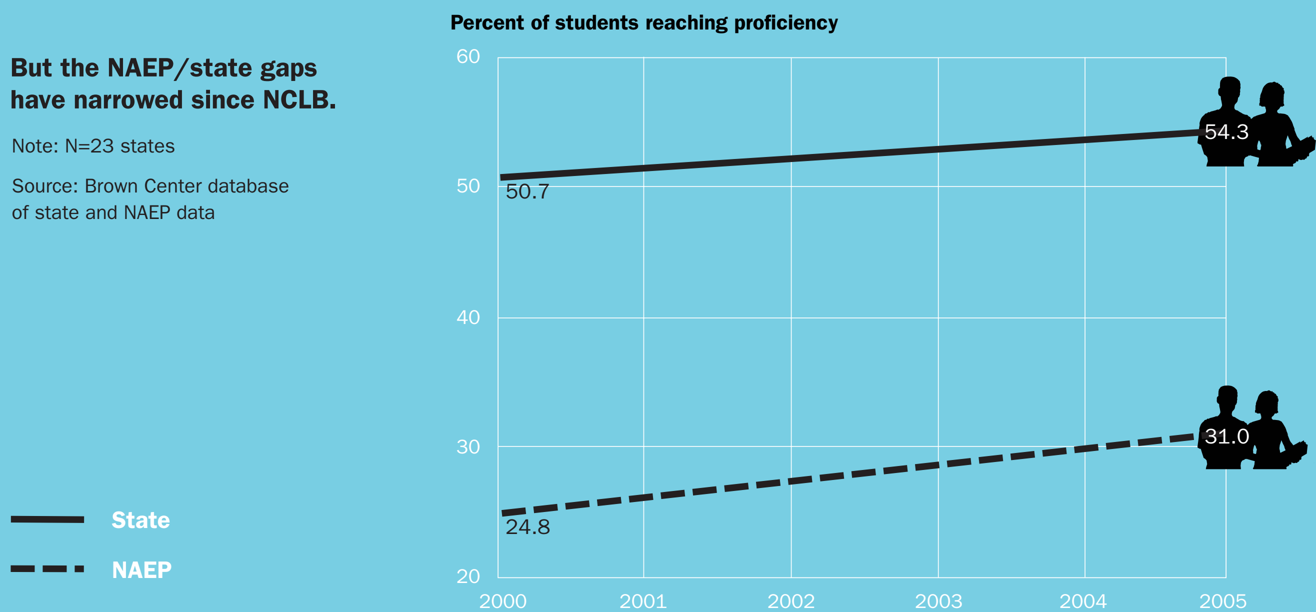
Comparison of state and NAEP proficiency

Fig **3-1**

But the NAEP/state gaps have narrowed since NCLB.

Note: N=23 states

Source: Brown Center database of state and NAEP data



The states report large gains in 4th grade math.

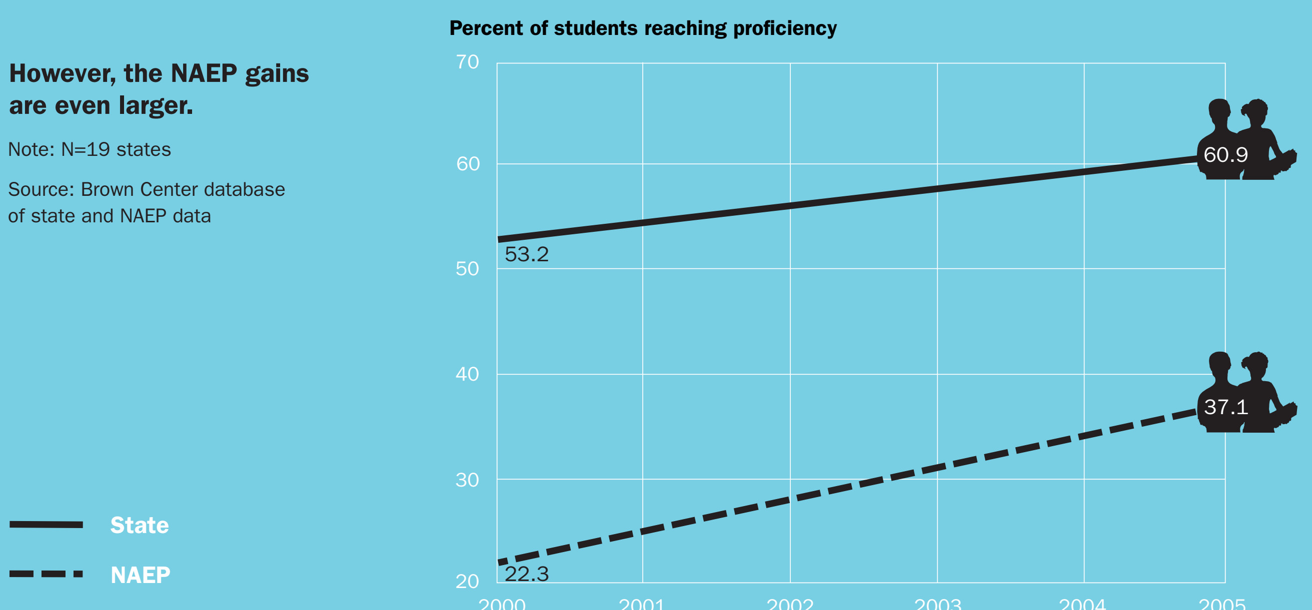
Comparison of state and NAEP proficiency

Fig **3-2**

However, the NAEP gains are even larger.

Note: N=19 states

Source: Brown Center database of state and NAEP data



Paul Peterson and Frederick Hess graded each state on whether the percentage of proficient students reported on state tests resembles the percentage of proficient students reported on NAEP. States with test scores similar to NAEP received high grades. States with scores discrepant from NAEP were given low marks. States are reporting much higher rates of proficiency than NAEP. On average, state proficiency rates in fourth grade reading exceed NAEP's by 36%. In 2006, Peterson and Hess repeated their analysis and gave seventeen states grades of D or F. Five states and the District of Columbia received A's. Peterson and Hess attribute the variation in proficiency rates to the autonomy states are granted under NCLB.[27]

These studies are fulfilling an important watchdog function. Given the discretion states are allowed under NCLB, it is imperative that states accurately measure the progress students are making toward attaining high standards. The studies of state test scores provide convincing evidence on three counts: differences between scores on NAEP and state tests exist, the differences are large, and the differences vary by state. But on a key question the studies are less persuasive. What explains the states' divergence from NAEP? All of these researchers assume that states are exploiting autonomy under NCLB to inflate test scores and that NAEP should be the benchmark by which state tests are judged. But the evidence adduced thus far only shows that differences with NAEP exist. The studies document the magnitude of these differences, not their origin, nor whether NAEP should be the final word on student learning.

Why Do NAEP and State Test Scores Differ?

To check whether NCLB is the source of the differences—or perhaps exacerbated them—we compared today's NAEP-state test gaps to the gaps that existed immediately before NCLB was passed. We first examined 2005 data and, like the aforementioned studies, computed the differences between the percentage of students that states reported reaching proficiency and the percentage as measured by NAEP. We then identified those states that participated in NAEP in 1998 (reading) or 2000 (math) and computed the NAEP-state test score differences for that year, using those calculations as measures of pre-NCLB status.[28] If NCLB is inducing states to inflate test scores, then the states' pre-NCLB differences with NAEP should be smaller than those after NCLB was enacted.

The data are presented in Figures 3-1 through 3-4. In math, the gaps have narrowed, and in reading they have widened. If states have exploited the autonomy of NCLB to artificially inflate proficiency rates, it is not evident in these data. One could make a weak case by limiting the argument to reading (see Figures 3-3 and 3-4). State-reported proficiency in eighth grade reading has climbed by about 13 percentage points while the NAEP proficiency rate has stayed the same. In fourth grade, state scores are up almost 10 points but only 2 points on NAEP. But the overall pattern does not make sense. Why would states inflate scores in reading when pre-NCLB scores in that subject were higher than in math? Look at eighth grade math (Figure 3-1). Gains on NAEP are actually outpacing state reported gains. Proficiency on NAEP

States report much higher proficiency rates than NAEP.

States are reporting large gains in 8th grade reading.

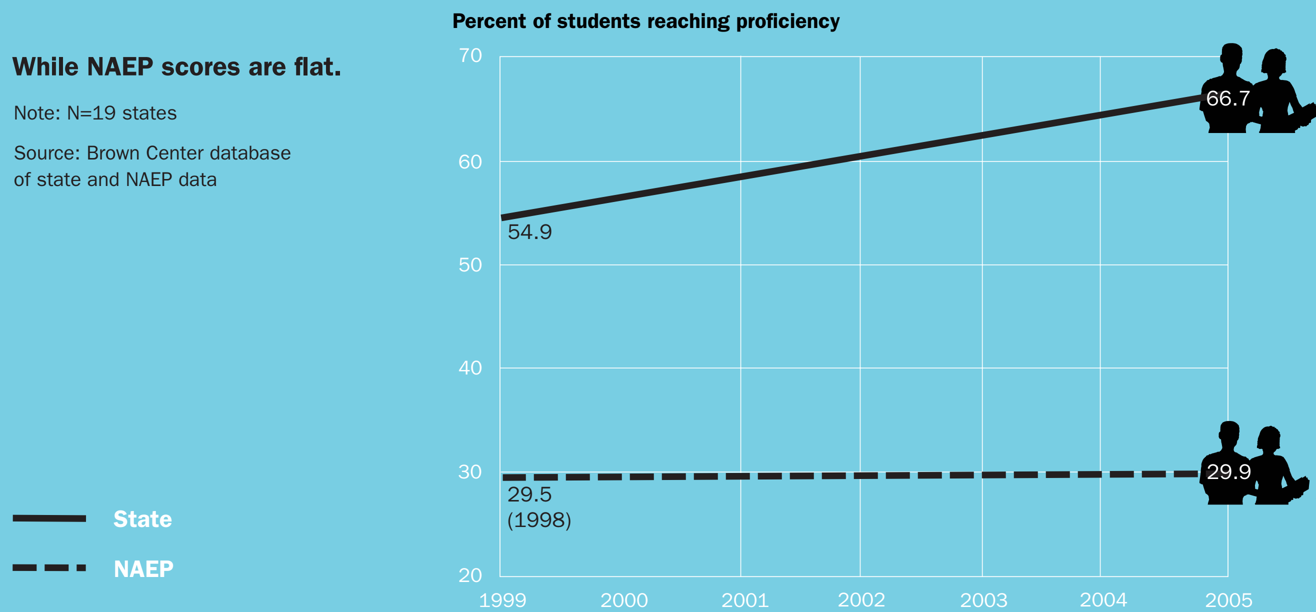
Comparison of state and NAEP proficiency

Fig **3-3**

While NAEP scores are flat.

Note: N=19 states

Source: Brown Center database of state and NAEP data



The State-NAEP gap is also widening in 4th grade reading.

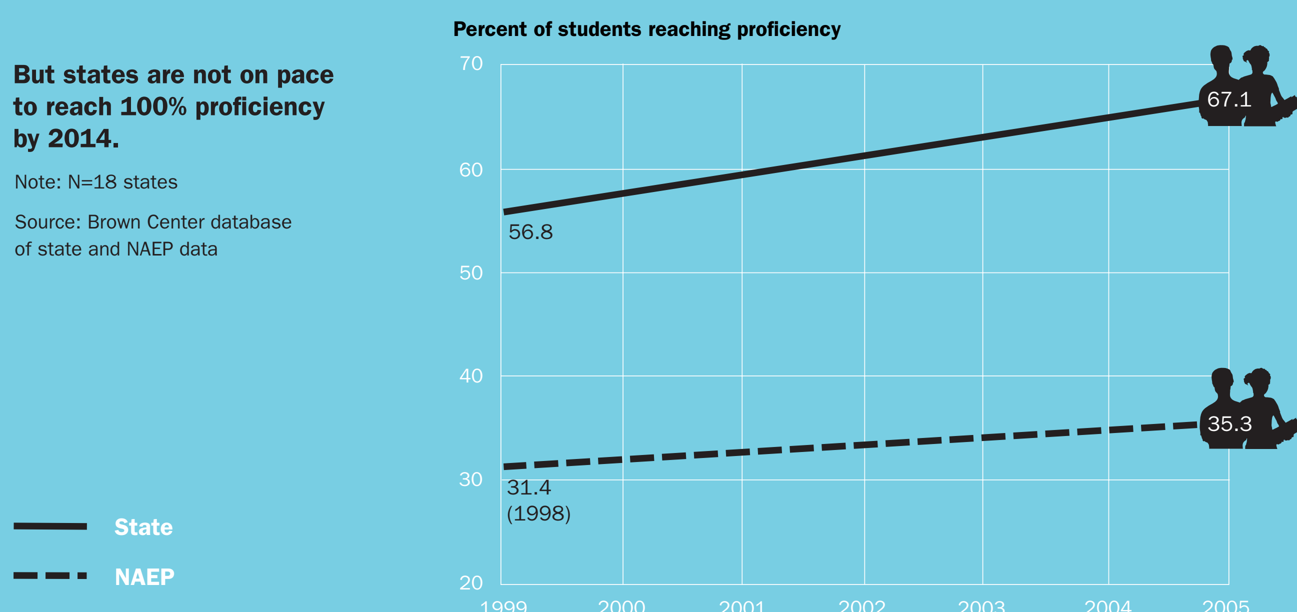
Comparison of state and NAEP proficiency

Fig **3-4**

But states are not on pace to reach 100% proficiency by 2014.

Note: N=18 states

Source: Brown Center database of state and NAEP data



has increased 6.2 percentage points and only 3.6 points on state tests.

One thing that the four line graphs make abundantly clear: states report much higher proficiency rates than NAEP. This is consistent with the findings of a study by Bruce Fuller and colleagues that examined test data from twelve states going back to the early 1990s.[29] States apparently are drawing their cut-points for proficiency—the minimum test score for students to be declared proficient—at much lower levels than NAEP. That was going on before NCLB, and it is still going on.

But it does not appear that states are "racing" to further lower cut-points in response to NCLB. In eighth grade math, proficiency increased from 50.7% to 54.3%, a gain of 3.6 percentage points in five years, or 0.7 points per year. Even if all of these gains were engineered by manipulating cut-points and absolutely no real progress has taken place, the rate of gain is not impressive. Recall that NCLB requires 100% proficiency by 2014. At the current rate, states will reach 100% proficiency in eighth grade math in 2069. And that assumes getting the last 10% of students over the proficiency bar will take the same amount of time as the other 90%. No one is racing—to the bottom or anywhere else.

If NCLB is not driving the differences in test scores, what is? One reason that scores are higher on state tests may be because students take them more seriously than NAEP. State tests matter. There are high stakes for both students and teachers. Scores are reported for each student and averaged to compute a publicly displayed score for schools. Several states hold students accountable for their performance by attaching consequences to test scores (for example, mandatory summer school, retention in grade). Schools offer special preparation and hold pep rallies to motivate students before state testing.

NAEP tests do not matter—to students or teachers. Recall the discussion from part one of this report. A small sample of schools is first selected randomly. Then the NAEP is given to a small, randomly selected sample of students within the selected schools. Average scores typically are not reported below the state level. How students or schools individually score on NAEP is never reported.

A second source of difference may lie in the content of the tests. What do the tests assess? State results are reported using the main NAEP test only, not the trend. As noted in part one of this report, the math tests of the trend and main NAEP assess different mathematical content, and the results from the two NAEP tests differ. It is dubious to conclude that a state is promulgating weak proficiency standards solely based on a discrepancy with NAEP when the two NAEP tests themselves do not agree. On the other hand, all but a few states—California and Massachusetts being the most notable exceptions—patterned their frameworks after the 1989 standards of the National Council of Teachers of Mathematics, the same model for the main NAEP framework. Theoretically, at least, the main NAEP and state tests are defining mathematics similarly.[30]

The question of content highlights the most significant omission of the studies analyzing state-NAEP differences. None of the studies cited above examined the learning assessed by NAEP. What does NAEP test? In the area of mathematics, in particular, should NAEP be the benchmark by which state results are judged? NAEP's reputation as "the nation's report card"

These results suggest that at both the fourth and eighth grades, NAEP does not reflect international standards in mathematics.

NAEP-TIMSS 8th grade test comparison: percentage of items classified at a lower grade level

Table 3-1

Content Area	NAEP	TIMSS
Number	27	4
Measurement	37	25
Geometry	43	3
Data	10	4
Algebra	18	0

Source: NCES (2006), *Comparing Mathematics Content in NAEP, TIMMS, and PISA 2003 Assessments*

rests on the test's technical advantages—it is the only test that has been regularly administered to a large, randomized national sample of students.

But serious concerns have been raised about NAEP's content validity—that it assesses the mathematics that children should learn. Knowledge of fractions, for example, is essential in eighth grade math. But NAEP contains very few items involving fractions and places an inordinate emphasis on facility with whole numbers. A 1997 U.S. Department of Education study compared the eighth grade NAEP to the leading eighth grade international assessment, TIMSS. Only 13% of NAEP items were found to involve fractions; on TIMSS the figure was 34%.[31] In 2006, the Department compared NAEP and TIMSS by classifying items by grade level. The most revealing comparison asked the question: what percentage of items on one test would be re-classified at a lower grade level on the framework of the other test? In other words, how many eighth grade NAEP items would be considered fourth grade items on TIMSS—and vice versa. As displayed in Table 3-1, NAEP appears to be much less demanding than TIMSS. In the content area called "number," 27% of eighth grade NAEP items would be re-classified at the lower grade level, but only 4% of the eighth grade TIMSS items. In geometry, 43% of NAEP items would be re-classified to a lower grade level on TIMSS. But only 3% of TIMSS geometry items would be so re-classified on NAEP.[32]

In light of this analysis, it is not surprising that the most recent TIMSS results do not correspond to NAEP results. The huge gains on NAEP reported since the mid-1990s are not showing up on TIMSS. American fourth graders achieved exactly the same score on TIMSS in 2003 as they had in 1995. Zero gain. During that same period, NAEP was indicating that fourth graders had gained more than a full school year's worth of learning. Eighth graders' gains on TIMSS also fell far short of gains on NAEP. These results suggest that at both the fourth and eighth grades, NAEP does not reflect international standards in mathematics. A charitable way of stating this is that TIMSS measures different mathematical skills and knowledge than NAEP, just as NAEP measures different abilities than state tests.

In particular, the NAEP algebra items have come under fire. Approximately 25% of eighth grade NAEP items address algebra. Some of these items require no knowledge of algebra for students to answer them correctly. A significant number of items test whether students can discern patterns, often in sequences of letters or other non-mathematical symbols, a skill that is pre-algebraic at best.[33]

In September 2005, a group of distinguished mathematicians reviewed the NAEP algebra items that are in public release, that is, items that are no longer part of NAEP tests and have been released to the public. Hyman Bass of the University of Michigan was generally positive about the test, although he did express reservations about the abundance of pattern items. Roger Howe of Yale University warned, "Given the level of these problems, I do not see how NAEP could claim to determine that any student is 'proficient' in algebra. How much algebra should be examined on the eighth grade NAEP, given the current state of the curriculum in the U.S., is a question for debate. However, one can safely say that these questions do not provide a thorough probe of Algebra I." Jim Milgram of Stanford, who has formally evaluated a number of state assessments and the tests of several other nations, offered the harshest verdict. He concluded "virtually all of the problems are at a low level relative to USA math standards," and at "even a lower level when compared to foreign problems." Milgram's most serious charge is that many NAEP items are mathematically incorrect. In Milgram's view, "On both the eighth grade and fourth grade exams 20% of the algebra problems are incorrect." The many mathematically incorrect items, Milgram added, "make it very hard to interpret NAEP test scores."[34]

Milgram's point, which is shared by other analysts who have evaluated NAEP math items, is that in order to derive meaningful levels of student proficiency from any test, one first must be able to answer the question: proficient at what? If the what of a test—which, in this case, refers to the content of NAEP—is deficient, then results from the test are meaningless. And they are meaningless no matter where proficiency cut-points are set. High cut-points will produce small numbers of proficient students, and low cut-points will produce large numbers of proficient students. That is just as true for bad tests as it is for good ones.

Most of the state standards have been judged to be poor. In 2005, the Fordham Foundation organized a team of mathematicians to evaluate state math standards. The mathematicians, in effect, downgraded states that follow NAEP by judging the 1989 NCTM standards, on which NAEP is modeled, as embracing bad mathematics. California and Massachusetts, the two states most divergent from NCTM and NAEP, received the highest grades from the mathematicians.[35]

Comparing their grades to the Peterson-Hess grades is interesting. Of the six states getting A's in Peterson and Hess's 2006 rankings, only one state, Massachusetts, received an A grade for its math standards. Two states, South Carolina and Maine, and the District of Columbia were awarded A's by Peterson and Hess but the mathematicians gave them D's. And two—Missouri and Wyoming— get A's from Peterson and Hess but received F's from the mathematicians. In sum, five of the six states receiving A's are, according to the Fordham experts, implementing math standards of poor quality and deserving of either D or F grades. Simply producing results similar to NAEP is not a good indicator of quality. Many of the states that report results similar to NAEP are giving tests that are mathematically deficient.

Most of the state standards have been judged to be poor.

Conclusion

Let's summarize what is currently known about whether states are honestly reporting test scores.

1. States report higher levels of student proficiency than NAEP. This was true before NCLB and continues to be true. It does not appear that states are racing to the bottom in response to NCLB.

2. In judging how states are responding to NCLB, one must consider both the content of state assessments and the cut-points that states use to define proficiency. Using a particular test, such as NAEP, to judge the quality of state tests assumes that the benchmarking test is measuring the correct content. In the case of the main NAEP in mathematics, that assumption is questionable. The argument that all of the prominent tests—state, NAEP, and TIMSS—measure the same knowledge and skills simply is not true, as the content analysis above demonstrates. Even the two NAEP tests do not measure the same mathematics.

3. States do vary in where they set cut-points for proficiency. To standardize the cut-points, Robert Linn has suggested taking the median score from a baseline year, say 2002, the year of NCLB's signing, and defining that score as proficient. Such an approach would eliminate the possibility of states lowering the bar to boost proficiency statistics.[36]

Many of the states that report results similar to NAEP are giving tests that are mathematically deficient.

ENDNOTES

1 See, for example, the 2000 Brown Center Report's discussion of items from the long-term trend and the 2004 Brown Center Report for an analysis of the mathematical content of the main NAEP. Tom Loveless, *The Brown Center Report on American Education* (The Brookings Institution, 2000), pp. 16-19. Tom Loveless, *The 2004 Brown Center Report on American Education* (The Brookings Institution, 2004), pp. 9-17.

2 For an historical account of test score trends in the twentieth century, see John Hillman Bishop, "Is the Test Score Decline Responsible for the Productivity Growth Decline?," *American Economic Review,* (American Economic Association), vol. 79, no. 1 (March 1989), pp. 178-197.

3 Marshall Smith as quoted in Lynn Olson, "A Decade of Effort," *Education Week,* vol. 25, no. 17 (January 2006), pp. 8-10.

4 Adjusted NAEP scores were computed weighting 2005 means for each racial/ethnic group and using 1992 population proportions. Each group's expected score was computed out of 97% (the total for these groups) and summed to create an overall expected value for 2005.

5 Richard Whitmire, "Boy Trouble," *The New Republic,* January 26, 2006. Peg Tyre, "The Trouble with Boys," *Newsweek,* January 30, 2006.

6 Sarah Mead, "The Truth about Boys and Girls," (Education Sector, June 2006). For critics see, for example, Christina Hoff Sommers, "Where the Boys Aren't," *The Wall Street Journal,* July 3, 2006.

7 National Institute of Child Health and Human Development, *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction,* (Government Printing Office, 2000).

8 Richard L. Allington, "Does State and Federal Reading Policymaking Matter?," in *The Great Curriculum Debate,* T. Loveless, ed., (The Brookings Institution, 2001), p. 271.

9 Louisa Moats, "Whole Language Lives On: The Illusion of 'Balanced' Reading Instruction," (Thomas B. Fordham Foundation, October 2002).

10 Kate Walsh, Deborah Glaser, and Danielle Dunne Wilcox, "What Education Schools Aren't Teaching About Reading and What Elementary Teachers Aren't Learning," (National Council on Teacher Quality, 2006).

11 Jay Campbell, Catherine Hombo, and John Mazzeo, *NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance,* U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, (Government Printing Office, 2000), figure 1.

12 Linda Dager Wilson and Rolf K. Blank, "Improving Mathematics Education Using Results from NAEP and TIMSS," (Council of Chief State School Officers, 1999).

13 National Commission on NAEP 12th Grade Assessment and Reporting, "12th Grade Student Achievement in America: A New Vision for NAEP," (NAGB, 2004).

14 National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989). *The Oprah Winfrey Show: Oprah's Special Report: American Schools in Crisis,* first broadcast 12 April 2006 by Harpo Productions, Inc. John M. Bridgeland, John J. DiIulio Jr., Karen Morison, "The Silent Epidemic," (Bill & Melinda Gates Foundation, 2006).

15 See Diane Ravitch, *Left Back: A Century of Failed School Reforms* (Simon and Schuster, 2006). See also William H. Kilpatrick, "The Project Method," *Teacher's College Record,* vol. 19, no. 4 (Fall 1918), pp. 319-335.

16 Ina Mullis, Michael Martin, Eugenio Gonzalez, and Steven Chrostowski, *Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades* (TIMSS & PIRLS International Study Center, 2004).

17 Nora Broege, Ann Owens, and Barbara Schneider, "Performing Well, But Feeling Bad: The Classroom Experiences of Adolescents," paper presented at the second annual conference of the Achievement Gap Initiative at Harvard University, June 2006.

18 See pages 121-132 in Maureen Stout, *The Feel Good Curriculum: Dumbing Down of America's Kids in the Name of Self Esteem* (Perseus Books, 2000).

19 The nation counts were computed using the TIMSS 2003 almanac for eighth grade. The almanac is available at: http://timss.bc.edu/timss2003i/userguide.html. "Successes" were counted as countries in which each subsequent group (agree a lot, agree a little, disagree a little, disagree a lot) had a lower score than the group preceding it.

20 In Tables 2-3, 2-4, and 2-5 several countries tied in the percent column. Ties between countries were broken using un-rounded data from the almanacs on the TIMSS 2003 User guide website. Available at: http://timss.bc.edu/timss2003i/userguide.html.

21 National Council of Teachers of Mathematics, *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 2000).

22 In only 1 out of 46 countries did the positive within country relationship between relevance and achievement hold true. A more relevant math curriculum was not associated with student achievement.

23 Herbert W. Marsh, "The Big-Fish-Little-Pond Effect on Academic Self-concept," *Journal of Educational Psychology,* vol. 79 (1987), pp. 280-295.

24 Tom Loveless, "Charter School Achievement and Accountability," in Peterson and West, eds., *No Child Left Behind? The Politics and Practice of School Accountability,* (The Brookings Institution, 2003).

25 Fordham Foundation, "Gains on State Reading Tests Evaporate on 2005 NAEP," October 19, 2005.

26 Kevin Carey, "Hot Air: How States Inflate Their Educational Progress Under NCLB," (Education Sector, May 2006).

27 The first sentence of Peterson and Hess (2005) reads, "It turns out that in complying with the requirements of No Child Left Behind, some states have decided to be a whole lot more generous than others in determining whether students are proficient at math and reading." Paul Peterson and Frederick Hess, "Johnny Can Read… in Some States. Assessing the Rigor of State Assessment Systems," (*Education Next,* Summer 2005). The second analysis can be found in Paul Peterson and Frederick Hess, "Keeping an Eye on State Standards, A Race to the Bottom?," (*Education Next,* Summer 2006).

28 Fourth and eighth grade reading scores were obtained from 1999 state tests.

29 Bruce Fuller, Kathryn Gesicki, Erin Kang, Joseph Wright, "Is the No Child Left Behind Act Working? The Reliability of How States Track Achievement," (Working Paper 06-1, Policy Analysis for California Education, 2006).

30 But the two outliers are not acting as anticipated. Massachusetts actually reports fewer students proficient in mathematics than the NAEP, and California is close. The two states' math frameworks are quite similar in de-emphasizing the mathematical content that NAEP embraces. In the case of these two states, then, divergence from the NCTM model may not matter as much as where the cut-points for proficiency are drawn.

31 Don McLaughlin, John Dossey, and Fran Stancavage, "Validation Studies of the Linkage Between NAEP and TIMSS Eighth Grade Mathematics Assessments," in U.S. Department of Education, National Center for Education Statistics, *Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A Technical Report* by Eugene Johnson and Eugene Owen, (Government Printing Office, September 1998) p. A-2.

32 Teresa Neidorf, Marilyn Binkley, Kim Gattis, David Nohara, *Comparing Mathematics Content in NAEP, TIMMS, and PISA 2003 Assessments: Technical Report,* (Government Printing Office, 2006), tables 10-14.

33 Tom Loveless, *The 2004 Brown Center Report on American Education* (The Brookings Institution, 2004), pp. 9-17.

34 Material quoted from Algebraic Reasoning: Developmental, Cognitive, and Disciplinary Foundations for Instruction, a conference held at the Brookings Institution (2005). Quotes are available at: http://www.brookings.edu/gs/brown/algebraicreasoning.htm.

35 Fordham Foundation, "The State of State Math Standards" (2005).

36 Robert Linn, "CRESST Policy Brief 8: Fixing the NCLB Accountability System," (The National Center for Research on Evaluation, Standards, and Student Testing, 2005).

Designed by Lipman Hearne, Chicago/Washington

THE BROOKINGS INSTITUTION

1775 Massachusetts Avenue, NW • Washington, DC 20036
Tel: 202-797-6000 • Fax: 202-797-6004
www.brookings.edu

THE BROWN CENTER ON EDUCATION POLICY

Tel: 202-797-6469 • Fax: 202-797-2480
www.brookings.edu/browncenter