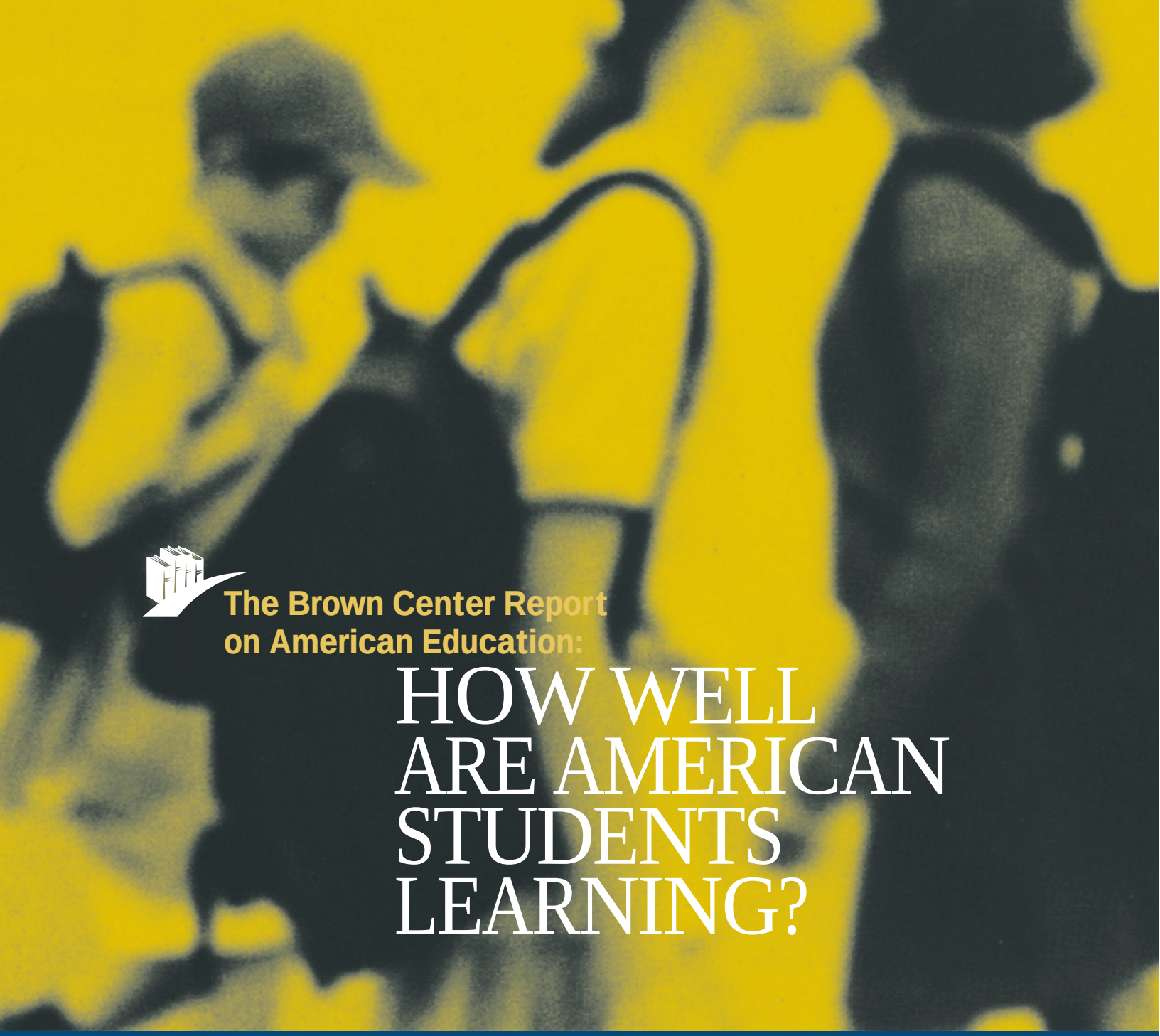


The Brown Center Report on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

Focus on Math Achievement

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 1-800-275-1447, fax 202-797-2960, e-mail BIBOOKS@brookings.edu, or visit online at www.brookings.edu.

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

The Brown Center Report
on American Education:

HOW WELL ARE AMERICAN STUDENTS LEARNING?

Focus on Math Achievement

September 2000
Volume I, Number 1

by:
TOM LOVELESS
Director, Brown Center on Education Policy

PAUL DIPERNA
Research Assistant

TABLE OF CONTENTS

THE BROWN CENTER REPORT ON AMERICAN EDUCATION

There was a time when student test scores primarily concerned two groups: individual parents, as they received reports on their children's progress, and real estate agents, as they helped home buyers compare schools in different neighborhoods. Test scores were difficult to obtain and rarely discussed in public. But now measures of student achievement are splashed across the front page of major newspapers, widely available on the Internet, and the subject of intense scrutiny and furious spin. Politicians closely watch test scores. From the race to the White House to the thousands of contests for local school boards, candidates stretch and bend scores to make them look as good or as bad as possible. Teachers unions and other organizations cite data to defend public schools and assure the public that all is well—or at least not as bad as everyone thinks. On the other side, critics of public schools publish voluminous studies documenting a steady decline in student performance.

What should the average citizen believe?

The purpose of this report is fourfold: to report on the direction of achievement in American schools, that is, to determine whether it's going up, down, or sideways; to figure out whether any change that is detected is big, small, or insignificant; to dig under the numbers and uncover the policies and practices influencing the direction of student achievement; and, finally, to figure out whether the public is getting the full story on student learning. Americans spend $350 billion each year on elementary and secondary education. They deserve an accurate,

nonpartisan, no-holds-barred, data-driven account of what they're getting for their money.

The Brown Center Report will appear annually, this being the first edition. Although varying in content from year to year, the report will be presented in the same three sections. The first section will use the latest and best evidence available to evaluate student achievement in America's schools. The second section will go into greater depth on a theme related to student learning. This year's theme is mathematics achievement. The third section will evaluate the impact of policies and practices on student learning. This year's topics are the use of calculators in math instruction and state and federal programs that single out exemplary schools for special recognition.

Part I THE NATION'S ACHIEVEMENT

FOR THE LONG-TERM PICTURE OF THE NATION'S PROGRESS, the best measure of student achievement is the National Assessment of Educational Progress (NAEP). The NAEP gauges learning in several academic areas by testing students of three different ages: nine, thirteen, and seventeen. For the sake of simplicity, and because these subjects are both foundational and of greatest interest to the American public, this report focuses solely on achievement in reading and math.

All three age groups made small gains in reading from 1971 to 1999 (see Figure 1). Nine year olds gained four scale score points, up from 208 to 212 on a scale that ranges from 0 to 500 points. But their 1999 scores were below those of 1980. Thirteen year olds also gained four points, from 255 to 259, scoring the same in 1999 as they had in 1980. Seventeen year olds gained the least (three points), inching forward from 285 to 288. Seventeen year olds' scores peaked in 1988 and suffered a small decline in the 1990s.[1]

In math, the picture is much brighter for all three age groups (see Figure 2). From 1973 to 1999, nine year olds gained thirteen points, improving from 219 to 232; followed by thirteen year olds, with a gain of ten points, moving from 266 to 276; and seventeen year olds, a gain of four points, up from 304 to 308. The math scores registered in 1999 were the highest on record for all three ages. In both subjects, achievement gains were greatest at age nine and shrank for older students.

Clearly, the story is not one of disastrous decline. Slow and steady gains are being made. Nor is it cause for national celebration. Today's nine year olds are better at reading and math than nine year olds in the early 1970s. But when older students are compared with their peers from the past, about two-thirds of the improvement has evaporated before the end of high school. The size of the gains and the pace of improvement are both important. This kind of analysis is best understood by expressing growth in standard deviations (SDs), a unit of measure commonly used by statisticians (see Table 1). The reading gain at age nine (+0.10) holds up through age thirteen (+0.11), then fades to

Clearly, the story is not one of disastrous decline. Slow and steady gains are being made. Nor is it cause for national celebration.

[In math] it would take a little more than eighty-three years before American eighth graders were performing at a level equal to their Japanese counterparts.

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. The number of students selected for a NAEP national sample for any particular grade and subject is 7,000 or more.

There are three NAEP test types: (1) the main NAEP gauges national achievement while also reflecting current practices in curriculum and assessment, (2) the long-term trend NAEP allows reliable measurement of change in national achievement over time, and (3) the state NAEP measures achievement of students in participating states. These assessments use distinct data collection procedures and separate samples of students.

Since 1990, the main and state math tests have been governed by a framework reflecting recommendations of the National Council of Teachers of Mathematics (NCTM). The long-term trend test consists of the same items and test procedures used in 1973.

a trace, +0.06, for seventeen year olds. The math gain also drops off steadily as students get older, from +0.38 to +0.30 to +0.13.

Are the gains big or small? Is progress fast or slow?

The gains are small to modest. In experiments designed to discover if particular practices affect academic achievement, statisticians generally regard SD gains greater than +0.50 as large, from +0.20 to +0.50 as modest, and less than +0.20 as insignificant. How big are these gains in the real world? Here's one illustration. The difference between the average height of fifteen and sixteen year old girls in the United States is 0.20 SD, barely noticeable to the naked eye. The difference between thirteen and eighteen year old girls is 0.80 SD, which no one could miss.[2] None of the NAEP gains are even close to 0.80. The largest gain, +0.38 SD for the math achievement of nine year olds, is only modest in size.

Evaluated by another method, the math gains for nine and thirteen year olds appear larger. Since the scale score difference between nine and thirteen year olds has averaged about forty-five points, an eleven point NAEP gain represents approximately one year's worth of learning for ages nine to thirteen. A thirteen point gain was registered by nine year olds, meaning that they have gained a little more than a year's worth of learning and probably know as much mathematics as a ten year old in 1973. Impressive. The difference between thirteen and seventeen year olds has averaged about thirty-four points, so eight and one-half points represent one year's worth of learning for this age group. By the same metric, thirteen year olds have also gained about one year's worth of learning in mathematics since 1973. Also impressive. In contrast, seventeen year olds have only gained a few months of math learning.[3]

Incremental growth occurred in the 1990s. In math, two point gains were registered by nine year olds from 1990 to 1999 and thirteen year olds from 1994 to 1999. Is the rate of progress as slow as it seems? Consider how long it would take to close the gap between Singapore and the United States in eighth grade mathematics. According to the Third International Math and Science

Change in Achievement (SD Units and Years of Learning) — Table 1

Reading (1971–1999)

	SDs	Years
Age 9	+0.10	+0.34
Age 13	+0.11	+0.53
Age 17	+0.06	+0.40

Math (1973–1999)

	SDs	Years
Age 9	+0.38	+1.16
Age 13	+0.30	+1.19
Age 17	+0.13	+0.48

Study (TIMSS), which included a math test given in several countries in the mid-1990s, thirteen year olds in Singapore score approximately 1.50 SDs higher than thirteen year olds in the United States. Let's assume that the United States starts closing this gap at a rate of 0.12 SD per decade, which happens to be the average rate of progress for thirteen year olds during the history of NAEP math testing. At this speed, it would take about 125 years to catch Singapore. The gap with Japan is about one full SD, so it would take a little more than eighty-three years before American eighth graders were performing at a level equal to their Japanese counterparts. These estimates assume that for the next century students in Singapore and Japan continue performing at the same level as they are now. If they were to raise their math skills, it would take even longer for American students to match them in achievement.[4]

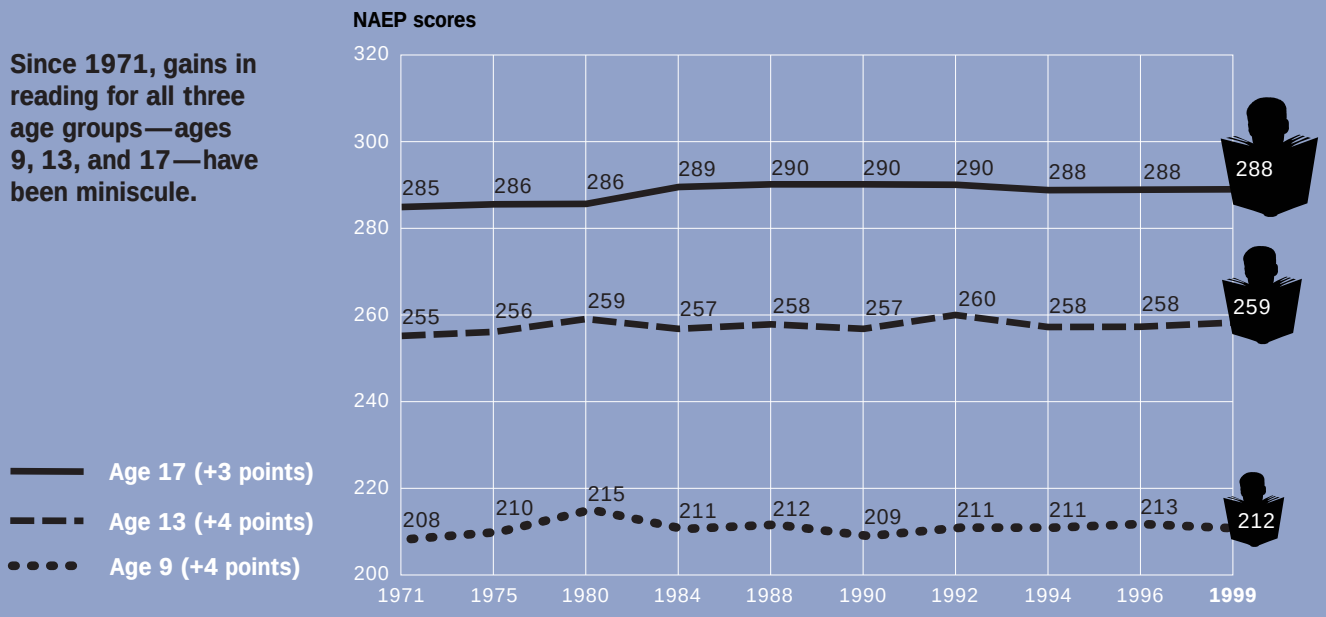
What about recent progress?

The 1990s do not stand out as a time of great strides forward in academic achievement. Although achievement continued to improve, it did not accelerate in the century's final decade. After several years of national debate over the quality of American schooling, critics question whether the tens of billions of dollars spent on educational reform have been worth it. Others point out that changing social conditions—sharp increases in child poverty, single-parent families, non-English speaking students, and incarcerated parents—have made schooling more difficult.

A shortcoming of the NAEP tests is that they are only given intermittently. Both subjects were tested in 1999, and results were released in August, 2000. Most states have begun administering their annual achievement tests, however, and many are posting

Reading scores are flat — Fig 1

Reading scores as measured by the National Assessment of Educational Progress (NAEP), 1971–1999. NAEP data are expressed as scale scores, ranging from 0 to 500.



State Test Results
(1998–1999)

Table 2

Reading

		Grade 4
Improvement	↑	14 (67%)
No Change	↔	4 (19%)
Decline	↓	3 (14%)
Total		21 states
		Grade 8
Improvement	↑	13 (48%)
No Change	↔	7 (26%)
Decline	↓	7 (26%)
Total		27 states
		Grade 10
Improvement	↑	12 (57%)
No Change	↔	4 (19%)
Decline	↓	5 (24%)
Total		21 states

NOTE: Data obtained from 35 states (and the District of Columbia) that administered the same achievement test in 4th grade, 8th grade, or 10th grade, in either math or reading.

scores on the Internet, offering researchers another tool to gauge the achievement trends reported by NAEP.

What are state tests saying?

We are not in the position to judge the validity of state tests. They are too new, extremely diverse in content, and employ different approaches to assessing student knowledge. However, what states are reporting to constituents about the direction of student achievement can be evaluated. State test results are tied increasingly to important outcomes. Based on test scores, schools may receive monetary rewards or states may impose sanctions. Students may be either promoted to the next grade or forced to repeat the same grade. Unlike the NAEP, the state tests really count.

From public data available on websites maintained by the fifty states and the District of Columbia, we found thirty-six states that posted reading and math scores in 1998 and 1999, allowing us to compare achievement in these two years.[5] Did scores go up, down, or remain the same? Do the states confirm or contradict the latest NAEP scores?

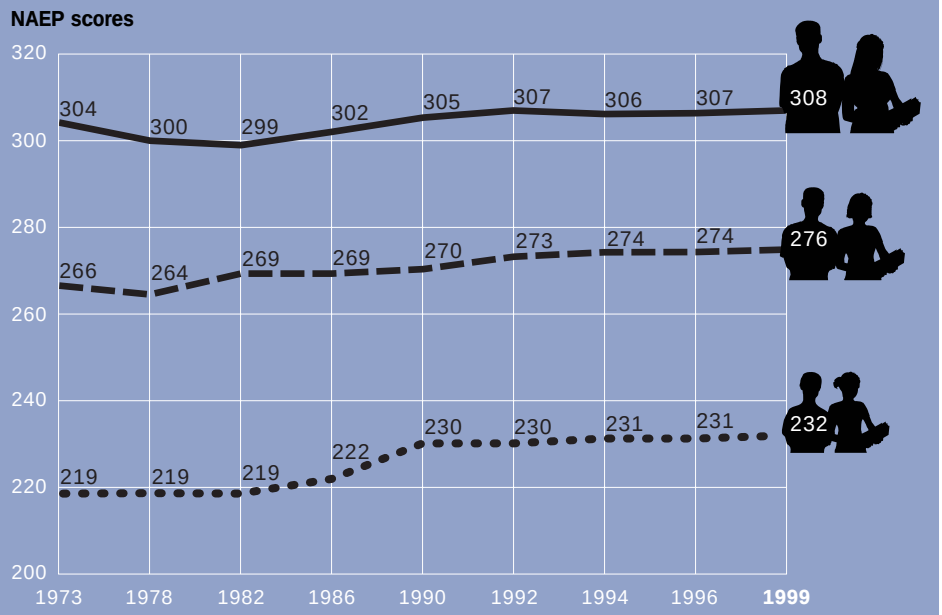
State tests suggest that achievement was still heading up at the end of the decade. More states reported gains than losses in both reading and math. In reading, 67 percent of the states that tested fourth graders reported a gain (see Table 2). The percentage was lower for eighth and tenth graders. In math, an overwhelming 86 percent of states showed a gain at fourth grade (see Table 3). Although these data are reporting on states, not students, the pattern of decline across grade levels in math is remarkably similar to that found in NAEP. A slump begins sometime after fourth grade and extends into high school.[6] It is more difficult for states to

Math scores are rising slowly

Mathematics scores as measured by the National Assessment of Educational Progress (NAEP), 1973–1999. NAEP data are expressed as scale scores, ranging from 0 to 500.

Fig 2

In 1999, youth in all three age groups—ages 9, 13, and 17—registered their highest scores on record, but gains were greatest at age 9 and least among the older student groups.



demonstrate growth in the middle and high school grades than the elementary grades. Possible explanations for the middle-grade slump are discussed below.

Does the type of state test matter?

Some states administer off-the-shelf tests purchased from publishers. Others have developed their own tests or give a commercial test that has been customized for their use. Does the type of test make a difference in the results? Yes, but only in reading (see Figure 3). States that developed their own reading tests were more likely to report improvement from 1998 to 1999. At the fourth grade, the difference is large. About 90 percent of the states with custom tests reported gains, compared to only 45 percent with an off-the-shelf test. Bear in mind that this comparison involves a small number of states—ten use custom tests and eleven use commercial tests to assess fourth grade reading—so the data are merely suggestive, not conclusive.[7]

Nevertheless, the fact that different kinds of state tests produce different results in reading is intriguing. It is too early to say why this is happening or to predict that it will continue, but there are several plausible explanations. It could be that when states develop their own tests, they make them easier. That would make a strong case for using tests created by external authorities, experts outside the system being tested. But easier tests should have produced higher scores in both 1998 and 1999—not affecting the gain over time. Perhaps customized tests are more engaging and have a stronger "comfort factor," the growth students may exhibit after becoming accustomed to a test's protocols. Or it could be that state-developed

State Test Results (1998–1999) — Table 3

Math

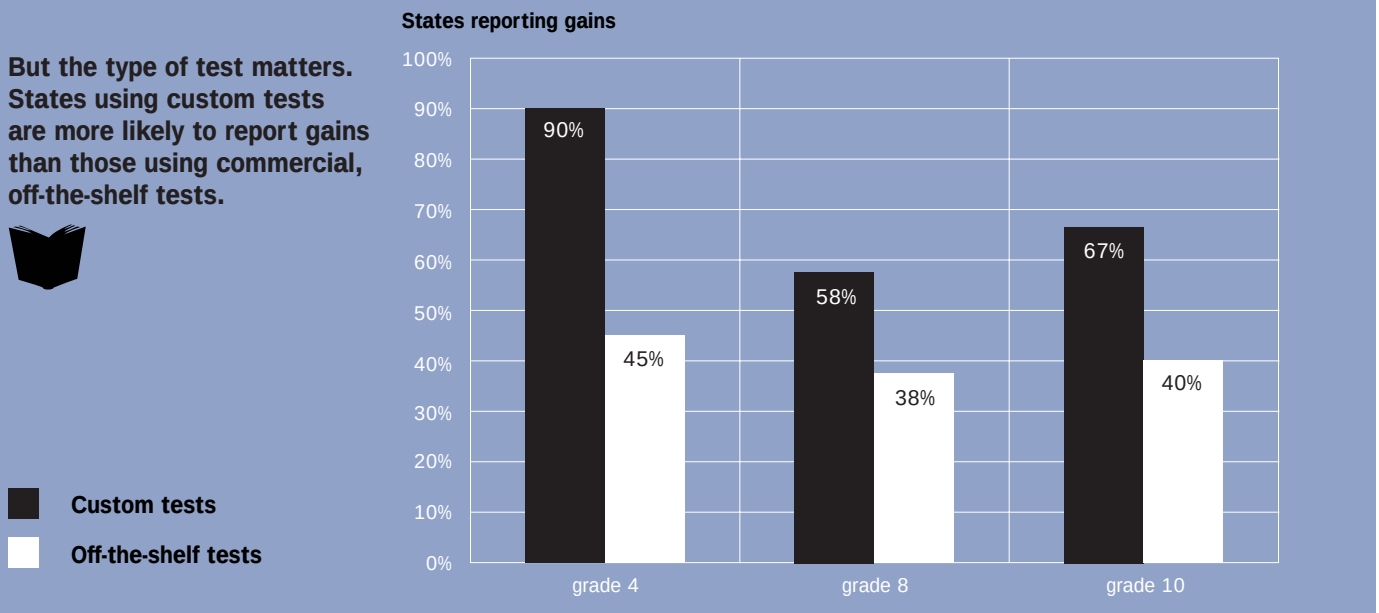
Grade 4		
Improvement	↑	18 (86%)
No Change	↔	1 (5%)
Decline	↓	2 (10%)
Total		21 states
Grade 8		
Improvement	↑	17 (61%)
No Change	↔	6 (21%)
Decline	↓	5 (18%)
Total		28 states
Grade 10		
Improvement	↑	12 (52%)
No Change	↔	3 (13%)
Decline	↓	8 (35%)
Total		23 states

NOTE: Data obtained from 35 states (and the District of Columbia) that administered the same achievement test in 4th grade, 8th grade, or 10th grade, in either math or reading.

States are reporting gains in reading . . . — Fig 3

Percentage of states reporting gains in reading in 1998–1999

But the type of test matters. States using custom tests are more likely to report gains than those using commercial, off-the-shelf tests.



The fact that different kinds of state tests produce different results in reading is intriguing.

tests are more in synch with the curriculum taught in classrooms ("alignment" is the buzzword), allowing low-performing schools in one year to pinpoint areas of deficiency and improve in the following year. If that were true, however, one would also expect math scores to be related to the type of test. But they are not (see Figure 4).[8]

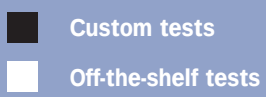
Achievement tests are growing in importance. Parents surely want to know if their children are likely to bring home improving test scores merely because of the type of test that is given instead of true improvement in reading skills. Teachers, principals, and policymakers would like to know too. In future years, with more state test data to analyze, researchers should be able to look into this issue in greater depth and confirm or deny whether custom tests hold an advantage over off-the-shelf achievement tests.

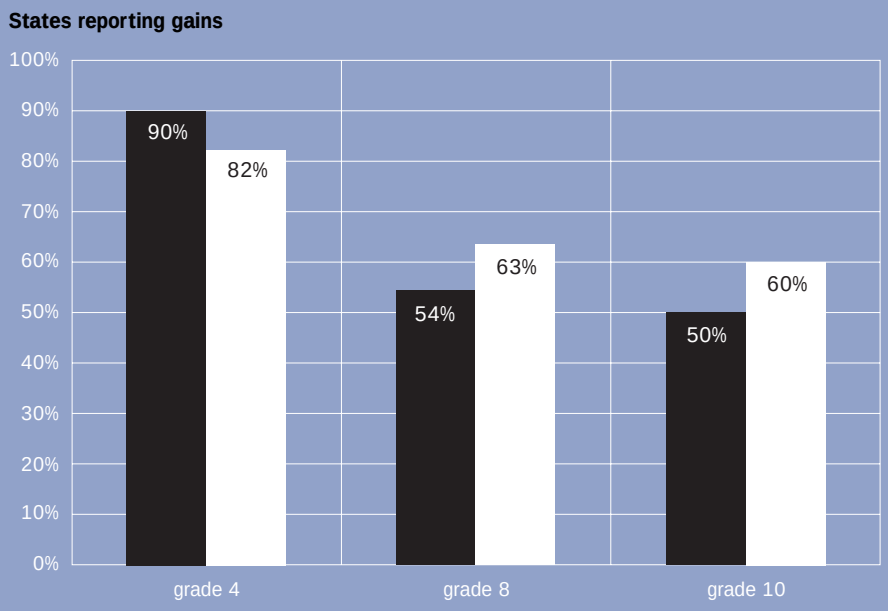
States are reporting gains in math

Percentage of states reporting gains in mathematics in 1998–1999

Fig 4

Unlike reading, math scores are not related to whether students take custom or commercial tests. In math, the differences between the two tests virtually disappear.





Part II

A CLOSER LOOK AT MATHEMATICS ACHIEVEMENT

A SLUMP IN MATH GAINS BEGINS AFTER FOURTH GRADE and extends through high school on both national and state tests. The middle-grade slump also appears on the most prominent international test, TIMSS, as the relative ranking of American students falls precipitously after fourth grade. No other country has a sharper drop in math ranking than the United States. A British publication, *The Economist*, concluded, "The longer children stay in American schools, the worse they seem to get."[9] This overstates the case. Older students have made gains in learning. The slump is not in absolute achievement, but in the pace of improvement. It decelerates after fourth grade.

No other country has a sharper drop in math ranking than the United States.

Why the middle-grade slump?

Researchers who have studied TIMSS data offer several theories for why this is happening. One argument is that the slump is an artifact of the TIMSS test. Many of the European countries in the TIMSS high school sample include students who are in their thirteenth or fourteenth year of schooling, the equivalent of the freshman and sophomore year of college in the United States.[10] This clearly is an unfair comparison. But it doesn't explain why the slump is also apparent on national and state tests in the United States, nor why it exists between fourth and eighth grades.

Another theory is that student motivation plays a role in performance. In the words of Mark Reckase of Michigan State University, the NAEP is a "drop from the sky test."[11] Students are given the test without warning or preparation, unlike the SAT or Advanced Placement tests. The NAEPs are also low stakes tests, meaning that students don't pay a price for poor performance or receive rewards for doing well. The slump may reflect young children's intrinsic motivation to do well on any test that their teacher happens to give them—and the inevitable waning of this innocent impulse in the teenage years. An analysis of extended response items on the 1996 NAEP revealed that six percent of fourth graders, thirteen percent of eighth graders, and twenty-five to thirty percent of twelfth graders left items blank or gave "off-task" answers (for example, used the test sheet for art work).[12]

A falling dropout rate since the early 1970s could also be depressing scores of older students. If students who would have left school two decades ago are now part of the NAEP sample of seventeen year olds—and if these students are also low achievers—their presence in school would hold down more recent scores. Like questions about the validity of TIMSS high school samples, however, this argument does not explain the portion of the slump that appears before high school.

Are tracking and ability grouping to blame?

A group of TIMSS researchers speculate that tracking and ability grouping are the culprits, pulling down U.S. performance by exposing too few students to advanced math.[13] Tracking is when schools group students by ability or prior achievement into separate classes and offer them a different curriculum. It is usually found in high schools and middle schools. Ability grouping refers to the same practice, but the groups are taught a different curriculum *within* the same class. Ability grouping is primarily an elementary school practice.

Although high-scoring Asian and European countries typically do not differentiate curriculum before high school, the case against tracking and ability grouping is weak. Meta-analyses, which combine the findings of many studies on the same topic, calculate that tracking has zero effect on achievement. Ability grouping's effect is consistently positive, especially in math.[14] In the United States, tracking in mathematics isn't commonplace until students are split off for algebra or pre-algebra, usually no earlier than seventh grade—and too late to be held responsible for an achievement decline that begins after fourth grade. In 1996, for example, about 40 percent of fourth grade teachers in NAEP said they form ability groups for math, but only 18 percent said students were assigned to their class based on ability (tracking). At the eighth grade, the figures were 24 percent for ability grouped math and about 60 percent for tracked math. At both grade levels, students in the tracked classes registered higher math scores than the untracked students.[15]

Other evidence from abroad casts doubt on the tracking indictment. Germany begins tracking in all subjects at age eleven, but it doesn't have the slump. The United States decline in international ranking is steeper in science than math.[16] But science is rarely tracked in American middle schools; 82 percent of the eighth grade science teachers in NAEP say they teach untracked classes.[17] In American primary grade classrooms, ability grouping is a near universal practice in the teaching of reading, unparalleled in other countries. Yet American fourth graders score near the top in reading, higher than any other American age group on an international assessment.[18] And when high-scoring European and Asian countries begin tracking—in high school—they track far more severely, separating students by school, not by classroom as in the United States. Yet they continue to outdistance American achievement in the high school grades.[19]

What about curriculum and instruction?

Other explanations point to particular aspects of curriculum and instruction, though some of the reasons offered for the poor U.S. results are contradictory. For example, some researchers have worried that the American math curriculum in grades five to eight covers so many topics that it is fragmented and shallow, "a mile wide and an inch deep." But researchers also complain that it focuses too narrowly on arithmetic (suggesting it may not be so fragmented after all) and recommend that geometry and statistics receive more attention (thereby adding to the diffuse nature of the curriculum).[20]

The TIMSS data do not include measures of students' previous learning. As a result, it is not possible to see how classroom practice affects changes in test scores. Whether curriculum and instruction affect achievement or vice versa is difficult to determine. For example, studies of Japanese and American math books show that eighth grade textbooks in the United States devote time to arithmetic. Japan's eighth grade books don't include arithmetic.[21] Are the books causing differences in math proficiency or responding to pre-existing differences? Can Japanese texts drop arithmetic in eighth grade because it's safe to assume that students have mastered it by then, whereas American texts can't make that assumption? Or are U.S. middle school textbooks holding students back by presenting mathematics that students have already learned?

Studies of instruction suffer from similar problems. TIMSS videotape studies of Japanese and American classrooms have detected national differences in teachers' instructional styles.[22] A thumbnail description: American teachers show students how

It is not teachers' instructional choices, per se, that are impeding United States achievement, but a cultural "system" of teaching that defines what teachers and students should be doing in classrooms.

to complete a particular procedure, then assign lots of practice problems that students complete while working alone. Japanese teachers pose a problem, work through alternative solutions that students generate, often with students meeting in small groups, then engage the entire class in additional problems that delve deeply into the topic.

A plausible hypothesis is that Japanese teachers employ strategies that are more effective and American teachers, less effective. An equally plausible position is that Japanese and American teaching styles are largely irrelevant to the two nations' achievement differences. Achievement data were not collected as part of the videotape study, so researchers are unable to link the observed teaching strategies to how much students learned. In the TIMSS data that allow for research on achievement, variables depicting teaching methods possess weak explanatory power, for example, accounting for less than 5 percent of the TIMSS score difference between a consortium of high-achieving school districts in the United States and the United States as a whole (the socioeconomic composition of classrooms explains 50 percent).[23] This is not surprising considering the vast literature on instructional methods that preceded TIMSS. Decades of studies, many with higher quality data than TIMSS, some even with randomized assignment of subjects to experimental conditions, have failed to single out the instructional approaches attributed to Japanese teachers as exceptional.[24]

Then what causes the slump?

James Stigler and James Hiebert, researchers who have studied the TIMSS videotapes, stress that it is not teachers' instructional choices, per se, that are impeding United States achievement, but a cultural "system" of teaching that defines what teachers and students should be doing in classrooms.[25] Classrooms are not hermetically sealed off from the national culture. Indeed, persuasive explanations for the middle-grade slump point to facets of American culture that discourage academic achievement in adolescence.

Differences in teaching could be reflecting differences in the importance of academic study at age thirteen in Japan and the United States. Almost two-thirds of Japanese eighth graders attend juku, special schools offering intensive after-school instruction in basic skills so that students may do well on high school entrance exams, given near the end of ninth grade.[26] This shifts the burden of teaching basic skills and reinforcing skills through constant review out of the regular classroom and onto the juku. Cultural incentives are geared toward achievement. Japanese students labor to get into the most prestigious high schools because the high school of attendance is highly predictive of the college and career that follow. The entrance exams loom immediately ahead for eighth graders. The Japanese exams touch each student's future in a way that no American thirteen year old can begin to comprehend, motivating the Japanese student to take the academic demands of eighth grade very seriously. Would instruction in American math classes be different in such an environment?

Other studies suggest that American schools could demand much more of students. A recent analysis from the University of Pennsylvania shows that students in the TIMSS countries that slump from fourth to eighth grade aren't assigned as many minutes of homework as students in the other countries.[27] In addition, countries that succeed in maintaining adolescents' achievement require students to demonstrate mastery of the curriculum before graduating. Working hard in school is reinforced by both school practice and education policy.[28]

Other aspects of teen culture also must assume some responsibility for the slump. The way American adolescents spend their

One test, the main NAEP, is telling us that students are getting better at math, while the other, the long-term trend NAEP, is saying that math achievement remains flat.

time—especially the huge number of hours working in part-time jobs, hanging out with friends, and participating in sports and other extracurricular activities—is unique in the world. Two-thirds of high school students hold down part-time jobs; one in six works more than twenty-five hours each week. In European and Asian countries, this is unthinkable. Teens already have a job: going to school. The American teenager's time is organized around activities that undermine the value of academic accomplishment.[29]

It is unlikely that the middle-grade slump is a mirage. American schools, teachers, parents, employers, and policymakers all contribute to making academic achievement more difficult to accomplish after fourth grade.

Two different NAEPs, two different scores

Let's return to NAEP test scores. The NAEP statistics that have been cited up to this point in the report are known as "long-term trend" scores. There are in fact two sets of national scores in NAEP, produced from two different samples. (There are also state scores, but they aren't relevant to the following discussion.) The "main" NAEP test, first administered in 1990, is given to a random sample of students across the nation. It is governed by a framework that may be altered from time to time to reflect changes in curriculum or testing practices. The long-term trend test (hereafter called the "trend") consists of a set of items that are given to a separate random sample. The trend sample always takes the exact same test under the exact same conditions, making it possible, in math, to monitor the nation's progress since 1973. The stable conditions and consistent test items serve as an anchor for measuring achievement over time. Otherwise, we wouldn't know if a score increase was because students had learned more or because of changes in the test.[30]

Press coverage rarely notes the bifurcated nature of NAEP, treating scores from the two tests as if they simply come from different pages of "the nation's report card," the nickname for NAEP. In February, 1997, when scores from the 1996 main NAEP were released, the headline in the *New York Times* declared, "National Tests Show Students Have Improved in Math." The main NAEP scores did indeed show improvement. But not the trend. When the trend scores were released later in the year, *Education Week* accurately reported, "None of the 1996 scores was different enough from the 1994 scores to be considered statistically significant." How confusing. Two NAEP tests testing the same subject in the same year produced different results.[31]

Can the same test report different results?

The following analysis compares the gains in math on the main and long-term trend assessments. Both samples are nationally representative. If the tests encompass the same mathematics, they should produce similar scores. If the scores are different, then a divergence in either test content or protocol may have caused the discrepancy. Keep in mind that the trend NAEP tests students at certain ages (nine, thirteen, and seventeen) while the main NAEP tests at certain grades (four, eight, and twelve).

Results from the two tests have diverged (see Figures 5, 6, and 7). From 1990 to 1996, fourth graders gained eleven points on the main test. In the same time period, the math performance of nine-year-old fourth graders declined by one point on the trend test, opening a twelve point scale score gap between the two tests. Eighth graders gained nine points on the main, but on the trend scores for thirteen-year-old eighth graders were unchanged, opening a nine point gap. Twelfth graders improved by ten points on the main, whereas seventeen year olds in

Two national tests give sharply different impressions of math performance

Math scores as measured by the National Assessment of Educational Progress (NAEP) for age 9 / grade 4, age 13 / grade 8, and age 17 / grade 12

The main NAEP test shows substantial gains in math scores between 1990 and 1996. The trend test shows very slight to no gains. Most recent data available are from 1996.

Fig 5

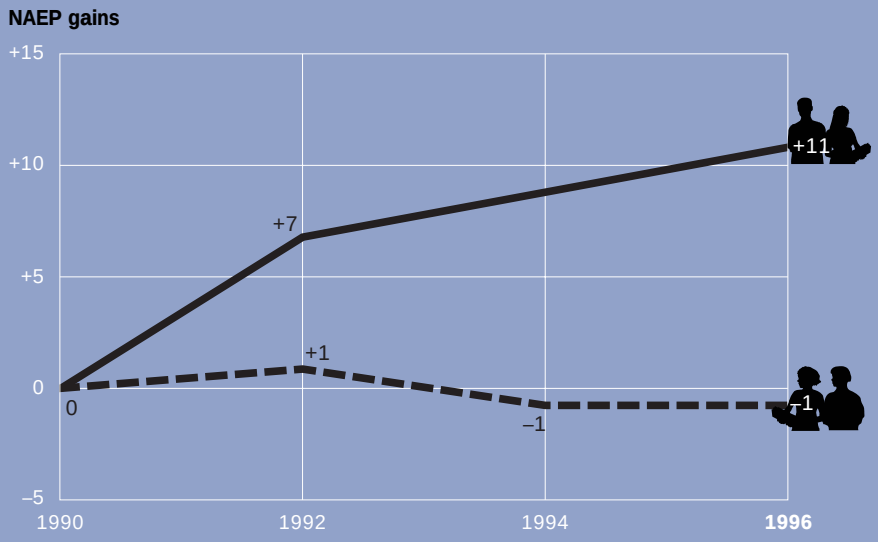


Fig 6

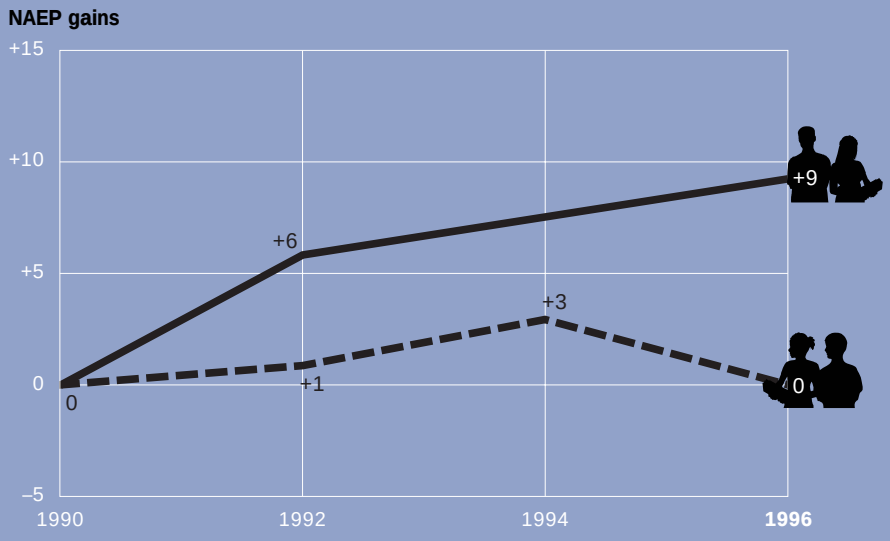
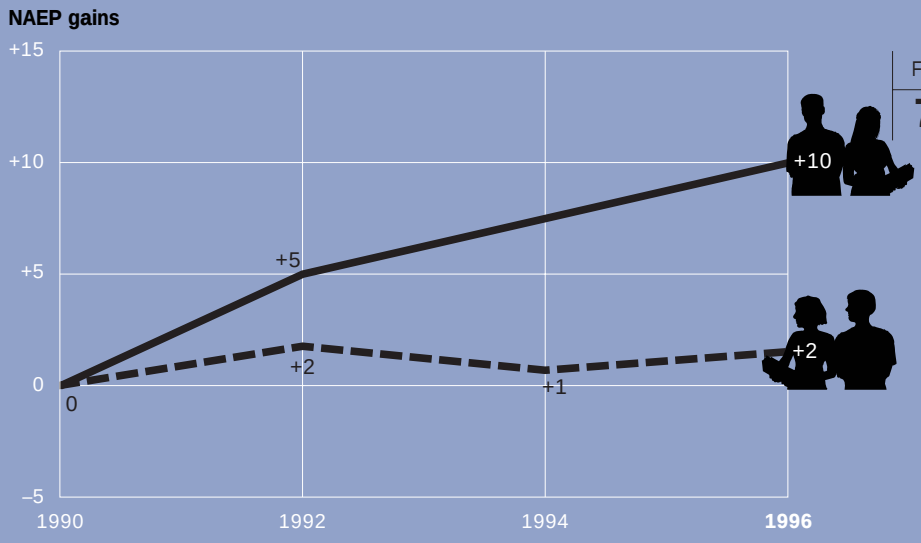


Fig 7



twelfth grade gained only two points on the trend, an eight point difference.[32]

The two tests give sharply different impressions of math performance. One test, the main NAEP, is telling us that students are getting better at math, while the other, the long-term trend NAEP, is saying that math achievement remains flat. To put the issue in a proper context, one can refer back to Figure 2 and see that—at all three grade levels—the gulf that developed between the two tests between 1990 and 1996 is approximately as large as the gains registered in nearly three decades of NAEP testing.

Why the discrepancy?

As pointed out above, the main NAEP was designed to keep abreast of changes in math curriculum. The innovations include short-answer and extended-response items on which students may receive partial credit. The main NAEP appears to contain a larger proportion of geometry problems than the trend, at least based on the items available to the public on the NAEP website (at fourth grade, 22 percent of items from the main are on geometry versus 6 percent of items from the trend; at 8th grade, 26 percent versus 12 percent; and at twelfth grade, 23 percent versus 15 percent).[33] Since 1990, calculators have been provided for some items on the main NAEP, and geometric shapes are given to students for questions that ask them to determine the relationship between simple and complex figures.

The changes were made to bring NAEP in line with standards issued by the National Council of Teachers of Mathematics (NCTM) in 1989. The NCTM standards have been controversial, triggering what the media has dubbed "the math wars" between NCTM's advocates and critics across the country. The changes in NAEP mirror many of the flash points in the math wars—the increased emphasis on new topics at the expense of computation, the use of calculators, the reliance on manipulative materials to convey mathematical concepts. These elements have been especially controversial as NCTM prescriptions have been adopted into classrooms.

Regardless of the position one takes in the NCTM debate, it is important to know that "the nation's report card" consists of two tests that are apparently measuring two different kinds of mathematics. One can argue that both tests are doing their jobs. The main NAEP reflects the latest thinking in math education; the long-term trend anchors the test so that today's student performance can be compared to students of the past. There is still a lot of overlap in the tests, but no one should again read a headline about NAEP math scores going up or down without asking which NAEP scores are being reported.[34]

What are kids learning in math?

What kinds of math are students learning? To answer this question, we gathered data on how students performed on various items on the long-term trend. For each item, the trend reports the math topic assessed and the percentage of students answering the item correctly. What kinds of math problems are students missing and answering correctly? If students were performing relatively better on NCTM-endorsed topics in 1996, this would add to the evidence that differences in content distinguished the two tests after 1990.

The analysis suggests that the main test is more oriented toward NCTM-like topics (geometry and problem solving) and the trend more toward pre-NCTM topics (arithmetic).

We clustered items by math topic and computed correct response rates for each cluster. Geometry and problem solving stand out as areas in which all three age groups improved (see Table 4). From 1990–

Change in Student Achievement on Math Clusters (1990–1996) — **Table 4**

Age 9

Cluster Type	Gain	Loss
Geometry (4 items)	+6	
Problem Solving (3 items)	+3	
Data Analysis (13 items)	+2	
Addition of Whole Numbers (7 items)	+2	
Division of Whole Numbers (4 items)	+2	
Multiplication of Whole Numbers (4 items)		-1
Subtraction of Whole Numbers (6 items)		-1

Age 13

Cluster Type	Gain	Loss
Integers (3 items)	+8	
Percents (14 items)	+7	
Problem Solving (2 items)	+7	
Algebra (6 items)	+6	
Decimals (7 items)	+3	
Geometry (12 items)	+3	
Data Analysis (10 items)	+2	
Addition of Whole Numbers (6 items)		-1
Subtraction of Whole Numbers (6 items)		-2
Fractions (4 items)		-3

Age 17

Cluster Type	Gain	Loss
Geometry (14 items)	+5	
Square Roots (2 items)	+5	
Problem Solving (2 items)	+4	
Integers (7 items)	+3	
Data Analysis (6 items)	+2	
Percents (15 items)	+1	
Algebra (7 items)	+1	
Decimals (9 items)		-1
Convert Decimals to Fractions (2 items)		-2
Multi-Step Problem Solving (3 items)		-5
Fractions (3 items)		-13

Percent of Students Correct on Math Clusters (1990–1996)

Table 5

Age 9

Cluster Type	1990	1996
Addition of Whole Numbers (7 items)	79%	81%
Subtraction of Whole Numbers (6 items)	76%	75%
Application (7 items)	69%	69%
Data Analysis (13 items)	67%	69%
Multiplication of Whole Numbers (4 items)	65%	64%
Division of Whole Numbers (4 items)	61%	63%
Problem Solving (3 items)	50%	53%
Geometry (4 items)	22%	28%

Age 13

Cluster Type	1990	1996
Addition of Whole Numbers (6 items)	94%	93%
Subtraction of Whole Numbers (6 items)	91%	89%
Data Analysis (10 items)	84%	86%
Decimals (7 items)	61%	64%
Geometry (12 items)	55%	58%
Algebra (6 items)	51%	57%
Fractions (4 items)	57%	54%
Problem Solving (2 items)	47%	54%
Integers (3 items)	36%	44%
Percents (14 items)	36%	43%

Age 17

Cluster Type	1990	1996
Addition of Whole Numbers (3 items)	97%	97%
Integers (7 items)	76%	79%
Decimals (9 items)	78%	77%
Data Analysis (6 items)	73%	75%
Problem Solving (2 items)	68%	72%
Geometry (14 items)	62%	67%
Percents (15 items)	64%	65%
Fractions (3 items)	76%	63%
Square Roots (2 items)	53%	58%
Algebra (7 items)	46%	47%
Convert Decimals to Fractions (2 items)	39%	37%
Multi-step Problem Solving (3 items)	30%	25%

1996, nine year olds made their greatest gains in geometry and problem solving and notched small increases in data analysis and addition and division of whole numbers. Thirteen year olds racked up impressive gains in several areas: integers, percents, problem solving, algebra, decimals, geometry, and data analysis. Troubling, however, is a drop in thirteen year olds' performance with fractions, which should be learned by fifth grade. Seventeen year olds scored solid gains in geometry, square roots, problem solving, integers, and data analysis. But the decline that thirteen year olds evidenced with fractions shows up as even more severe with seventeen year olds.

Areas of greatest gain are also students' weakest areas of performance, most notably with the two youngest age groups (see Table 5). Nine year olds registered impressive gains on geometry items, for example. But their knowledge of geometry remains abysmal. On the four items tapping knowledge of geometry, the correct response rate in 1996 was 28 percent. The four areas of greatest gain for thirteen year olds (integers, percents, problem solving, and algebra) are four out of their five weakest areas.

Table 5 delivers some discouraging news about American students' knowledge of arithmetic, the most fundamental branch of mathematics. Addition and subtraction with whole numbers appear to be under control by age thirteen. But those are the only areas of arithmetic where mastery is indicated at thirteen, just as students stand at the threshold of high school. The percentage of students correctly answering items is disappointingly low for decimals (64 percent), fractions (54 percent), integers (44 percent), and percents (43 percent). The cluster called "algebra" (57 percent) tests such rudimentary skills as adding monomials (e.g., 6n + 9n = ?), a concept usually covered in a pre-algebra class.

On this test, the average thirteen year old performs at about a 50 percent proficiency level on fractions, decimals, percents, and integers. Seventeen year olds score about 70 percent on the same content. That isn't good. About one-fourth of students take algebra in eighth grade. Many people now call for *all* eighth graders to take algebra. But students must first learn arithmetic, including a thorough mastery of fractions, decimals, integers, and percents. The mastery of arithmetic is non-negotiable. This is not because those who insist upon it are pre-historic thinkers. It's simply the way math works. In math, if you don't learn arithmetic, you are not only incapable of using mathematics in everyday life, but you also can't possibly move on to learn algebra and other advanced mathematics. Students who try to learn higher math without a solid grounding in arithmetic might be able to memorize formulas—they might even be able to go through the motions on some mathematical procedures—but they'll never gain a deep conceptual understanding of the subject.

What are the policy implications?

The policy implications are crystal clear. Calling for all eighth graders to take an algebra course is putting the cart before the horse. A more sensible goal is for all students to master arithmetic by the end of eighth grade, if not before. National goals should focus on learning. The courses students take are merely means to that end. President Clinton's recent call for all students to be able to read by the end of third grade has moved the nation. Thank goodness he didn't declare a goal of enrolling all students in a reading course by the end of third grade. It is time for a national commitment to all students learning arithmetic. Only if students master arithmetic, can learning algebra follow.

Part III

POLICIES AND PRACTICES AFFECTING ACHIEVEMENT

— Calculator Use

— Exemplary Schools

CALCULATORS: DO THEY HELP OR HURT MATH ACHIEVEMENT?

The use of calculators in elementary school classrooms generates intense debate. The National Council of Teachers of Mathematics (NCTM) first expressed its support for calculators in 1974. It reissued the endorsement in 1980, calling for schools to "introduce calculators and computers into the classroom at the earliest grade practicable."[35] In 1989, the NCTM recommended that calculators be used in grades K–4, admonishing schools, "clearly, paper and pencil computations cannot continue to dominate the curriculum." In 1990, the National Research Council issued "Reshaping School Mathematics," a report that urged "the replacement of most paper-and-pencil drills with calculator-based instruction" starting in kindergarten. Because calculators "diminish the role of routine computations," the report advised, "young children can instead be given activities with calculators that emphasize discovery and exploration."[36]

Critics of calculators believe they may impede learning, especially when used by students who haven't memorized basic facts (for example, 2+2, 6x7, 14-9) or learned how to add, subtract, multiply, and divide on paper. The risk is that calculators will become a crutch for students. Worse yet, young children may never acquire a deep understanding of how numbers work if, on first exposure to mathematical operations, they merely push buttons to arrive at answers.[37] Surveys show that professors in schools of education believe calculators should be used more often in teaching math. But teachers want them used less, and a large majority of the public thinks that they shouldn't be used at all with young children.[38]

What does the research say?

Research thus far hasn't resolved the dispute. Suydam's 1979 review of the literature concluded that calculators do not undermine basic skills. Two meta-analyses of the research give qualified support for calculators. Hembree and Dessart (1985) studied 79 research reports from programs implemented in grades K–12. They found that using calculators had neither a positive nor negative effect on the paper-and-pencil skills of low- and high-ability students. For average students, the effect was positive in most grades, but negative for fourth graders. Hembree and Dessart concluded that for grades other than fourth, calculators positively affect math achievement.[39]

Brian A. Smith (1997) reviewed twenty-four studies published from 1984 to 1995 and also found a positive effect for calculators. Few of the studies involved students at the fourth grade or lower, however, and none of these examined calculators' effect on the acquisition of computation skills. Smith limits his recommendations accordingly. He cautions that calculators should only be used "on a limited basis" in the elementary grades, for exploratory purposes and problem-solving activities.[40]

Advocates of calculators would be more persuasive if the calculator studies were of higher quality.

Meta-analysis is excellent for summarizing research findings but is limited by the quality of original studies. The literature on calculators supports their use. Advocates of calculators would be more persuasive, however, if the calculator studies were of higher quality. Most of the studies were short in duration, lasting only a few weeks, and lacked sufficient controls to equalize comparison groups or to screen out other influences on student outcomes. Some of the studies trained teachers in the experimental groups (those with calculators) but not in the controls. Others allowed students in the non-calculator group to be aware that students in other classrooms were using calculators during the experiment. Letter grades or scores on teacher-made tests sometime served as outcome variables. This does not instill confidence in the studies' findings.

Are test scores related to calculator use?

The test results cited in this report, from NAEP and TIMSS, provide an interesting perspective on the calculator issue. On both tests, students are asked how often they use calculators in class. And on both tests, calculator use is correlated with lower math scores. Nine year olds who report that they use calculators in class every day have the lowest NAEP scores of any response category (see Figure 8). Students who use calculators only once or twice per month have the highest scores. A similar pattern is evident on TIMSS (see Table 6). Frequent calculator use is negatively correlated with math achievement in several countries. A vast majority of students in the highest-scoring nations (Japan, Singapore, Korea) report that they never use calculators in math class.

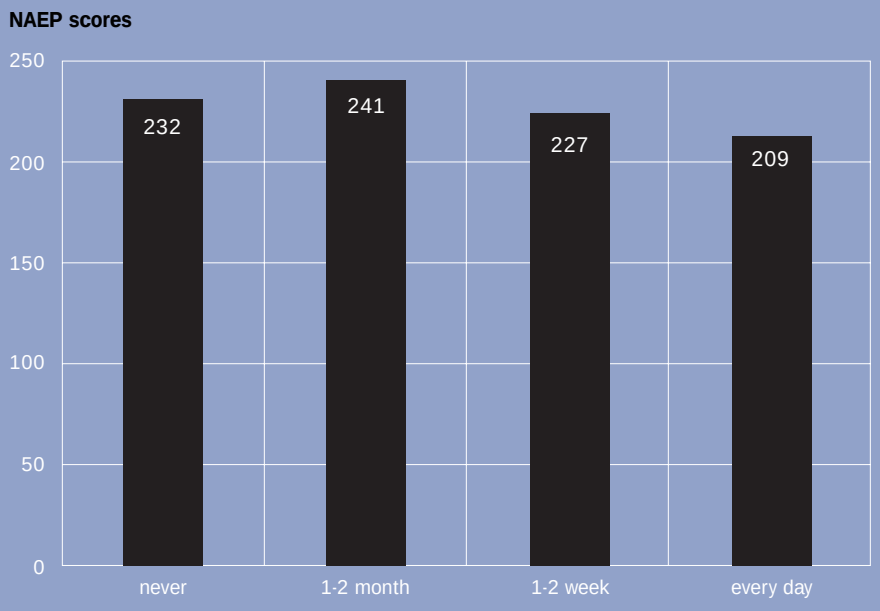
Causality cannot be inferred from these data. Low student achievement may just as

Everyday calculator use is correlated with lower test scores

Math scores as measured by the National Assessment of Educational Progress (NAEP), 1996. Students in grade 4 were asked, "How often do you use a calculator for math classwork?" NAEP data are expressed as scale scores, ranging from 0 to 500.

Fig 8

Fourth graders who use calculators every day score the lowest on NAEP math tests. But the data do not prove causality.



Low student achievement may just as easily "cause" calculator use as the other way around.

easily "cause" calculator use as the other way around. Imagine a teacher facing mandates that students know how to convert fractions to decimals and solve multi-step problems using percents. But on the first day of school, the teacher finds that students can't even add or subtract whole numbers. Out come the calculators.

Great care must be taken when interpreting studies that try to gauge the effects of an educational practice without taking into account students' initial test scores. Evaluated inappropriately, classroom practices intended to be compensatory can appear harmful. Teachers may use a computer drill and practice program with low-achieving students, for example, but when these students' low test scores are compared to those of students using computers for different purposes, it does not mean that the drill and practice software "caused" the low achievement.

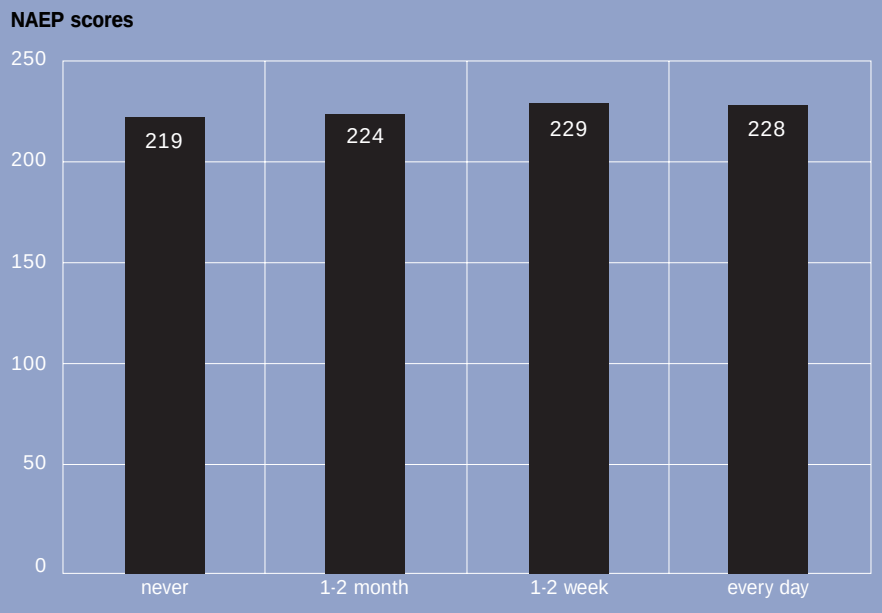
Two more important statistics are found in the NAEP data. First, the negative correlation between calculator use and NAEP scores evaporates when teachers are asked about the frequency of calculator use (see Figure 9). Teachers who say that their students use calculators every day have students with high test scores. What's going on? Doesn't this contradict the student data in Figure 8? Not necessarily. Teachers were asked, "How often do your students use calculators in class?" Only about 5 percent of teachers responded "every day." Teachers may have estimated how often the class as a whole uses calculators in a lesson. But 33 percent of students answered "every day" to the question, "How often do you use a calculator for math classwork," which could include independent seat work, group activities, or even games. Also consider that if a teacher reports a frequency of once or twice per month, all of the students are coded that way, even though a significant portion of the class might use calculators more often.

Teacher reports of calculator use show a different story

Math scores as measured by the National Assessment of Educational Progress (NAEP), 1996. Fourth grade teachers were asked, "How often do students use a calculator?" NAEP data are expressed as scale scores, ranging from 0 to 500.

Fig 9

Teachers who say that their fourth grade students use calculators every day have students with high test scores. This contrasts with the trend in the student-reported data (Figure 8).



A second statistic pertains to equity. African American and Hispanic students are about twice as likely as white students to report that they use calculators every day (see Figure 10). With daily calculator use also associated with lower math scores on the NAEP, this raises a troubling new perspective on the "digital divide" that deserves serious attention. We need to test and verify the benefit of new technologies before they become central elements of classroom practice. Providing access to new technologies, only to learn later that they hinder learning, does not advance the cause of educational equity.

What are the policy implications?

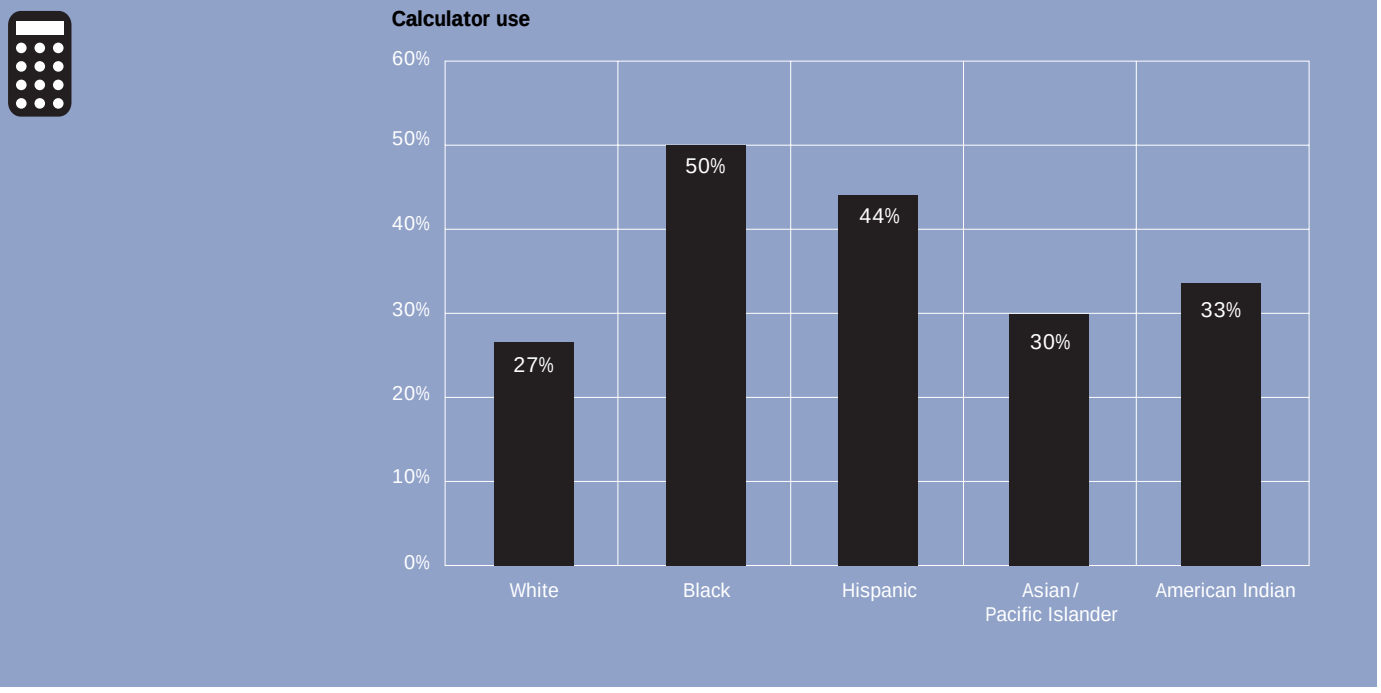
More information is needed to sort out the cross-currents in the NAEP data on calculator use. And, as already pointed out, more high-quality research needs to be conducted on this topic before anyone can declare with much confidence that calculators are helpful or harmful in learning mathematics. The National Science Foundation has financed the creation of math programs, later endorsed by the Department of Education, that promote calculator use in the early grades. Little is known about the impact of calculators on the basic computation skills of children below fourth grade. Both agencies should officially adopt a neutral stance on the question and support the kind of scientifically sound research that could objectively evaluate the claims of both calculator advocates and their critics.

The NAEP testing procedures warrant an additional comment. As mentioned in the second section of this report, calculators are provided to students for a portion of the

Black and Hispanic students are more likely to use a calculator every day

Students reporting everyday calculator use for math classwork in grade 4 (NAEP), 1996.

Fig 10



Frequency of Calculator Use in Math Classwork
(Student Responses, TIMSS 1994–95, Grade 4)

Table 6

Never

Country	% Students	Mean Score
Australia	25	545
Canada	51	532
England	15	510
Hong Kong	95	593
Israel	24	522
Japan	89	602
Korea	93	616
Netherlands	90	579
New Zealand	18	495
Norway	89	510
Singapore	96	634
United States	34	534

Some Lessons

Country	% Students	Mean Score
Australia	67	556
Canada	43	546
England	74	524
Hong Kong	3	492
Israel	60	541
Japan	11	561
Korea	5	579
Netherlands	10	592
New Zealand	61	512
Norway	8	498
Singapore	3	511
United States	53	565

Most Lessons

Country	% Students	Mean Score
Australia	8	512
Canada	6	493
England	11	474
Hong Kong	2	-
Israel	16	525
Japan	1	-
Korea	2	-
Netherlands	0	-
New Zealand	21	475
Norway	3	429
Singapore	1	-
United States	13	507

main and state NAEP tests. The NAEP website posts sample items, and a few that fourth graders take are listed below. They are startling in their simplicity. Research is needed to address the following questions: Why do fourth graders need to use calculators on these items? What skills or knowledge are being measured when students are allowed to use calculators on such simple math problems? How are main NAEP scores affected by calculator use, and have calculators contributed to the main NAEP scores' divergence from the long-term trend scores? [41]

1. Kitty is taking a trip on which she plans to drive 300 miles each day. Her trip is 1,723 miles long. She has already driven 849 miles. How much farther must she drive?
   (A) 574 miles (C) 1,423 miles
   (B) 874 miles (D) 2,872 miles

   Did you use the calculator on this question?
   ☐ Yes ☐ No

2. A whole number is multiplied by 5. Which of these could be the result?
   (A) 652 (C) 526
   (B) 562 (D) 265

   Did you use the calculator on this question?
   ☐ Yes ☐ No

3. Every hour, a company makes 8,400 paper plates and puts them in packages of 15 plates each. How many packages are made in one hour?
   (A) 560 (C) 17,857
   (B) 8,385 (D) 126,000

   Did you use the calculator on this question?
   ☐ Yes ☐ No

4. Martha planted 32 seeds. She put 8 seeds in each row.
   How many rows did she plant?
   Which of the following could Martha use to solve the problem correctly?
   (A) 32 + 8 (C) 32 x 8
   (B) 32 - 8 (D) 32 / 8

   Did you use the calculator on this question?
   ☐ Yes ☐ No

ARE "EXEMPLARY SCHOOLS" TRULY EXEMPLARY? The federal government's Blue Ribbon Schools Program (BRSP) was launched in 1982. Secretary of Education Terrence Bell explained in *Education Week*, "We are not setting out to find the best schools in America. We are simply looking for distinguished schools that are doing an exceptionally fine job."[42]

Since then nearly 4,000 schools have been selected, receiving a flag that they proudly wave as a sign of national recognition. The awards alternate—elementary schools one year, middle and high schools the next.

The selection process has several stages. Public schools apply to their state education agency, private schools to The Council for American Private Education. These agencies then screen applicants and forward nominations to the U.S. Department of Education. The Bureau of Indian Affairs and Department of Defense also nominate schools under their authority. Some states nominate candidates for the national award in conjunction with their own school recognition program. The Department of Education convenes the National Review Panel to evaluate the quality of the schools as represented in the application packets. About half of the applicants are selected for site visits to confirm information in the application. For the 2000 Blue Ribbons, 198 out of 202 visited schools won an award.[43]

In recent years, the Department of Education has characterized the BRSP as a catalyst for self reflection, planning, and goal setting, all in an effort to support local school reform. The Department also advertises the program as encouraging schools to follow "best practices," educational strategies that are solidly grounded in research. These objectives are blended with the notion of educational excellence to produce an identity for the program that is summarized on the BRSP website (www.ed.gov/offices/OERI/BlueRibbonSchools) as follows:

> Since 1982 the Blue Ribbon Schools Program has celebrated many of America's most successful schools. A Blue Ribbon flag waving overhead has become a trademark of excellence, a symbol of quality recognized by everyone from parents to policymakers in thousands of communities.
>
> The emerging secret of the Blue Ribbon Schools Program is its power to stimulate and focus school improvement initiatives. "The Blue Ribbon nomination package pulls together what is cutting edge in education today," says one educator. "The school that goes through the process is examining itself in terms of what works in the best schools in the country."
>
> "Regardless of the direction you're going with in school improvement, the Blue Ribbon program gives you a vehicle to get on track. It gives you a framework and standards so you know where you stand," says one principal. Schools are finding that the richness and scope of the Blue Ribbon nomination process allows them to reflect, "not just on the surface level, but down deep." One educator says, "If you want a tool for

The bottom line is this: about one-fourth of these schools can honestly claim that their Blue Ribbon stands for academic excellence.

Achievement of 1999 Blue Ribbon Elementary Schools
(Number of Schools)

Table 7

State	Total	Top 10%	Bottom 50%
California	39	12	9
Pennsylvania	10	4	0
Michigan	9	1	3
Indiana	4	0	2
Washington	4	1	1
Illinois	3	1	1
New Mexico	1	0	1
Total	70	19	17

NOTE: Test scores from 1998–99 school year, adjusted for socioeconomic status (SES). Public schools only. Blue Ribbon awards given in 1999.

Achievement of Federal Blue Ribbon Schools in Pennsylvania (Number of Schools by Decile)

Table 8

Decile	Elementary Schools	Middle Schools	High Schools
10th	4	2	2
9th	5	1	1
8th	1	0	1
7th	0	0	0
6th	0	0	0
5th	0	0	0
4th	0	0	1
3rd	0	0	0
2nd	0	0	1
1st	0	0	0
Total	10	3	6

NOTE: Adjusted test scores (PSSA) from 1998-99 school year, which state officials compute to compare schools of similar socioeconomic status (SES). Public schools only. Blue Ribbon awards given in 1998 (middle schools and high schools) and 1999 (elementary schools).

school improvement, there's nothing out there better than the Blue Ribbon Schools Program. It's the best you can find."

How do Blue Ribbon Schools stack up?

Notwithstanding the fact that Blue Ribbon Schools are selected on criteria that include but are not restricted to achievement, how do they compare on state tests in reading and math? One would not expect winners to be the highest achieving schools, perhaps, but certainly schools waving the BRSP's "trademark of excellence" should be near the top in academic achievement.

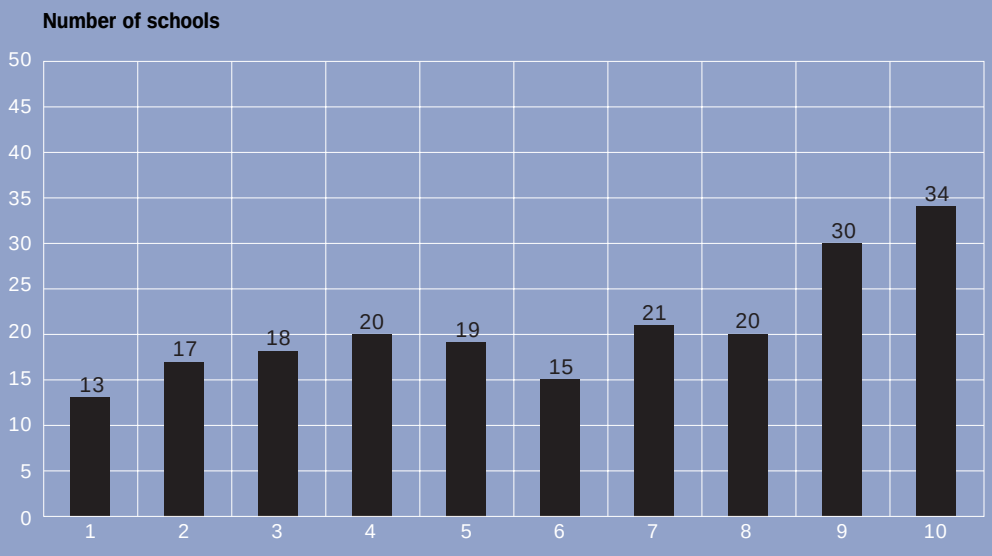
The following analysis examines the academic performance of Blue Ribbon schools in seven states, using the results from each state's test of reading and math achievement. To put everyone on an equal playing field, we statistically adjusted scores to compare schools serving students of similar income levels (that is, we used an SES-adjusted score). This is standard practice when comparing school test scores. Finding out that Blue Ribbon schools in poor neighborhoods score lower than the average school in wealthy neighborhoods is hardly surprising. In the analysis, if the state had already computed SES-adjusted scores for schools, we used the state's adjusted score. If it hadn't, we computed our own using the percentage of each school's students participating in the free and reduced lunch program (an income-based program). The important thing to know is that the following analysis compares Blue Ribbon schools to schools that serve students of similar socioeconomic levels.[44]

The results are striking (see Table 7). Only nineteen of these seventy Blue Ribbon elementary schools score in the top 10 percent

California's Distinguished Schools aren't necessarily high achievers

Fig 11

More than a third of California's 1998 award-winning elementary schools scored below average for schools of similar demographic characteristics.



California computes a "Similar Schools Rank," which compares schools of similar demographic characteristics and ranges from 1 (lowest performing) to 10 (highest performing). Test scores are from SAT-9 in the 1998–99 school year.

of similar schools in their respective states. Seventeen schools score in the bottom 50 percent, meaning their students score lower on reading and math tests than the average school with a similar population. And the remaining thirty-four schools score somewhere between these two groups, from the sixth to the eighth deciles, above average but not particularly outstanding. The bottom line is this: about one-fourth of these schools can honestly claim that their Blue Ribbon stands for academic excellence. Half of the schools can claim above-average academic standing, but not at the highest levels of performance. For the final one-fourth of schools, the Blue Ribbon may stand for an educational quality that is deserving of honor, but it is a quality quite distinct from superior academic achievement.

Let's look more closely at a few states' award winners, with an eye toward discerning the priority state officials give to achievement before they send nominees on to the federal level. Pennsylvania's selection process appears tilted toward high achievement, as its Blue Ribbon Schools are concentrated in the top two deciles of performance (see Table 8). Nine out of ten Blue Ribbon elementary schools, all three middle schools, and three out of the six honored high schools rank among the top 20 percent of similar schools in the state.[45] Michigan's award winners, on the other hand, include three elementary schools that score slightly below average, in the fourth and fifth deciles (see Table 9). These schools are located in relatively wealthy neighborhoods. Fewer than 10 percent of their students are on free lunch. Even though the schools' scores aren't bad—72 percent, 73 percent, and 77 percent of students score at a satisfactory level on Michigan's test—there are other Michigan schools with the same socioeconomic profile, with 95 percent or more

Achievement of Federal Blue Ribbon Schools in Michigan (Number of Schools by Decile)

Table 9

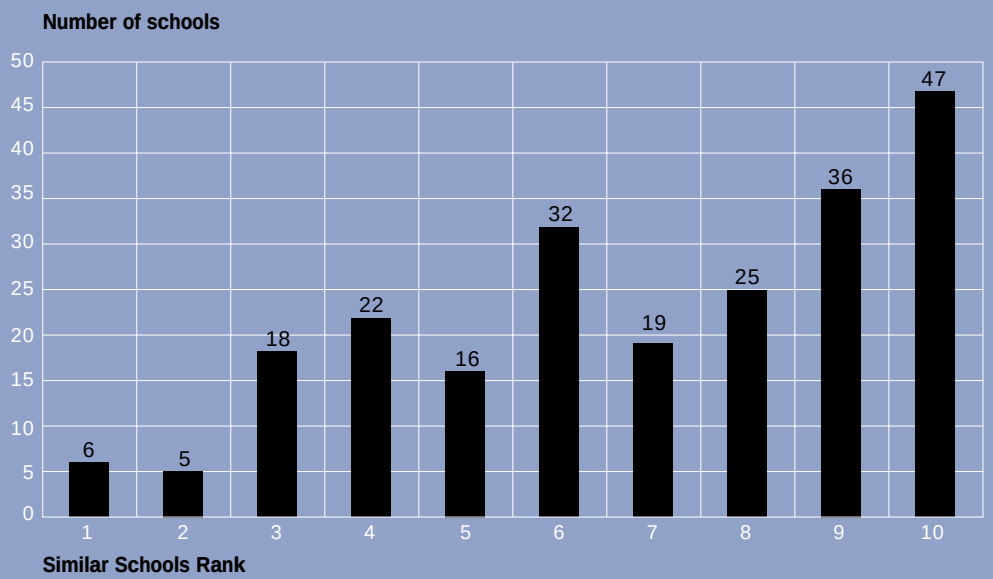
Decile	Elementary Schools	Middle Schools	High Schools
10th	1	0	0
9th	1	1	1
8th	1	0	0
7th	1	0	0
6th	2	0	0
5th	1	0	0
4th	2	0	0
3rd	0	0	0
2nd	0	0	0
1st	0	0	0
Total	9	1	1

NOTE: Adjusted test scores (MEAP, HSPT) from 1998-99 school year, which we computed to compare schools of similar socioeconomic status (SES). Public schools only. Blue Ribbon awards given in 1998 (middle schools and high schools) and 1999 (elementary schools).

In 2000, high achievement still wasn't necessary in California's program

Fig 12

California's 2000 award-winning elementary schools are slightly higher achievers than the 1998 Distinguished Schools.



California computes a "Similar Schools Rank," which compares schools of similar demographic characteristics and ranges from 1 (lowest performing) to 10 (highest performing). Test scores are from SAT-9 in the 1998–99 school year.

Achievement of Federal Blue Ribbon Schools in California (Number of Schools by Similar Schools Rank)

Table 10

Similar Rank Schools	Elementary Schools	Middle Schools	High Schools
10th	12	0	4
9th	10	0	2
8th	5	1	2
7th	1	2	2
6th	2	0	3
5th	4	1	1
4th	4	1	0
3rd	1	0	1
2nd	0	0	0
1st	0	0	1
Total	39	5	16

NOTE: California's Similar Schools Rank ranges from 1 (lowest performing) to 10 (highest performing). Public schools only. Blue Ribbon awards given in 1998 (middle schools and high schools) and 1999 (elementary schools). Rank based on SAT-9 scores for 1998–99 school year.

of students performing at the satisfactory level, and with no Blue Ribbon.[46]

How about state recognition programs?

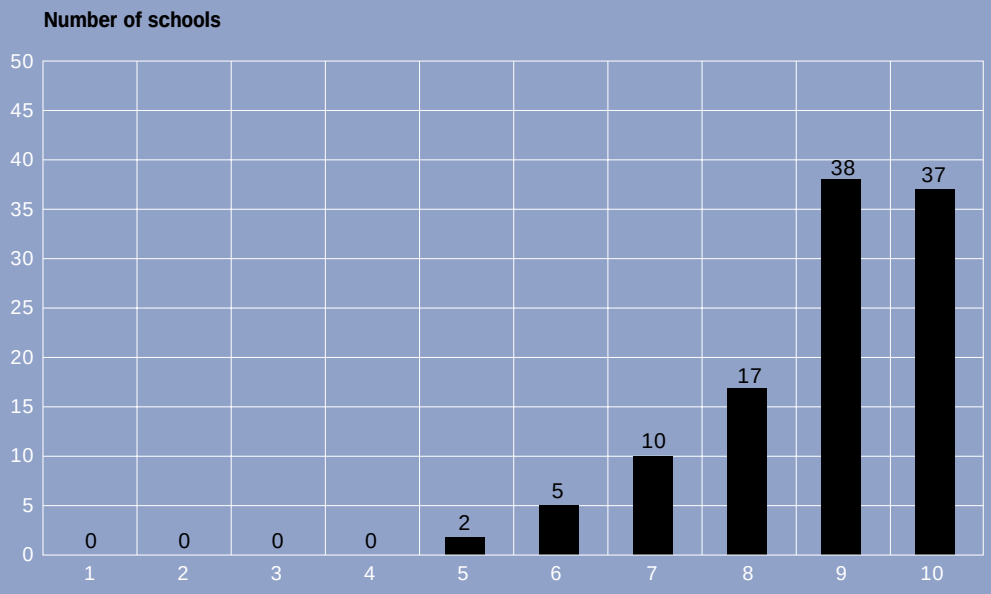
California runs its own exemplary school program in coordination with the BRSP. The state annually identifies 5 to 10 percent of schools as Distinguished Schools and presents each school with a special flag and plaque at a statewide awards ceremony. Governor Gray Davis recently proposed tying the awards to performance on the state's testing program, with scores adjusted for student background (comparable to the analysis here). The process that selected the 2000 winners dates back to 1985. It relies on a rubric of preferred practices, "designed to reflect the consensus of the education community regarding quality education." Using the rubric, a panel of experts screened applicants on criteria that included: standards and graduation requirements, leadership, curriculum and instructional practices, support services, technology, professional development, parent and community involvement, and school safety.[47]

Rewarding schools based on a rubric inevitably favors what schools do over what schools accomplish. The academic performance of the 1998 California Distinguished Schools (elementary schools) illustrates the point (see Figure 11). The winning schools were all over the map on the state's 1999 tests. California uses its own ranking system, from one to ten and called the "Similar Schools Rank," which compares schools of similar demographic characteristics. Almost one-third fell in the top two deciles of achievement, but a surprising thirty schools placed in the lowest two deciles. High achievement was obviously not a strict qualification for becoming a Distinguished School. This policy carried

Indiana's Four Star Schools are overwhelmingly high achievers

Fig 13

About 70% of Indiana's Four Star Schools in 2000 scored in the top two deciles of academic performance. In contrast to California, only 2 of 109 schools scored below the average for schools of similar demographic characteristics.



Based on ISTEP+ scores for 1998–99 school year.

over to the state's Blue Ribbon schools as well (see Table 10).

The intent of Governor Davis to emphasize achievement in California's new state program should change this pattern. The 2000 Distinguished Schools still include a large number of low achievers, but the group as a whole has slightly higher test scores than the 1998 winners (see Figure 12). Contrast California's award winners to Indiana's, which seems to stress high academic achievement in its school recognition program. Almost 70 percent of its Four Star Schools achieved in the top two deciles of academic performance on the state's 1999 test of academic skills (see Figure 13).

What are the policy implications?

As pointed out earlier in this report, many states are now rewarding and sanctioning schools based on test scores in academic subjects. School recognition programs, whether at the federal, state, or local levels, should be changed to acknowledge the era of high standards and strong academic emphasis that American education has entered. Under current procedures it is theoretically possible for a school to receive a Blue Ribbon Schools award from the federal government at the same time it is placed on academic probation by state authorities. To become a model for states to follow, the federal Blue Ribbon Schools Program should be reformed along the following lines:

1. Eliminate the self-selecting application process, which encourages self-promotion and all-out campaigns for the award. Federal officials should collect achievement data from the states, objectively screen for the best nominees using technically valid procedures, and make the appropriate statistical adjustments to ensure that schools serving different student populations are treated fairly. The current system gives schools with high inputs (schools in wealthier communities whose students are high scoring from day one) an advantage.
2. Drop the rubrics. The practices included in the Blue Ribbon application packet are indeed "cutting edge," but that can be a problem. Many have not been rigorously tested or verified by high-quality research as contributing to student achievement. Schools that use unpopular or out-of-fashion approaches but manage to teach children how to read and to do mathematics well should receive a Blue Ribbon award before cutting-edge schools with mediocre levels of student learning. Judge schools by their accomplishments, not their practices.
3. Place high academic achievement front and center as the defining characteristic of an excellent school. In the current BRSP application packet for elementary schools, academic achievement is the last criterion discussed, eighth on the list of eight characteristics. Officials may want to award schools for other accomplishments—most improved in math, excellent professional development, increased attendance, effective parent outreach. But these awards should be labeled for the specific quality they are applauding. When the public hears of a group of schools being designated "exemplary" by a government program, people should be able to assume that those schools have attained levels of excellence in reading and mathematics.[48]

When the public hears of a group of schools being designated "exemplary" by a government program, people should be able to assume that those schools have attained levels of excellence in reading and mathematics.

THIS REPORT HAS ASSESSED THE DIRECTION of student achievement in the United States, evaluated the size and significance of gains and losses in achievement test scores, explored questions about the quality of information the American public is receiving on academic progress from state and federal testing programs, and analyzed two policies and practices associated with student achievement. The report reaches seven major conclusions:

- The academic achievement of American students has risen since the 1970s but only at a snail's pace. Gains in reading are exceedingly small; gains in mathematics are significant. Younger students, ages nine and thirteen, have made greater progress than seventeen year olds in both subjects.
- State tests confirm that achievement continued to rise from 1998 to 1999. States that write their own tests were more likely to report reading gains than those using commercial, off-the-shelf tests. The sample of states is very small, however, so more data are needed to reach any meaningful conclusions.
- It is unclear why test score gains are more difficult to accomplish with older students. Arguments that the middle-grade slump is an artifact of testing, caused by tracking, or exacerbated by a particular style of classroom instruction lack supporting evidence. The diminished status of academic achievement among American teenagers is the most persuasive explanation, and, unfortunately, subordinating achievement to other aspects of teen life is reinforced by schools, families, business, and public policy.
- A clear picture of national achievement in mathematics is complicated by the divergence of the two national NAEP tests—the long-term trend and the main—in the 1990s. The two tests appear to assess different mathematics, with the long-term trend NAEP placing greater emphasis on arithmetic and the main NAEP on geometry and problem solving.
- Student performance in geometry and problem solving improved in the 1990s. Performance in arithmetic remained static or declined slightly. Results for thirteen year olds suggest large numbers of students have not mastered the basic arithmetic skills that are necessary before moving on to algebra.
- Research generally favors the use of calculators in classroom instruction. However, little is known about the impact of calculator use on young children's learning of basic skills. More high-quality studies are needed in fourth grade and earlier. Fourth graders who say they use calculators every day score significantly lower on the NAEP math test than other fourth graders. The Department of Education and the National Science Foundation should adopt a neutral stance on the issue, especially given the cautionary signals in the federal government's own NAEP data.
- Schools designated as exemplary by federal and state awards programs are not always exemplary in academic achievement. High achievement should be the distinguishing characteristic of schools receiving such awards. Awards for other accomplishments should be labeled for the quality deserving honor.

ENDNOTES

1. Rounding to the nearest whole number was performed in the final step of calculations.

2. Frederick Mosteller, Richard J. Light, and Jason A. Sachs. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size," *Harvard Educational Review,* vol. 66, no. 4 (Winter, 1996), pp. 797–842.

3. Neal and Johnson estimate one year's worth of schooling to be equivalent to 0.22 to 0.25 standard deviation on the Armed Forces Qualification Test, about the same as 13–17 year olds on the NAEP. See Derek A. Neal and William R. Johnson, "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy*, vol. 104, no. 51 (1996), pp. 869–895.

4. Alan Krueger estimates the average rate of gain on all NAEP tests to be about 0.07 SD per decade. This is the same as the average rate of gain on the six math and reading tests that we analyze. Alan B. Krueger, "Reassessing the View that American Schools Are Broken," *Federal Reserve Board of New York Economic Policy Review* (March, 1998), pp.29–43.

5. To be included, states had to use the same test in the same subject area and grade level in 1998 and 1999.

6. We looked at 5th grade scores in the states testing that grade and found evidence of the slump already underway. Data posted on the Brown Center website at www.brookings.edu/browncenter.

7. Indiana reported custom and off-the-shelf scores. Indiana scores are reported for each test type in Figures 3 & 4. However, Indiana is counted only once in Tables 2 & 3.

8. Another argument is that off-the-shelf tests tend to be norm-referenced, whereas custom tests tend to be criterion-referenced. Since norm-referenced tests mathematically constrain the distribution of scores, they may cause the discrepancy in Figure 3. If that were the case, however, the same pattern should appear for mathematics scores. It doesn't.

9. "America's Education Choice," *The Economist*, (April 1, 2000), p. 17.

10. Gerald Bracey, "The TIMSS 'Final Year' Study and Report: A Critique," *Educational Researcher,* vol. 29, no. 4 (May, 2000), pp. 4–10.

11. Mark Reckase, "The Controversy over the National Standards Set by the National Assessment Governing Board," Paper presented at The Brookings Papers in Education Policy Conference, Brookings Institution, Washington, May 16, 2000, p.27.

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. William H. Schmidt, et al., *A Summary of Facing the Consequences: Using TIMSS for a Closer Look at United States Mathematics and Science Education.* (Kluwer Academic Publishers, 1998).

14. Robert E. Slavin, "Ability Grouping and Student Achievement in Elementary Schools: A Best Evidence Synthesis," *Review of Educational Research*, vol. 57 no. 3 (1987), pp. 293–336.

15. Data obtained from the following page on the NAEP website: www.nces.ed.gov/nationsreportcard/TABLES/index.shtml. For the 4th grade, refer to "NAEP 1996, 1992, and 1990 National Mathematics Results—Data Almanacs; Grade 4 Teacher Data," pp. 499, 511. For the 8th grade, refer to "NAEP 1996, 1992, and 1990 National Mathematics Results—Data Almanacs; Grade 8 Teacher Data," pp. 439, 445, 448, 451. Note that these scores are unadjusted for other factors that may influence achievement. My own study of two states tracking policies found detracking more prevalent in low income schools. If true nationally, this could produce a spurious correlation of higher test scores and tracked classes. See Tom Loveless. *The Tracking Wars, (*The Brookings Institution Press, 1999).

16. Ina V. S. Mulis, Michael O. Martin, Albert E. Beaton, Eugenio J. Gonzales, Dana L. Kelly, and Teresa A. Smith, *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's Third International Mathematics and Science Study (TIMSS),* Center for the Study of Testing, Evaluation, and Educational Policy, Boston College (1998), pp. 53–56; Schmidt, et al.

17. Obtained from the following link on the NAEP website: www.nces.ed.gov/nationsreportcard/TABLES/index.shtml. For the 8th grade, refer to "NAEP 1996 National Science Results—Data Almanacs; Grade 8 Teacher Data," p. 499.

18. Marilyn Binkley, Keith Rust, and Trevor Williams, eds., *Reading Literacy in an International Perspective,* U.S. Department of Education, (National Center for Educational Statistics, 1996).

19. Mulis, et al.

20. Schmidt, et al.

21. Geoffrey Howson, *Mathematics Textbooks: A Comparative Study of Grade 8 Texts.* TIMSS Monograph series, (Pacific Educational Press, 1995).

22. James W. Stigler and James Hiebert, *The Teaching Gap,* (Free Press, 1999).

23. Blase Masini, et al., "An HLM Estimation of School Productivity Effects in The First in the World Consortium," Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April, 2000.

24. The effects of instructional methods are reviewed in Herbert J. Walberg and Jin-Shei Lai, "Meta-Analytic Effects for Policy," *Handbook of Educational Policy,* (Academic Press, 1998), pp. 419–453. Also see Herbert J. Walberg, "Syntheses of Research on Teaching," *Handbook of Research on Teaching.* 3d ed., (Macmillan Publishing Company, 1986), pp. 214–229.

25. Stigler and Hiebert, *The Teaching Gap*.

26. U.S. Department of Education, National Center for Education Statistics, *Pursuing Excellence,* (NCES 97–198, U.S. Government Printing Office, 1996).

27. Erling E. Boe, Henry May, Christine S. Leow, and Gema Barkanic, "The Rise and Fall of National Performance in Mathematics and Science: Changes in Relative Standing from the Fourth Grade to the Final Year of Secondary School," Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, April 2000.

28. Boe, et al., "The Rise and Fall of National Performance in Mathematics and Science: Changes in Relative Standing from the Fourth Grade to the Final Year of Secondary School." Also see John Bishop, "The Effect of National Standards and Curriculum-Based Exams on Achievement," Working Paper #97–01 (Center for Advanced Human Resource Studies, Cornell University, 1997).

29. Laurence Steinberg, *Beyond the Classroom*, (Simon and Schuster, 1996).

30. National Assessment Governing Board (NAGB), *Mathematics Framework for the 1996 National Assessment of Educational Progress,* U.S. Department of Education, 1996.

31. Peter Applebome, "National Tests Show Students Have Improved in Math," *New York Times*, February 28, 1997, p. 15. Linda Jacobson, *Education Week,* September 3, 1997, *Education Week* online archives: www.edweek.com/edsearch.cfm.

32. Finding appropriate comparison groups is a challenge because the main test is given to students of a particular grade and the trend test to students of a particular age. About 33 percent of nine year olds are below the fourth grade and have never been exposed to fourth grade material. If below-grade-level students are included in the comparisons, the gap between the two tests shrinks at all three grade levels, but the report's substantive conclusions remain the same.

33. Refer to the following NAEP website: nces.ed.gov/nationsreportcard/TABLES/index.shtml. Follow the link to "Test Question Data" for the 1996 NAEP main and 1996 NAEP long-term trend.

34. A move to eliminate the trend test was recently defeated by NAGB. See David J. Hoff, "Test-Governing Panel Contemplates Halting Trend-Data Collection," *Education Week*, March 31, 1999, *Education Week* online archives.

35. Susan Walton, "Add Understanding, Subtract Drill," *Education Week*, July 27, 1983, *Education Week* online archives.

36. Quoted in Robert Rothman, "Guide for Implementing New Math 'Vision' Issued," *Education Week*, February 21, 1990, *Education Week* online archives.

37. Mark Clayton, "Calculators in Class: Freedom from Scratch Paper or 'Crutch,'" *Christian Science Monitor,* May 23, 2000, available online in the *Christian Science Monitor* archives:http://www.csmonitor.com/.

38. Only 38 percent of education professors think calculators will hamper the learning of basic arithmetic, Steve Farkas and Jean Johnson, *Different Drummers: How Teachers of Teachers View Public Education (*New York: Public Agenda, 1997), p.11. But 86 percent of the public rejects the use of calculators in early grades, and 73 percent of teachers want students to memorize the multiplication tables and learn pencil-and-paper arithmetic before using calculators. See Steve Farkas and Jean Johnson, *Given the Circumstances: Teachers Talk about Public Education Today,* (New York: Public Agenda, 1996), p. 19.

39. M. N. Suydam, "The Use of Calculators in Pre-College Education: A State-of-the-Art Review." Columbus, OH: Calculator Information Center, 1979. (ERIC Document Reproduction Service No. ED 171 573). Ray Hembree and Donald J. Dessart, "Effects of Hand-Held Calculators in Pre-College Mathematics Education: A Meta-Analysis," *Journal for Research in Mathematics Education,* vol. 17, no. 2, (1986), pp. 83–99.

40. Brian A. Smith, *A Meta-Analysis of Outcomes from the Use of Calculators in Mathematics Education,* Texas A & M doctoral dissertation, December, 1996, p. 101.

41. Some of the questions may be designed to find out if students know when it is appropriate to use calculators.

42. "National News Roundup," *Education Week,* January 19, 1983, *Education Week* online archives.

43. Alan Richmond, "In the Age of Accountability, a Blue Ribbon Means a Lot," *Education Week*, May 24, 2000, *Education Week* online archives.

44. For each school, we first computed a composite achievement score by averaging math and reading scores. We then regressed the composite scores on each state's SES variable (usually the percentage of students in the free/reduced lunch program) and used the residuals (the amount each school scored below or above the expected value for a school of similar SES) as our SES-adjusted score. The adjustment lowers the relative ranking of Blue Ribbon Schools with SESs above the state mean (those serving students from wealthier families) and raises it for schools with SESs below the mean (those serving students from poorer families). Some states do not release scores of schools serving small numbers of students (fewer than 10) or schools where an inadequate percentage of total enrollment took the test. We dropped these schools from the analysis, as well as those schools on which we didn't have SES data.

45. Three schools serving grades K–4 were dropped from the analysis. Pennsylvania does not give the PSSA test in the 4th grade.

46. One elementary school was dropped from the analysis because there was no data on the Michigan Department of Education website.

47. See "1998 California School Recognition Program: Distinguished Elementary School Application Scoring Rubric," www.cde.ca.gov/ope/csrp.

48. Under current policy, schools qualify by accomplishing any of the following on a nationally-normed test: 1) scoring two-thirds of a standard deviation above the mean (about the 75th percentile); gaining one-third of a standard deviation in the previous five years (about 13 percentile points around the mean); the school's "majority" group scoring two-thirds of a standard deviation above the mean. See "National Review Panel: Elementary Scoring Guidelines, 2000–2001," Blue Ribbon Schools Program, Office of Educational Research and Improvement, U.S. Department of Education (June, 1999).

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: 202-797-6406 • Fax: 202-797-2973
www.brookings.edu/browncenter