Working Paper 06-1

Is the No Child Left Behind Act Working?

The Reliability of How States Track Achievement

PACE

Policy Analysis for California Education
University of California, Berkeley and Davis
Stanford University

2006

Executive Summary

Debate is well under way regarding the efficacy of the *No Child Left Behind* (NCLB) *Act,* including whether this bundle of federal rules and resources is prompting gains in student achievement. Spirited conversation will intensify as the Congress discusses how to adjust and reauthorize this ambitious set of school reforms. Both state and federal gauges of student achievement will inform this debate.

We first ask whether state testing systems provide an accurate and consistent indication of the share of fourth-grade students who are deemed "proficient" in reading and mathematics. We compare the states' own estimates against federal determinations, based on results from the National Assessment of Educational Progress (NAEP). We report, for the first time, on comparable state testing results and trends since 1992.

We detail, looking across 12 diverse states, a small improvement in the percentage of children achieving proficiency in *reading,* based on NAEP results between 1992 and 2005. But states estimated much higher shares of students reaching proficiency, compared with the NAEP results. We then detail how children made greater progress in *math* proficiency over this 13-year period. Yet again we discovered that state test results exaggerate the annual rate of improvement, compared with the federal NAEP results.

This historical disparity between state and federal estimates of proficiency levels has not previously been illuminated over this range of states. But the phenomenon is not new. The gap does not stem simply from NCLB's unintended incentive for states to set low cutoffs for defining which students are deemed proficient. Instead, we show that states have long claimed that a much higher share of students are proficient relative to NAEP results, even before NCLB created the incentive for states to set a low bar.

Second, we report on achievement change during the three school years following enactment of NCLB (in January 2002). We find that some states have maintained their apparent momentum in raising the percentage of fourth-graders proficient in math, while reading performance leveled-off or slipped in several states, as gauged by state and NAEP exams. Our analysis places earlier reviews of state test scores, post-NCLB, in the context of leveling NAEP scores after 2002.

Third, we find that two states with weak accountability systems prior to NCLB (Arkansas and Nebraska) did experience gains in math proficiency after enactment of NCLB but not in reading. We discuss how adjustments to federal reforms could help all states devise student assessment systems that yield more consistent benchmarks of children's achievement over time.

Working Paper 06-1

Is the No Child Left Behind Act Working?

The Reliability of How States Track Achievement

Bruce Fuller
Kathryn Gesicki
Erin Kang
Joseph Wright

University of California, Berkeley
2006

Policy Analysis for
California Education

Table of Contents

Is the No Child Left Behind Act Working?

The Reliability of How States Track Achievement

Backdrop

President Bush began to claim in 2004 that NCLB already had shown its effectiveness, spurring children to achieve at higher levels. During his weekly radio address to the nation Mr. Bush (2004a) said, "We have recently received test results that show America's children are making progress." The ambitious federal reforms had been signed into law just two years earlier. By early fall, as the election campaign was heating up, Bush (2004b) became even more upbeat. "We're making great progress. We're closing the achievement gap," he said in a speech delivered in King of Prussia, Pennsylvania.

Analysts scrambled to pinpoint the evidence on which the White House was basing its claims. The Administration cited one number: a nine-point gain in the share of fourth-graders deemed proficient in math, as gauged by the 2003 National Assessment of Educational Progress (NAEP), relative to proficiency levels in 2000. But the 2003 exam was administered during the first school year subsequent to Bush's January 2002 signing of NCLB. Math scores had begun their ascent back in the 1980s, likely buoyed by states' earlier accountability reforms (Loveless, 2003).

Our research group had charted state test scores from the late 1990s through 2004 for 15 large and medium-size states for which data were readily available. We could discern no consistent gains in reading scores since passage of NCLB (Fuller, 2004; PACE, 2004). A few days later, the U.S. Department of Education released its own collection of test scores from 40 states. Secretary of Education Rod Paige (2004) said, "The PACE authors would like readers to conclude that No Child Left Behind has failed—this on the basis of their flawed study." We never argued that NCLB had "failed," simply that the evidence to date did not substantiate the bold claims made by the Administration.

Rather sobering news arrived one year later when NAEP scores, stemming from the spring 2005 testing, were released. These results offered a glimpse into how the nation's students had performed over the three school years following enactment of NCLB. Reading scores among fourth-graders were flat, with 31 percent of the nation's children at or above proficient in 2002, 2003, and 2005 (NAEP, 2005). The share of eighth-graders proficient or above in reading had fallen two percentage points. The percentage of fourth-graders proficient in math continued to climb between 2003 and 2005, but math scores at the eighth grade had reached a plateau.

One bright spot, the comparatively stronger performance by black fourth-graders, was cause for celebration within the Administration. Bush said, "It shows there's an achievement gap in America that's closing" (Dillon, 2005a). But veteran Washington analyst, Jack Jennings, referring to earlier state-led efforts, said, "The rate of improvement was faster before the law.

There's a question as to whether No Child is slowing down our progress nationwide." Ross Wiener at the Education Trust said, "There's been a discernible slowdown in progress since '03" (Romano, 2005).

Most recently, the Education Trust analyzed test results from 32 states—again relying on the states' own definitions of proficiency—and observed apparent progress between 2003 and 2005 in reading and math at the elementary school level (Hall & Kennedy, 2006). Yet these analysts pointed out that the NAEP results tell a different story for recent years.

How to Assess the Efficacy of NCLB?

Some argue that it's still too early to assess the effects of Washington's massive school reform effort. Others emphasize that we must delve into the implementation steps operating across federal, state, and local levels to understand whether these ambitious policies are truly touching the motivation of teachers and students alike. One evaluation of NCLB implementation, led by researchers at the RAND Corporation, is beginning to illuminate action among these organizational layers (Stetcher, Hamilton, & Naftel, 2005).

Still, as the episodes of dueling test scores vividly portray, the political pressure to glean trends from state and federal NAEP results will persist, and this quest will intensify as Washington begins to review NCLB within the congressional reauthorization process.

Another pressing question is whether the federal reforms have somehow advanced the momentum of states' earlier accountability efforts, or at least have improved government's capacity to accurately track student progress. It also may be that NCLB is being felt in a few midwestern or southern states that were slow to implement standards-based accountability programs (e.g., Carnoy & Loeb, 2002).

We begin by reporting how state reading and math scores and corresponding NAEP results among fourth-grade students have moved since 1992. This allows the opportunity to see how state and NAEP results were moving prior to the 2002 inception of NCLB.

This analysis leads to the fundamental question of whether state test scores offer accurate gauges of student progress over time, especially when trying to infer discrete effects of federal NCLB reforms. One argument against relying on NAEP scores is that these exams are not aligned to any state's curricular standards. But the yawning gaps between federal and state test results—reported as the percentage of children "proficient" or above in reading and math—make it difficult to attribute these differences solely to the lack of alignment.

As we assembled state testing data, going back to 1992, we discovered a lack of institutional memory, disinterest or limited state capacity to track change over time, and fairly frequent changes in testing regimes. This included shifts in where states set the achievement bar (cut-point) in defining which children are deemed "proficient". This leads to interrupted time-series and, in the absence of the careful equating of scores, difficulty in relying on state tests to track achievement over time. We will discuss how improvements might be crafted by Washington and the states to advance confidence in state-level testing programs. We can learn much from the few states where test results track more closely to NAEP results.

We do *not* aim to determine the discrete effects of state or federal school accountability reforms. Our analysis simply offers an unprecedented set of test-score trends, contrasting state and federal gauges across 12 diverse states. Note also that we will distinguish between shares of fourth-graders deemed by state or federal NAEP authorities (*levels*), versus *trends* in the percentage-of-students-proficient that are reported over time.

Policy Challenge 1—Clarifying Learning Aims, Strengthening Assessment

The contemporary story of state assessment efforts begins with a telling irony, emerging in the wake of the Reagan Administration's 1983 report, *A Nation at Risk.* When Capitol Hill leaders asked Daniel Koretz at the Congressional Budget Office to track student performance since the post-war period, he came back with some good news, at least for younger students. After little progress in the 1950s and 1960s, third and fourth graders had shown steady improvement on the Iowa Test of Basic Skills (ITBS) through much of the 1970s.

Even SAT scores, after falling by a third of a standard deviation during the immediate post-war period, as the GI Bill propelled a more diverse range of students into higher education, had floated upward from the late 1960s forward (Congressional, 1986). This would foreshadow an equally remarkable shift in the demographic mix of students taking standardized tests in decades to come.

Still, Koretz rightfully emphasized that students taking the ITBS and the SAT were not representative of the nation's children. Nor did state testing regimes meet minimal criteria for yielding valid and reliable data on student achievement over time. To do this, state assessment systems would have to provide "annual or nearly annual scores," equate scores to make them comparable over time, and test comparable groups of students over time (Congressional, 1986:100). The fact that SAT results were not adjusted to take into account the influx of diverse GIs now taking the SAT was an obvious case in point. Koretz, in short, detailed how the nation had no assessment technology in place by which the progress of children and schools could be reliably tracked over time.

The rise of so-called *systemic reform*—emphasizing clearer learning standards and stronger pupil assessment—was to remedy these institutional shortcomings. Reform designers such as Michael Cohen (1990) at the National Governors Association, along with Mike Smith and Jennifer O'Day (1991) at Stanford University, began to articulate a state-led model of organizational change which called on states and districts to sharpen what children were to be learning, and to align assessment to transparent standards.

The restructured system was to focus with undistracted intensity on achievement *outcomes,* largely measured by state education departments, rather than remaining preoccupied with how to best mix school inputs and classroom practices. This focus on outcomes led to an unprecedented commitment to attaching high stakes to particular tests, including high school exit exams (Stetcher, 2002).

This streamlined school institution was to advance public accountability. State spending on education had increased by 26 percent in real, inflation-adjusted dollars between 1980 and 1987 and "policymakers are expressing increased concern over accountability, asking if investments made in previous years are paying off in terms of performance" (Cohen, 1990:254).

This version of "performance-based accountability" did not intend to narrow pedagogical practices through centralized regulation, nor dumb-down curricular content. Instead, Cohen proposed standards and forms of assessment that would advance complex forms of teaching and learning, citing Lauren Resnick's (1987) work in cognitive science. "Studies demonstrate that skilled readers, even at the early elementary grades, are able to comprehend what they read not simply because they have acquired the basics, but because they intuitively and automatically rely on what we think of as higher-order skills," Cohen said (261).

Several states in the 1980s already had engaged in curricular reform and policy efforts, nudging students to take more challenging courses (Blank & Schilder, 1991). But the new focus on state-crafted standards required expert panels to delineate what children should learn at each grade level, rather than only toughening course requirements or pushing uniform curricular packages.

Stronger links between state-set learning standards and new exams suggested the need to shift from norm-referenced to criterion-referenced testing (Elmore, Abelmann, & Fuhrman, 1996), a move made by several states beginning in the early 1990s, then incorporated into the No Child Left Behind Act. The NAEP governing board, long before, had chosen to report the share of students reaching basic, proficient, or advanced levels, beyond simply releasing average scale scores.

Policy Challenge 2—Politically Insulating Student Assessment

By the mid-1990s many governors and legislatures were advancing various renditions of accountability reform. A few states, such as California and Kentucky, did operationalize the original model of system reform, specifying how higher-order thinking skills and invigorating forms of pedagogy were to be advanced through new forms of testing.

State leaders, however, also understood that new state testing regimes must yield unambiguous benchmarks of student performance. This proved to be the more widely shared concern. Fully 48 states by 1999 had put in place statewide testing programs, assessing children in at least one grade in elementary, middle, and high schools (Goertz, Duffy, & Carlson Le Floch, 2001).

Policy makers traveling this path toward crisp learning aims and more transparent achievement data soon collided with several immovable obstacles in the road. Each continues to limit states' ability to devise stable testing systems which yield comparable benchmarks of achievement over time. Together, these constraints suggest that relying only on state-reported test results may paint an incomplete picture of students' actual progress.

First, while most states have succeeded in articulating clear learning objectives in basic subject areas, the specific domains covered in state tests vary dramatically. Once these domains are

decided upon within a state, similar test items may be used year after year, teachers may teach to a limited range of domains, or test items are quietly circulated among teachers (for review, Stetcher, 2002).

This leads to inflated results where higher shares of students surpass cut-points as they become more familiar with items that appear on the standardized exams. In addition, as states have moved to criterion-referenced testing, the cut-points for determining which students are deemed "proficient" are set at varying levels across states. Within a given state, cut-points also shift over time (Catterall, Mehrens, Ryan, Flores, & Rubin, 1998; Linn, 2001).

A portion of these factors likely explained how fourth-graders' reading performance during the first two years of the Kentucky Instructional Results Information System (KIRIS, 1992-1994) rose dramatically, by a stunning three-quarters of a standard deviation (Koretz & Barron, 1998). Student performance was *unchanged* for the same student cohorts on the NAEP reading assessment.

Kentucky actually varied its test forms year to year, so teaching-to-the-test did not fully explain the apparent inflation of results. The setting of cut-points may play a stronger role in such cases, where state officials deem a higher share of students are "proficient," compared with the federal NAEP benchmark. We do not assume that where the NAEP governing board currently sets cut-points is necessarily appropriate, nor validated against actual knowledge demands pressed by colleges or employers (Linn, 2000).

How cut-points are set, along with a state's emphasis placed on certain curricular topics, may recognize the mastery of basic domains by low-performing students. This may allow a larger group of students (at the low end of the distribution) to cross over the proficiency cut-point, while failing to differentiate students who achieve at higher levels. Overall, the gap between state test results and achievement on the NAEP can be driven by multiple factors, as Koretz (2006) emphasizes. And separating the actual mastery of certain curricular topics by particular cohorts of students from "inflation," under which real learning is illusory, becomes a slippery analytic task.

Researchers at the RAND Corporation analyzed the rise in scores on the earlier Texas Assessment of Academic Skills (TAAS), a study that achieved notoriety after being released two weeks prior to the 2000 presidential election. The substance of the study was eye-opening and followed the pattern seen in the Kentucky case, although in Texas the data were not analyzed at the item level as Koretz and Barron had done in Kentucky.

According to state results, fourth-grade reading scores climbed between 1994 and 1998 by fully 0.31 standard deviation (SD) for white fourth-graders, 0.49 for blacks, and 0.39 for Latinos. These gains were detected on the NAEP but at lower levels of magnitude in terms of corresponding effect sizes: 0.13, 0.14, and 0.14 SD (Klein, Hamilton, McCaffrey, & Stecher, 2000).

The RAND team warned states and schools to avoid "coach(ing) to develop skills that are unique to specific types of questions that are asked on the statewide exam… (and) narrowing

the curriculum to improve scores." They also emphasized that the TAAS results were "biased by various features of the testing program (e.g., if a significant percentage of students top out or bottom out on the test, it may produce results that suggest that the gap among racial and ethnic groups is closing when no such change is occurring)."[1]

The second constraint facing states is that changing a testing program typically leads to a decline in mean scores, as the factors that inflate results are temporarily suspended. That is, teachers don't know the test items to which they must teach, questions likely align with a new set of curricular domains and constructs, and the novel format of a new test may constrain student performance. Linn (2000) has documented this saw-tooth pattern of, first, steady gains under testing system A, and then a sharp decline when the state shifts to test B. The pattern often repeats as a third test replaces the second (also, Koretz, Linn, Dunbar, & Shepard, 1991; Koretz, in press, for case studies).

Third, state authorities often set proficiency cut-points at levels that are far below the thresholds established by the NAEP governing board. When *Education Week* (2005) analysts compared the percentage of fourth-graders deemed proficient in reading under state versus NAEP standards for 2003, not one state education department in the nation set their bar above the hurdle established by NAEP designers.[2]

Take the case of Alabama, where the state determined that 77 percent of all fourth-graders could read "proficiently" or above in 2003, compared with just 22 percent as assessed by the NAEP. This wide gulf between state and NAEP definitions of proficiency equaled 30 percentage points for New York state, 54 points in Tennessee, and 18 points in California. Our historical analysis reported below reveals that Massachusetts offers a rare case where the percentages of children deemed proficient in reading and math by the Commonwealth have been within 10 percentage points of the shares estimated from NAEP results, going back to the mid-1990s.

The final constraint on how states shape their testing systems does act to improve the reliability of exam results. It also may narrow what children learn in school. Remember the original advocates of systemic reform, like Cohen, foresaw an assessment process that would encourage more complex thinking skills. But states that have ventured into this terrain—previously unexplored on a statewide scale—often become lost in a morass of political hazards and technical difficulties.[3]

The original KIRIS tests, for example, included open-ended items and short essays. But these came to be collected only for a subset of students within a matrix-sampling approach and were never used within the accountability system (Koretz & Barron, 1998). Writing portfolios also were included in the original KIRIS exams, but then dropped after a few years (Massell, Kirst, & Hoppe, 1997).

In California, state schools chief Bill Honig devised the Classroom Learning Assessment System (CLAS) in the early 1990s, focusing on higher-order thinking and assessment of writing, not only reading, proficiencies. But questions were raised over the reliability of open-ended items, along with innovative questions that delved into students' backgrounds and attitudes. Scores

dropped in many middle-class communities. Governor Pete Wilson responded by abolishing the CLAS, then placing a moratorium on all state testing (Colvin, 1995; McDonnell, 2005).

In North Carolina, McDonnell and Choisser (1997) document how the state's original eight learning objectives, including attention to the child's "self-sufficiency" and "responsible group membership," were reduced to learning aims which could be more reliably, and less controversially, gauged over time.

This bureaucratic drift toward narrower specification of a simpler range of learning goals served to strengthen the measurement properties of testing systems—while shrinking the range of knowledge and social behavior that would come to be pressed by the state. The technical requirements of sound and efficient testing, implemented on a mass scale and in politically safe ways, came to shape what children are to learn (Stetcher, 2002).

Is NAEP the Answer?

Given these constraints on state-run testing programs, should parents and policy makers rely instead on NAEP results to accurately track student progress?

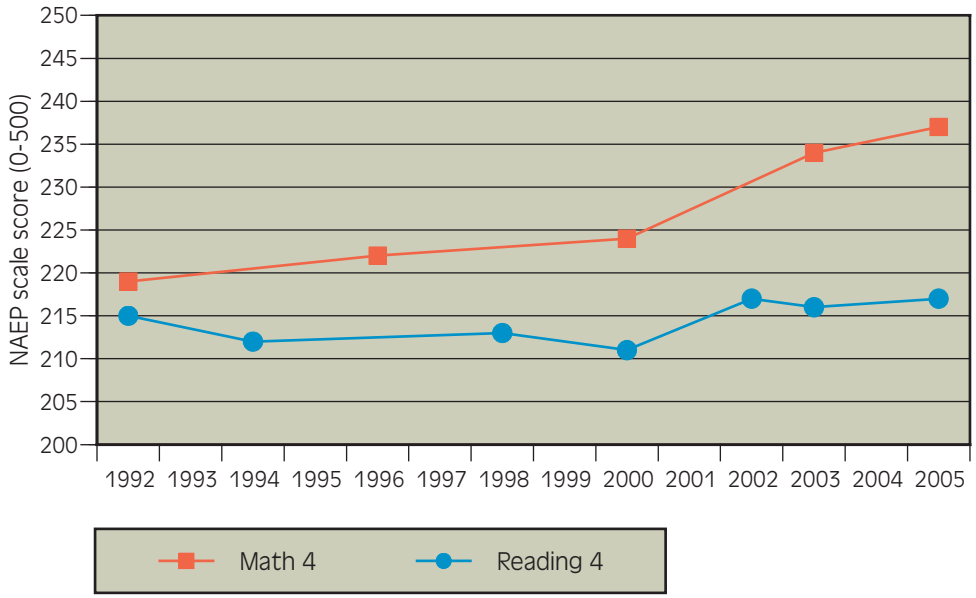
No, some reformers respond, since the NAEP isn't a cornerstone of the original theory of systemic reform. Students and teachers must know what is expected of them—that is, which domains of knowledge are to be addressed in the classroom—if they are to feel efficacious in raising achievement. It's state or district curricular panels, so the argument goes, that are best positioned to articulate learning standards, then align standardized tests to gauge progress.

In addition, the NAEP is built from random samples of students in each state. So, not all students take the NAEP, nor is it intended to inform parents or students about how individual youngsters are performing. The NAEP mainly assesses proficiencies in reading and mathematics. Exams are administered periodically in civic, geography, history, science, and writing, but with neither the frequency nor the coverage that would suffice for state policy makers.

Then there's the argument that the proficiency hurdle set by the NAEP governing board is simply too high. Cut-points tend to be set by curriculum experts at demanding levels, certainly relative to the thresholds established by state officials, applying the same label of students being "proficient" for a particular grade level (Linn, 2000; Koretz, 2006). Nor has the NAEP board shown much interest in tracking achievement in ways that take into account the changing demographic character of America's students (Zilbert, 2006).

To fill-out this backdrop, let's look at how fourth-grade NAEP scores have moved since 1992, displayed in Figure 1. These scale scores have climbed for students' math performance. The share of fourth-graders proficient in math rose from 18 percent to 32 percent between 1992 and 2003, and then another four percentage points (although statistically insignificant) between 2003 and 2005 (Perie, Grigg, & Dion, 2005). In contrast, reading scores have remained remarkably flat. The mean scale score barely inched upward, with the percentage of fourth graders proficient in reading increasing from 30 to 31 percent between 1992 and 2005 (Perie, Grigg, & Donahue, 2005)

FIGURE 1 National Trends in Student Achievement on the NAEP



Note: Trends between 1992 and 2005 reflect statistically significant increases for math but not for reading in grades 4 and 8. Data from 1996 to 2005 reflect the use of accommodations for students with disabilities and English-language learners. Accommodations were not permitted in 1992 and 1994.
Source: Editorial Projects in Education Research Center (Olson, 2006).

States undoubtedly will continue to stand by their testing programs. So, it's important that we learn more about how these gauges are performing over time, especially when placed against NAEP benchmarks. Next, we examine these questions, turning to historical data on 12 diverse states.

Parallel Play—Gaps in State and Federal Test Score Trends

Selecting states. Uncovering annual state test scores, going back to the early 1990s, requires lots of digging. Documents and electronic bulletins from state education departments focus on the current picture, or look back only two or three years. Many states change their testing regimes every four or five years.

So, to look longitudinally at trends, comparing state and NAEP test results, requires selecting a manageable sample of states. After choosing 12 diverse states, we spent 14 months working with state education officials, fellow researchers, education associations, and searching newspaper archives to construct time-series that reflected comparable test results. This took us back a decade prior to passage of NCLB (January 2002) and three years hence, aiming to match state-level NAEP scores in reading and math over the period, 1992-2005.

How to select states for such in-depth tracking of student performance has proven controversial in recent years. When the *Education Trust* aimed to track elementary school test results

over recent years, 2003-2005, they could locate comparable statewide scores for 32 states. The U.S. Department of Education's analysis (2004) of elementary and middle-school results drew from 25 states, looking back just one or two years.

We do not attempt to generalize our findings to the nation from the results found in the 12 sampled states. We do compare the historical patterns that we discovered, 1992-2005, to those detailed by other investigators who have drawn from recent state test results. For example, *Education Week* analysts (Skinner, 2005) detailed the gaps between the percentages of students deemed proficient or above, based on state versus NAEP tests in one year, 2003.

In sampling states we took into account four criteria. First, we endeavored to select a diverse range of states that might reveal differing patterns of fourth-graders' performance over time. To achieve this variation we first took into account the population size and urban or rural character of the state. Second, we aimed to ensure a geographically dispersed set of states, making selections from the East, upper Midwest, the South, lower parts of the Midwest, and the West.

Third, we sampled states to ensure variation in the intensity of their state accountability programs during the 1990s. At least three research teams have reviewed the presence and strength of accountability structures (Carnoy & Loeb, 2002; *Education Week,* 2005; Goertz & Duffy, 2001). We selected states that ranged from having strong to weak accountability systems prior to the passage of NCLB. An important possibility is that NCLB may yield significant gains for students in states that entered 2002 with weak accountability programs.

Fourth, we selected states in which state education departments or other sources could help locate historical data on student performance. These time-series typically began when states shifted to reporting results in terms of percent proficient or above in core subjects, typically in the mid-1990s. We also drew from statewide test scores published by academic researchers or in major newspapers. Most data points came from officials inside state education departments, including documents and statistical reports provided by education staffers (Appendix 2 lists data sources).

Table 1 displays 12 states that met our selection criteria. They vary geographically and in terms of enrollment size. The intensity of their accountability systems, as judged by independent analysts, differed widely during the 1990s and into 2004. The Carnoy and Loeb (2002) index took into account how many grade levels of students were tested, the severity of sanctions faced by "underperforming schools," whether an exit exam was in place, and the extent to which achievement data were publicly available. These policies yielded an index score between zero and five (column 2).

The *Education Week* (2005) analysts took into account a similar, although more numerous, set of indicators to arrive at a letter grade for each state's accountability program, as gauged in 2003 and 2004.[4] The two accountability indices proved to be highly related. Carnoy and Loeb, for instance, awarded a zero to Iowa and Nebraska for their weak or non-existent accountability programs in the 1990s. The *Education Week* team gave these states an 'F' and 'D',

TABLE 1. Enrollment size and intensity of state accountability and performance information systems

	Enrollment size, national rank[1]	Education Week's overall grade for standards and accountability (2004)[2]	Carnoy & Loeb accountability intensity index score (2000)[3]	Data items collected, via the state accountability system[4]
Arkansas	34	C	1	7
California	1	B+	4	16
Illinois	5	B	2.5	12
Iowa	32	F	0	0
Kentucky	26	A	4	5
Massachusetts	16	A	2	13
Nebraska	37	D	0	0
New Jersey	10	B	5	11
North Carolina	11	B	5	6
Oklahoma	27	B+	1	8
Texas	2	C+	5	8
Washington	14	B	1	12

[1] National Center for Educational Statistics (2005).

[2] Education Week (2004).

[3] Carnoy & Loeb (2002).

[4] Hurst, Tan, Meek, Sellers, & McArthur (2003). Data items inventoried by the U.S. Department of Education included student performance indicators, indicators of school quality, attributes of teachers and principals, and aspects of curricular programs, parent involvement, along with school-level expenditures (a total of 35 possible items).

respectively. Under Carnoy and Loeb, Kentucky, New Jersey, and North Carolina earned index scores of 4, 5, and 5, respectively, while *Education Week* awarded these states grades of A, B, and B.

We can't formally associate, with a sample of just 12 states, the intensity of accountability systems with test score trends. But we can focus, in part, on three states which displayed weak accountability regimes—Arkansas, Iowa, and Nebraska—to see if their achievement trends differ from the other states.

Differing trend lines. We first examined the extent to which state scores have moved in parallel fashion with NAEP scores over the 1992-2005 period for comparable time-series. Some states did not establish cut-points and convert raw scores into estimates of percent proficient until the late 1990s. For earlier years we occasionally plot percentile scores.

Table 2 presents summary results, and below we detail trends for illustrative states (Appendix 1 includes displays for all 12 states). Column 1 reports the mean gap between the percentage children deemed "proficient" or "advanced" (the latter designation varies), according to state testing systems versus the share determined proficient or above from NAEP results. The summary statistics reported are the means derived from annual percentages reported by states, or the actual and interpolated values stemming from NAEP data.

We see—for these fourth-grade reading scores—that state cut-offs determining which students are proficient or above are set lower, often much lower, than thresholds set by the NAEP governing board. This pattern was in place long before enactment of NCLB.

The first two columns report the mean difference between the percentage of children deemed proficient (or advanced) under state versus NAEP testing systems. This was calculated by taking into account state test results for each year (between 1992 and 2005) in which the percentage of children reaching "proficiency" were reported, then subtracting the share deemed proficient by the NAEP (annual scores for the NAEP were interpolated between actual data points). The footnotes specify the exact years for which comparable data are available by state.

In Kentucky, for example, the average share of fourth-graders reported as proficient or above in reading is 31 percentage points higher under the state testing system than the share determined under the NAEP over the 1992-2005 period. State tests in Massachusetts have yielded the closet share of proficient or above, compared with NAEP results: in reading the average annual gap between the two gauges equals 10 percentage points, and just a one percentage point difference in math.

But the historical gap between state and federal results is dramatic in other states. In Oklahoma, for instance, the disparity between percentage proficient or above, gauged by state versus NAEP results, averaged 48 percentage points in reading and 60 points in math since 1996. The state-NAEP gap in reading has averaged 55 points in Texas and 51 points in math since 1994. Such gaps were earlier reported by Skinner (2005) for state and NAEP results in 2003. Our new results show that this gulf between state and NAEP results has long existed for both reading and math.

TABLE 2. Comparing trends between state and NAEP test results for fourth-graders, 1992-2005 (Percentage of students at or above proficient, state and NAEP definitions)

States sorted by accountability intensity grade[1]	Average annual mean gap in percentage of 4th-graders proficient and above (state minus NAEP)		Average annual percentage point gain for recent continuous pre-NCLB series				Average annual percentage point gain, post-NCLB (2002-2005)			
	Reading	Math	Reading		Math		Reading		Math	
			State	NAEP	State	NAEP	State	NAEP	State	NAEP
Kentucky[2] (A)	31	16	1.3	0.7	2.7	0.7	2.7	0.3	3.0	1.9
Massachusetts[3] (A)	10	1	3.0[3]	1.1	1.3	1.5	-1.3 [3]	-1.0	0.3	3.8
California[4] (B+)	19	24	3.0[4]	0.2	4.0	0.9	3.7 [4]	0.0	4.3	2.3
Oklahoma[5] (B+)	48	60	-0.7	-0.3	-1.1	0.7	2.3	-0.3	4.3	2.8
Illinois[6] (B)	35	47	0.7	--[6]	2.0	4.0	1.2	-1.0	1.8	1.3
New Jersey[7] (B)	42	36	7.9[7]	0.3	2.8	1.2	-0.1	-0.4	0.2	2.7
North Carolina[8] (B)	43	54	1.6	0.7	2.8	2.3	2.1	-1.0	1.3	1.5
Washington[9] (B)	32	13	3.5	1.0	6.1	2.1	4.6	0.3	3.0	2.7
Texas[10] (C+)	55	51	2.4	0.4	4.6	1.5	1.5[10]	0.3	5.5[10]	3.2
Arkansas[11] (C)	26	27	5.0	0.8	4.5	2.1	3.3	1.3	6.3	4.0
Nebraska[12] (D)	44	48	3.5	0.3	--[12]	0.9	1.5	0.0	3.5	1.8
Iowa[13] (F)	38	45	-0.1	-0.1	0.1	0.7	1.1	-0.7	1.9	1.4
Unweighted means	35	36	2.6	0.4	2.7	1.5	1.9	-0.2	2.9	2.4

TABLE 2 Notes

[1] As scored by Education Week (2004).

[2] Kentucky Core Content Test (KCCT), 2000-05 (grade 4 reading and grade 5 math). The base year for the post-2002 NAEP trend for math is interpolated.

[3] Massachusetts Comprehensive Assessment System (MCAS), 2001-2005 (grade 4 English language arts), 1998-2005 (grade 4 math). The base year for the post-2002 NAEP trend for math is interpolated.

[4] Stanford-9, 1998-2001, percent above the national norm; California Standards Test (CST) 2001-2005 (grade 4 English language arts and math). The base year for the post-2002 NAEP trend for math is interpolated.

[5] Oklahoma Core Curriculum Test (OCCT), 1996-2005 (grade 5 reading), 1995-2005 (grade 5 math) percent satisfactory or above; 1999-2002 "traditional" students only, 2003-2005, and "regular education" students. The base year for the post-2002 NAEP trend for math is interpolated.

[6] Illinois Goal Assessment Program (IGAP), 1992-1998 (grade 3) percent meeting or exceeding state goals; Illinois Standards Achievement Test (ISAT), 1999-2005 (grade 3), percent meeting or exceeding Illinois Learning Standards. NAEP testing in reading did not begin until 2003. The base year for the post-2002 NAEP trend for math is interpolated.

[7] Elementary School Proficiency Assessment (ESPA), 1999-2002; New Jersey Assessment of Knowledge and Skills (NJ ASK), 2003-04. The state does not distinguish between ESPA and NJ ASK when analyzing its data, so we have followed suit. Score values are "general" scores (combined minus ESL and special education) as opposed to "total" scores (all students). Between 2000 and 2001 a 24.2 point gain on the ESPA reading exam was reported with no published changes in the cut-off values. The base years for the post-2002 NAEP trend for math and reading are interpolated.

[8] End-of-grade testing program, 1993-2005, percent at achievement levels III and IV. The base year for the post-2002 NAEP trend for math is interpolated.

[9] Washington Assessment of Student Learning (WASL), 1997-2005. The base year for the post-2002 NAEP trend for math is interpolated.

[10] Texas Assessment of Academic Skills (TAAS), 1994-2002, percent meeting minimum expectations; Texas Assessment of Knowledge and Skills (TAKS), 2003-2005, percent at panel's recommendation for minimal proficiency. TAKS began in 2003, thus the base year for the post-2002 state gain is 2003. The base year for the post-2002 NAEP trend for math is interpolated.

[11] Arkansas Benchmark Exams, 1998-2005. Score values are "combined" scores (all students) as opposed to "general" scores (combined minus ESL and special education). In 2005 new cut-off values were set, but scores relative to the previous cut-points were obtained. The base year for the post-2002 NAEP trend for math is interpolated.

[12] School-based Teacher-led Assessment Reporting System (STARS), 2001 and 2003 (reading), 2002 and 2004 (math), percent meeting or exceeding standards. State testing in math did not begin until 2002, thus no data are available for pre-NCLB state math gain. The base year for the post-2002 NAEP trend for math is interpolated.

[13] Iowa Test of Basic Skills (ITBS), 1994-2005. In 2000 new norms were set for the exam but only used for the 2001-2003 and 2002-2004 biennium averages (the post-2002 gains are derived from these two biennium values). The base year for the post-2002 NAEP trend for math is interpolated.

The disparity between state and NAEP results is also apparent in the year-to-year levels of change reported by the states. This suggests that gaps between the dual testing regimes are established by differences in cut-points, and then trend lines diverge as state results suffer from inflation over time.

Columns 3 through 6 (in Table 2) report the annual change in the percentage of children deemed proficient or above under state test results for the most recent continuous time-series, along with estimated annual changes in the NAEP. Kentucky's reading scores, for instance,

showed a 1.3 yearly point increase prior to NCLB enactment in the percentage of fourth-graders deemed at least proficient in reading, based on state test scores. But for the NAEP, the annual increase averaged 0.7 percentage points.

Two states have shown what appear to be dramatic gains, including New Jersey, where the average annual gain equals 7.9 percentage points in reading. But this includes a 24.2 point increase in the percentage of children deemed proficient or above between 2000 and 2001. Similarly, Arkansas reported a jump of 19 percentage points in the share of fourth-graders proficiency or higher in reading between 2001 and 2002, contributing to this sizeable gain when annualized.

Annual progress in math performance is, at times, more consistently gauged by state exams, relative to NAEP results (columns 5 and 6). North Carolina's mean annual gain of 2.8 percentage points tracks well against a mean rise of 2.3 percentage points for the NAEP. But the inflation of test results over time is vivid for some states. Washington's state tests yield mean annual gain of 6 percentage points in the share of fourth-graders determined to be proficient or above, compared with a 2 point gain each year when gauged by the NAEP.

The mean changes across all 12 states (bottom row) show that students in our sample performed slightly better on the NAEP in reading and about the same in math, compared with national trends over the 1992-2004 period.

Gains After the Enactment of NCLB?

The final four columns in Table 2 report average yearly changes over the three school years after passage of NCLB. The base period is pegged at spring 2002, the first testing season after President Bush signed the Act.

In the post-NCLB period the Kentucky reading test yielded a 2.7 percentage point yearly gain in terms of the percent proficient or higher in reading, on average. But the NAEP results showed just a 0.3 percentage point gain per annum. California reported a 3.7 point average yearly gain, but the NAEP results showed no discernible progress.

Texas education officials continued to report a 1.5 point increase per year in the share of students proficient or above in reading post-NCLB, even as the NAEP results showed almost no discernible change. In contrast, NAEP math scores continued to climb, registering a 3.2 percentage point gain each year, post NCLB. Still, the Texas testing system yielded a picture of even more robust growth of fully 5.5 percentage points per year, on average, again in terms of the share of fourth-graders proficient or above. In short, the pattern of inflated state results has generally persisted during the three years since enactment of NCLB.

State test results for math have yielded a more consistent gauge of progress since 2002, compared with NAEP results. Still, state test results show inflated levels of progress. Nebraska's state testing results show a 3.5 percentage point yearly gain, on average, in the share of students deemed proficient or higher, compared with a 1.8 point increase on the NAEP math exam. California reports a 4.3 percentage point gain each year, post-NCLB passage, while the NAEP shows this improvement is 2.3 percentage points a year.

Another important pattern is revealed in Table 2. Mean annual gains in reading, reported from state test scores, continued to climb after 2002. But NAEP proficiency levels hit a flat plateau or declined in some cases. State math scores after passage of NCLB continued to climb, as did NAEP scores. Yet mean annual gains in math tended to range higher for state test results, compared with the slower pace of progress revealed by NAEP scores.

Some analysts argue that NCLB created a distorted incentive for states to set low cut-points when determining percent proficient, given the Act's mandate of universal proficiency by 2013. Schools in states like California are disproportionately sanctioned when boards of education set challenging learning standards. This means that each annual stair step toward universal proficiency in California is a lot steeper, compared to schools situated in states that set lower standards and thus lower proficiency cut-points.

Differing patterns among states. Figure 2 displays common patterns observed among the 12 focal states. The top panel looks at fourth-grade reading scores for North Carolina. The top trend line shows the percentage of children deemed proficient or above, according to state test scores. The lower curve reports the share proficient as reported by the NAEP. Wide gaps between the two gauges of student proficiency are clearly observed.

In addition, we see that the percentage defined as proficient or above climbs steadily from 1992 forward, under state definitions. But only slight gains are revealed by NAEP results. This illustrates the tandem effect of setting a lower cut-point initially, followed by inflation of state results, presumably mixed with greater mastery of curricular material by some students.

The middle panel shows the same two trend lines for Texas. Note the break in the series for the share of fourth-graders proficient or above in reading, due to a change in the state testing regime. Once Texas put in place a new test, the Texas Assessment of Knowledge and Skills (TAKS), the share deemed proficient dropped, in part because state education officials raised the cut-point for the minimal level of achievement required to reach proficiency. This illustrates the saw-tooth pattern commonly seen in states that change testing programs.

The third panel, displaying results for Iowa, shows another pattern seen in some focal states: flat student performance on both state and NAEP exams over time.

Figure 3 shows state and NAEP results for math performance. The top panel, again focusing on North Carolina, shows a more encouraging pattern: the percentage of students deemed at least proficient rose through 2003 for both state and NAEP results, before leveling-off. Texas also displays solid progress in math, based on NAEP results.

These gains in Texas were large over the period, especially for African-American and Latino students. The latter group, in 2005, performed at the same level on the NAEP as that achieved by white students 15 years earlier (Treisman, 2005). State test results accelerated at a more rapid pace under the earlier test, then slowed and followed an erratic pattern after the state switched to the TAKS exam.[5]

FIGURE 2. Percentage of fourth-graders proficient or above in reading, according to state and NAEP testing

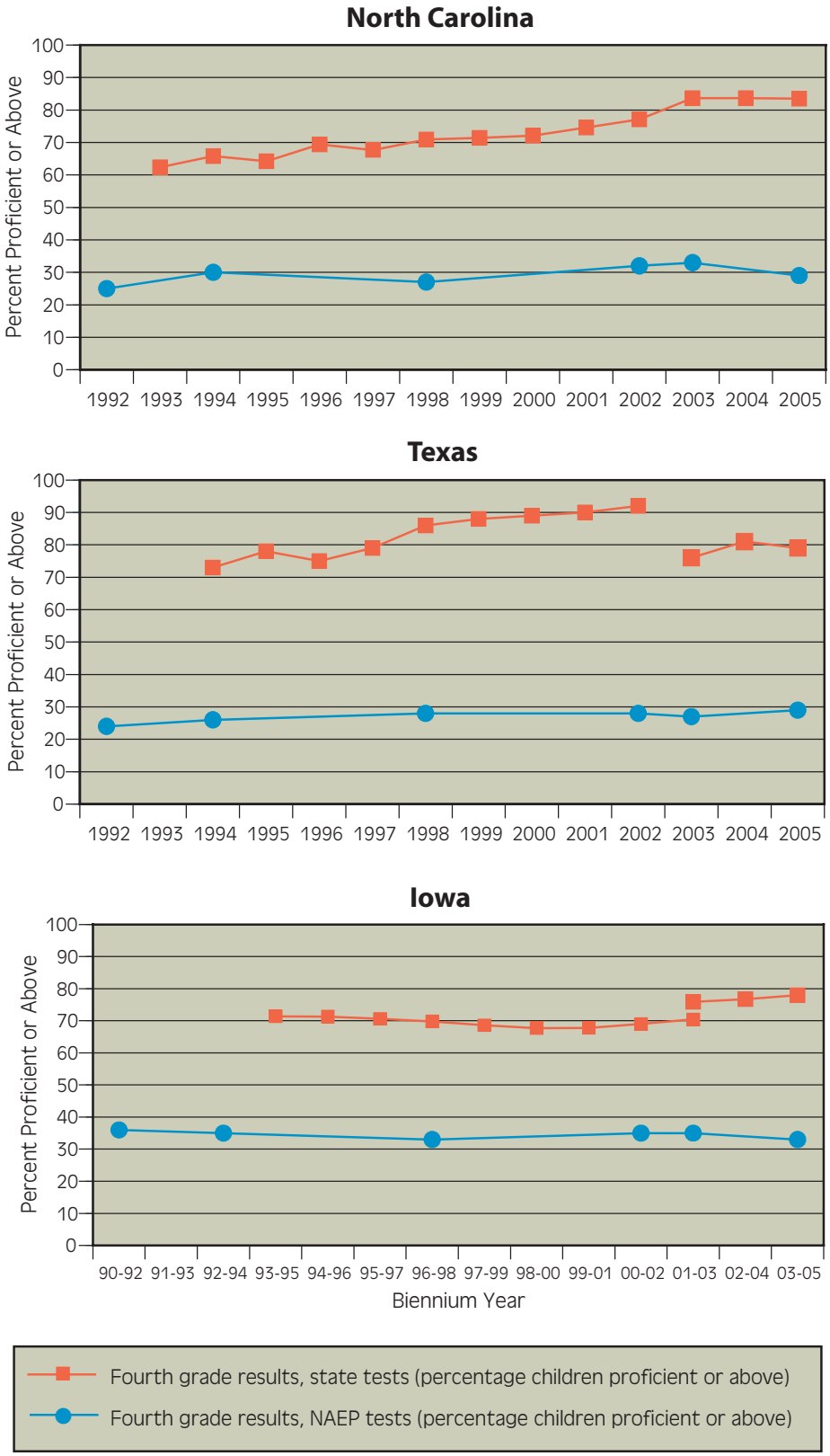
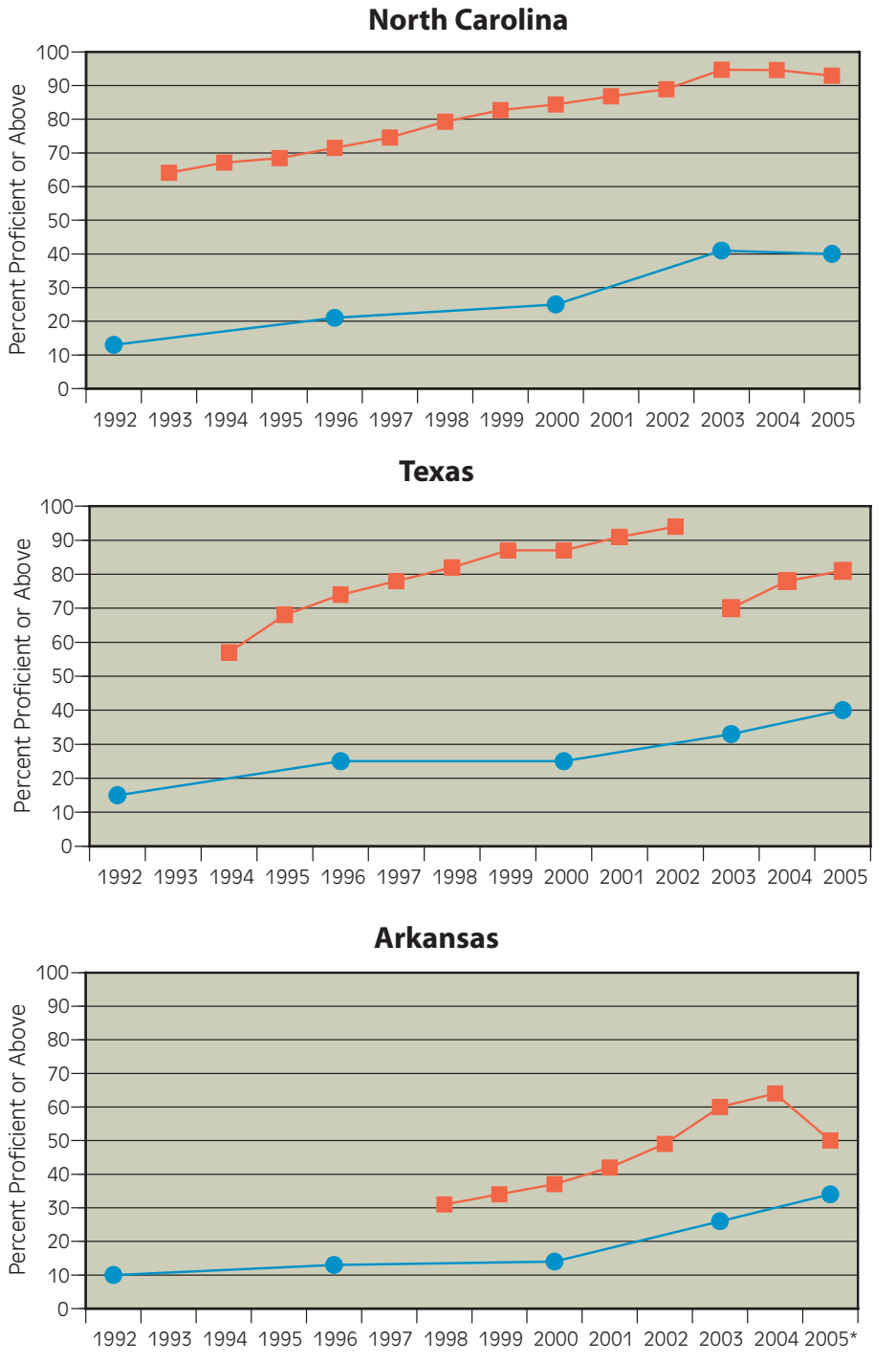


FIGURE 3. Percentage of fourth-graders proficient or above in math, according to state and NAEP testing



*Decline, at least in part, due to the state's decision to raise the standard above which children are "proficient".

Fourth grade results, state tests (percentage children proficient or above)

Fourth grade results, NAEP tests (percentage children proficient or above)

The difference between the rate of progress in fourth-grade math, again contrasting state versus NAEP results, is even more dramatic for Arkansas (bottom panel). The percentage of children deemed at least proficient under the Arkansas exam climbed from just over 30 percent to almost 70 percent between 1998 and 2004. The NAEP results also showed improvement in math proficiency. But the share of fourth-graders deemed proficient or above under the NAEP grew from 18 to 31 percent during this period.

Quantifying the association between state and NAEP results. To statistically characterize the covariation between state and NAEP test score results, we aggregated annual scores for all 12 states. This provided a maximum of 87 data points for both reading and math scores over an average of seven consecutive years of comparable time-series data within a state.

We found that the mean share of fourth-graders proficient or above in reading equaled 68 percent, according to state tests, but just 31 percent based on NAEP results. These are unweighted arithmetic means. The same pair of percentages for math equaled 65 percent of fourth-graders proficient or above, stemming from state tests, but just 30 percent proficient or above, according to NAEP results.

We then calculated a simple correlation coefficient (r) between state and NAEP *reading* scores, which equaled -0.27, indicating that the two sets of time series actually diverged modestly from one another over time (pooled years, n, equaled 87). Put another way, state test results can account for just six percent of the total variance in NAEP scores in a single-factor regression model. State *math* scores are not statistically associated with NAEP math scores over time: the bivariate correlation (n=80) equaled just 0.02.

This high degree of independence between the two sets of gauges appears to be related to the inflated rate of improvement in state results, starkly contrasting the flatter trend lines observed for NAEP results. Remember also that state scores bump around more, especially in jagged saw-tooth fashion whenever state officials change their exam systems.

If the problem was simply rooted in where states set initial cut-points, we would still see a tighter correspondence between the movement of state and NAEP scores over time. But the fact that state scores are moving independently suggests either that inflation in state scores is occurring, or that the kind of learning tapped by state tests is largely missed by the NAEP assessment. The latter argument seems unlikely.

These correlational results are similar to the weak correlation observed among all 50 states between the share of fourth-graders deemed proficient or above in reading, according to state testing versus NAEP results, compiled for 2003 (*Education Week*, 2005:84; Dillon, 2005b).[6]

Caution is urged by Koretz (2006) and Sigman and Zilbert (2006) when comparing year-to-year changes in student performance between state and NAEP test results, especially if the data series begins at a low or high point in the distribution of raw test scores. When a proficiency cut-point is set near either tail of the distribution, the comparative proportions of students who must cross over the cut-point can vary considerably. This is why we report on the full

13-year time series for most states, although it suggests that care is required before making strong inferences regarding the post-NCLB period, since only three years of observations are available to date.

Finally, we might expect to see stronger progress in fourth-graders' ability to reach the NAEP's "basic" level of achievement, rather than proficient, given that some states focus their accountability regime on low-performing students. This is true in some states, such as California, where the share of fourth-grade students scoring at basic in reading climbed from 48 percent to 60 percent between 1992 and 2005, compared to a rise from 19 percent to 21 percent scoring at proficient over the same period (Perie, Grigg, & Donahue, 2005).

Yet nationwide the trend lines for basic and proficient levels are close to identical for fourth-grade reading (flat) and math (sloping upward). Even if Washington decides in the context of NCLB reauthorization that current NAEP cut-points for proficiency are too challenging—pegged too far above the states' cut-points—the labeling of "basic" and "proficient" could be made more consistent between the tandem assessment programs (Linn, 2000; Koretz, 2006).

Conclusions and Policy Options

Taken together, these findings illuminate the difficulty in answering the bottom-line question: Is NCLB working?

Recent claims based solely on state test results—either pre or post-NCLB—assert gains in some states and other states where fourth-graders have reached a plateau in reading or math performance (Education Trust, 2004; PACE, 2004; Paige, 2004).

Yet we have detailed how state results consistently exaggerate the percentage of fourth-graders deemed proficient or above in reading and math—for any given year and for reported rates of annual progress, compared with NAEP results. For reading, this gulf between the two testing systems has actually grown wider over time. Any analysis conducted over the 1992-2005 period based solely on state results will exaggerate the true amount of progress made by fourth-graders. Again, the counter argument is that the fourth-grade NAEP is insufficiently sensitive to learning gains which are somehow uniquely detected by state tests.

State policy makers will likely stand by their testing regimes, given that some exam systems are closely aligned with curricular standards. But it remains unclear whether it's this tighter alignment that is driving higher estimates of the share of fourth-graders who are proficient, or whether the bar set for state tests is simply being set too low, relative to national standards. Then, teachers and students adapt to state tests in ways that inflate actual levels of substantive learning. Both factors—low cut-points and inflated scores over time—are likely at work in many states.

State education officials at times create tests that are differentially sensitive to improvements at the low level, as revealed earlier in Texas (Klein, Hamilton, McCaffery, & Stecher, 2000). And states periodically change their testing programs, leading to jagged, saw-tooth

trend lines—where the share of students defined as proficient shoots upward or falls precipitously, compared with the steady trend lines that characterize NAEP results.

States should not be discouraged from carefully gauging progress at the low end. State and NAEP testing officials could do more to inform the public on how student demographics are changing and implications for interpreting achievement trends.

The rising proportion of English learners or students of color should not be used as an excuse for insufficient progress. But even the interpretation of NAEP trends is constrained by our inability to understand how achievement is moving, net the prior effects of student and family characteristics. In California, for instance, gains in fourth-grade NAEP scale scores for every subgroup of color were strong, even though state averages remained flat, 1998-2005, given steady change in student composition (Sigman & Zilbert, 2006).

The 12 focal states did show some gains in reading over the 1992-2001 period, the decade prior to NCLB. Mean annual changes in NAEP scores ranged between -0.3 percentage points in Oklahoma (in the share of fourth-graders proficient or higher), to 1.1 points in Massachusetts. These modest inclines were far smaller than average gains reported by state testing programs. The inflated character of state test scores, following the establishment of low cut-points relative to the NAEP, also was apparent for math results.

Returning to the question—is NCLB working—the news after enactment is less encouraging. Many states continued to show progress after 2001, although the rate of growth fell, compared with the pre-NCLB period, even when gauged by their own testing results.

NAEP reading scores, in general, hit a plateau or declined over the three school years following enactment of NCLB. Gains in fourth-grade math were still apparent after 2001, but here too the gains reported by states were significantly higher than the pace of progress indicated by NAEP scores. Slowing rates of achievement growth, post-NCLB, also have been found in at least one tracking study of large numbers of students distributed across several states (Cronin, Kingsbury, McCall, & Bowe, 2005).

Could improvements in NCLB rules and resources encourage states to develop more consistent student assessments, yielding truly comparable achievement levels over time?

The answer is certainly, yes, at a technical level. Washington could first help raise confidence in state testing programs—and wider acceptance of NAEP results—by advancing a consensus as to where the proficiency bar should be set. State tests might be formally benchmarked to the NAEP, along with more transparent reports on the comparative rigor of state tests.

To combat the dumbing-down of cut-points that define proficiency, policy makers might return to the original vision of the challenging standards envisaged by the early architects of systemic reform: designing assessments that encourage analytic and writing skills and higher-order thinking. We should learn from the lessons around technical soundness and political sustainability taught by the earlier episodes from California and Kentucky. Adding rigor and

more complex skills to state tests could be motivating for teachers and students alike, and it would bring them into closer parity with proficiency levels yielded by NAEP assessments.

The federal government might provide resources to state education departments to conduct stronger equating exercises to link old and new tests. It's understandable that states may periodically want to alter who designs and runs their testing programs. But the inability of states to track achievement over time invites federal intervention and heavier reliance on the NAEP. It's difficult to see how fourth-graders could be making real advances in reading and component skills without the NAEP detecting such learning.

If state education officials believe their tests are more closely aligned with curricular standards, or somehow more informative than the NAEP, they should address the technical problems that weaken the credibility of their current systems. Otherwise, the relevance of state testing programs may diminish as parents, voters, and policy makers struggle to discern whether school reforms are, in reality, raising student achievement.

The U.S. Department of Education took a notable step in 2005—inviting states to propose a "growth model" under which schools would be recognized for raising achievement levels no matter where they started on the stair case toward universal proficiency (Anderson, 2005). This intriguing adjustment raises a different set of issues, especially whether states should be awarded greater flexibility when it comes to gauging student progress.

On the other hand, the new policy discussion around rewards for growth opens the door to bring state and NAEP proficiency standards into closer alignment. This conversation will likely occur within the broader debate around the proper role of Washington and the states within a federalist system of school governance.

State officials might engage this discussion with serious ideas for how to improve their assessment regimes. Otherwise, the credibility of their testing results—and state leaders' claims of progress—may suffer. And Washington officials may want to closely inspect these achievement trends, put forward by the states, before declaring victory.

Acknowledgments

We warmly thank Jack Jennings, Dan Koretz, Deb Sigman, Brian Stecher, and Eric Zilbert for their careful comments on earlier drafts of this paper. Funding for PACE's work on improving accountability policies is generously provided by the Hewlett Foundation. Special thanks to Mike Smith for his steady encouragement. The Noyce Foundation supports PACE's work on how to improve state-led accountability programs. Special thanks to Ann S. Bowers and Amy Gerstein. Robust discussions with Kati Haycock continue to sharpen our analysis. Much appreciation is expressed to Mike Kirst, recently retired as codirector of PACE, for his support as we worked through the evidence. Any errors of fact or interpretation are solely the authors' responsibility. Related research appears on pace.berkeley.edu.

Appendix 1 State and NAEP achievement trends for 12 states

FIGURE A1. Arkansas – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

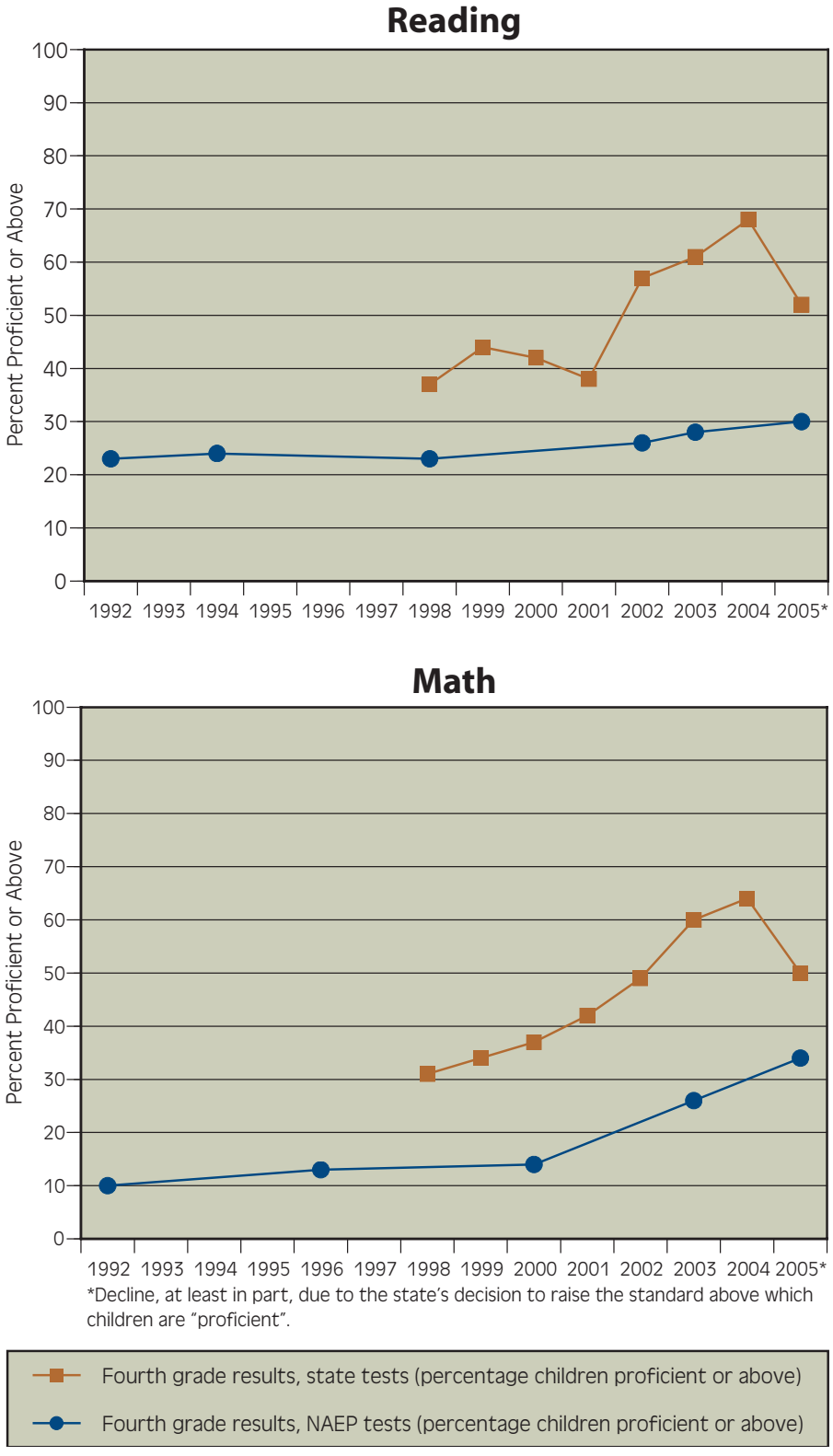
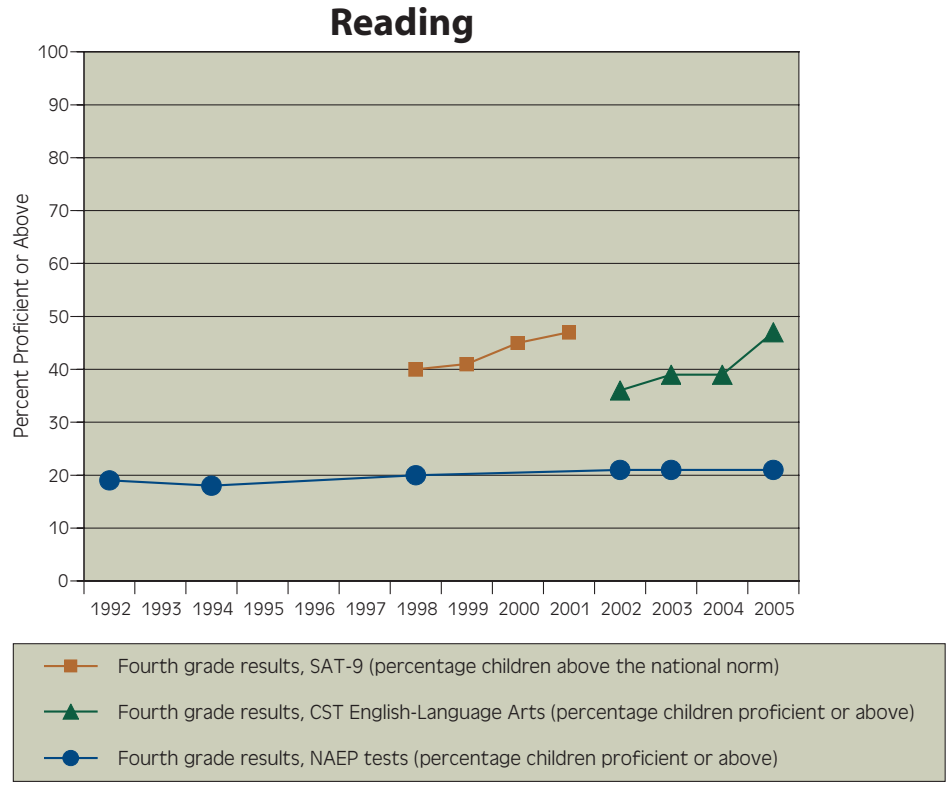


FIGURE A2. California – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing



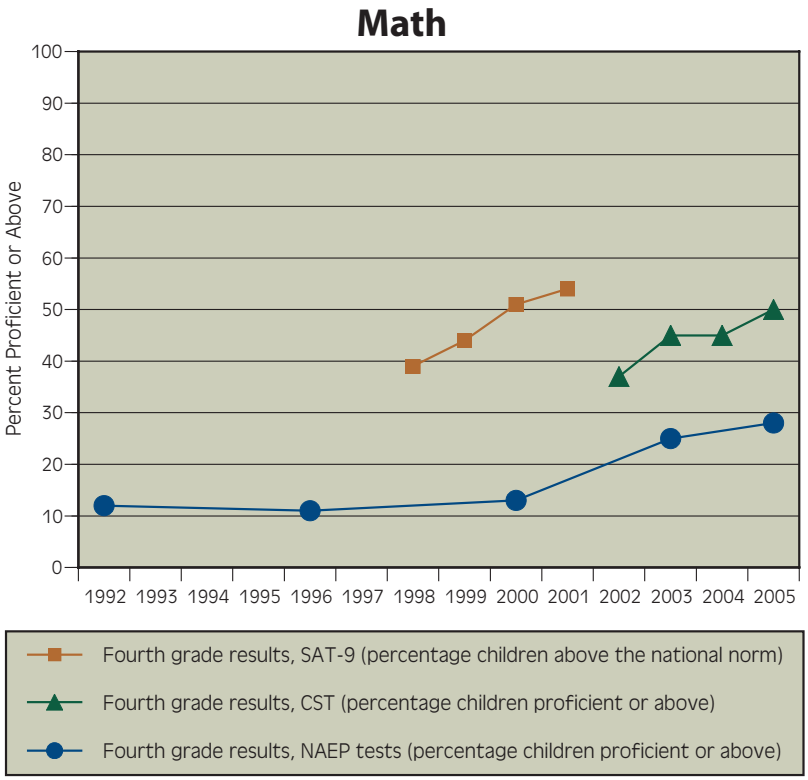


FIGURE A3. Illinois – Percentage of third- or fourth-graders proficient or above in reading and math, according to state and NAEP testing

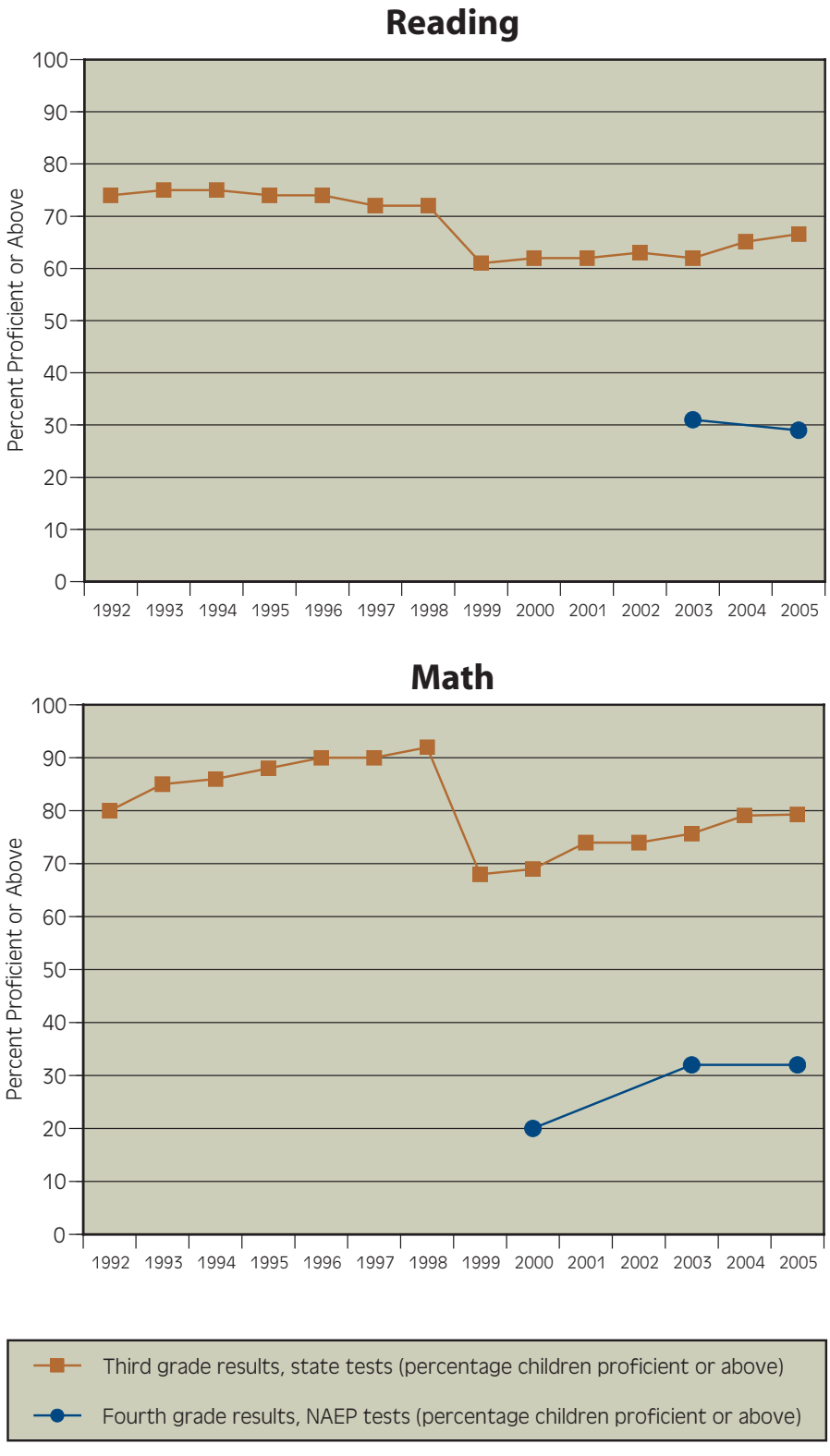
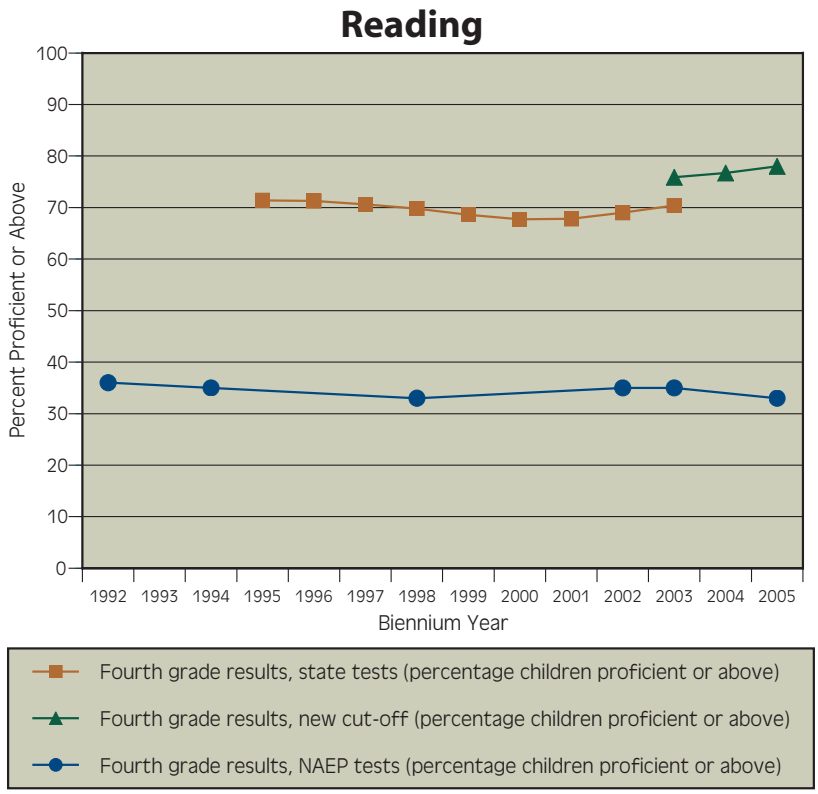


FIGURE A4. Iowa – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing



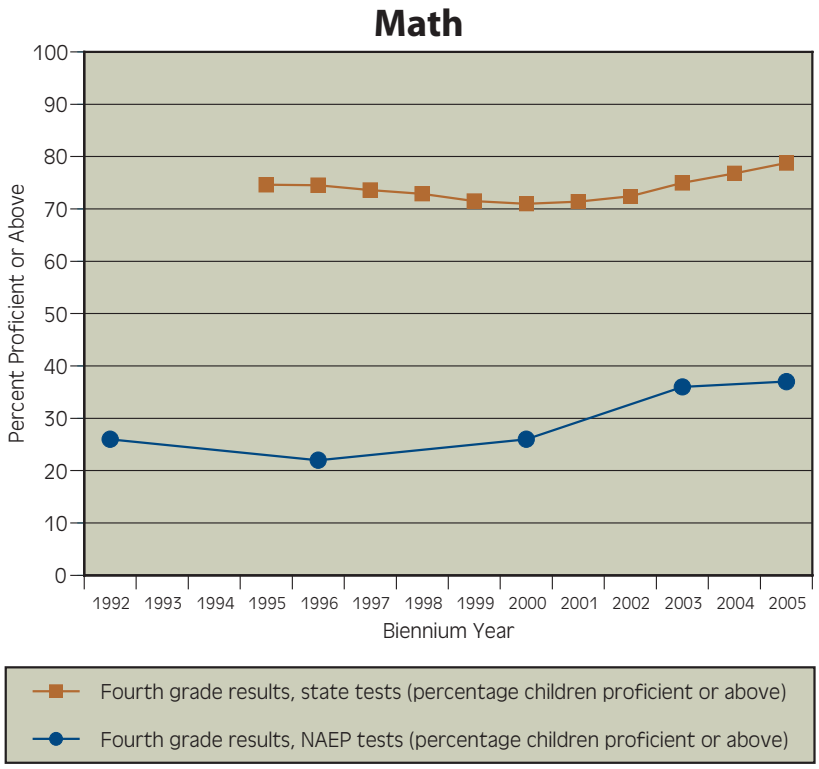
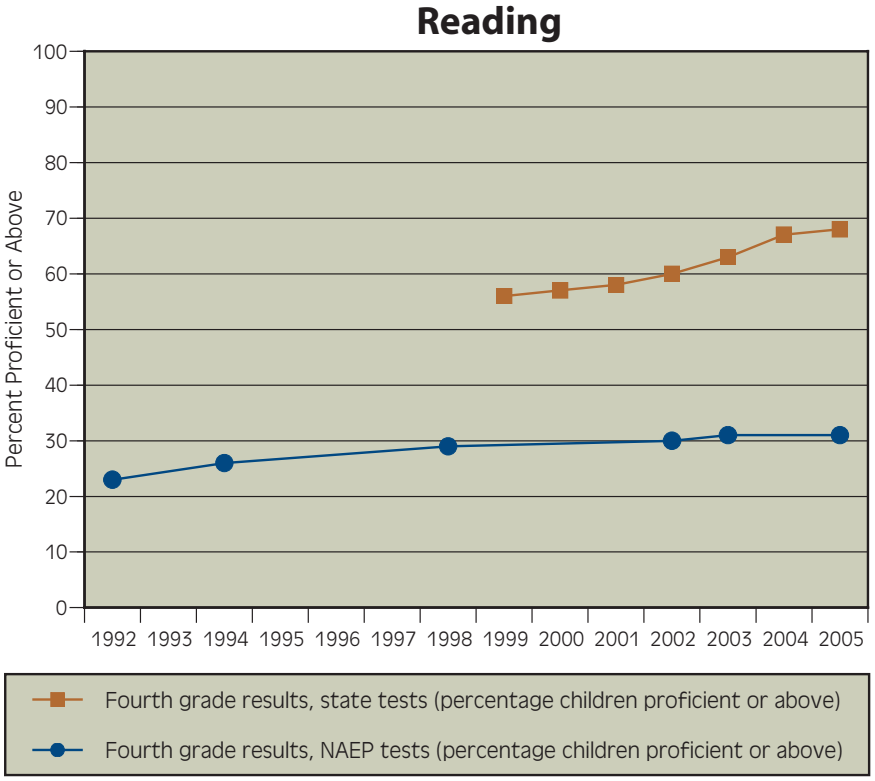


FIGURE A5. Kentucky – Percentage of fourth- or fifth-graders proficient or above in reading and math, according to state and NAEP testing



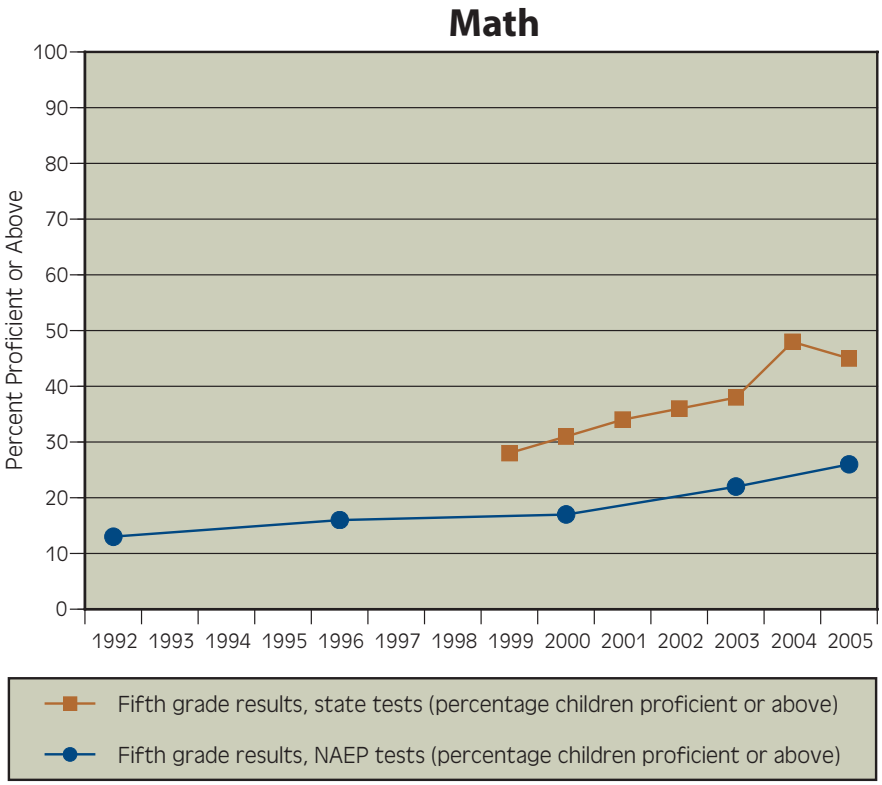


FIGURE A6. Massachusetts – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

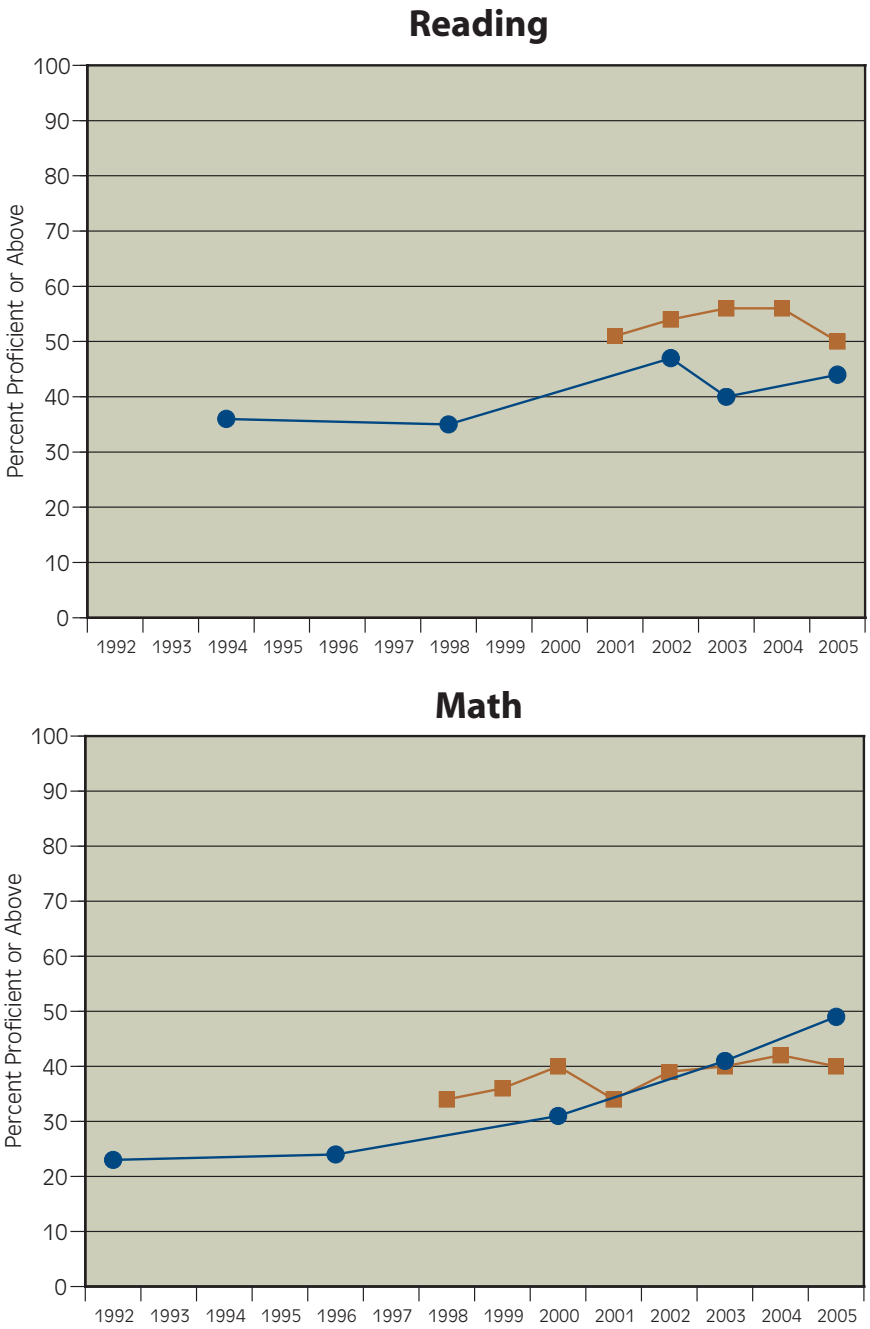


FIGURE A7. Nebraska – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

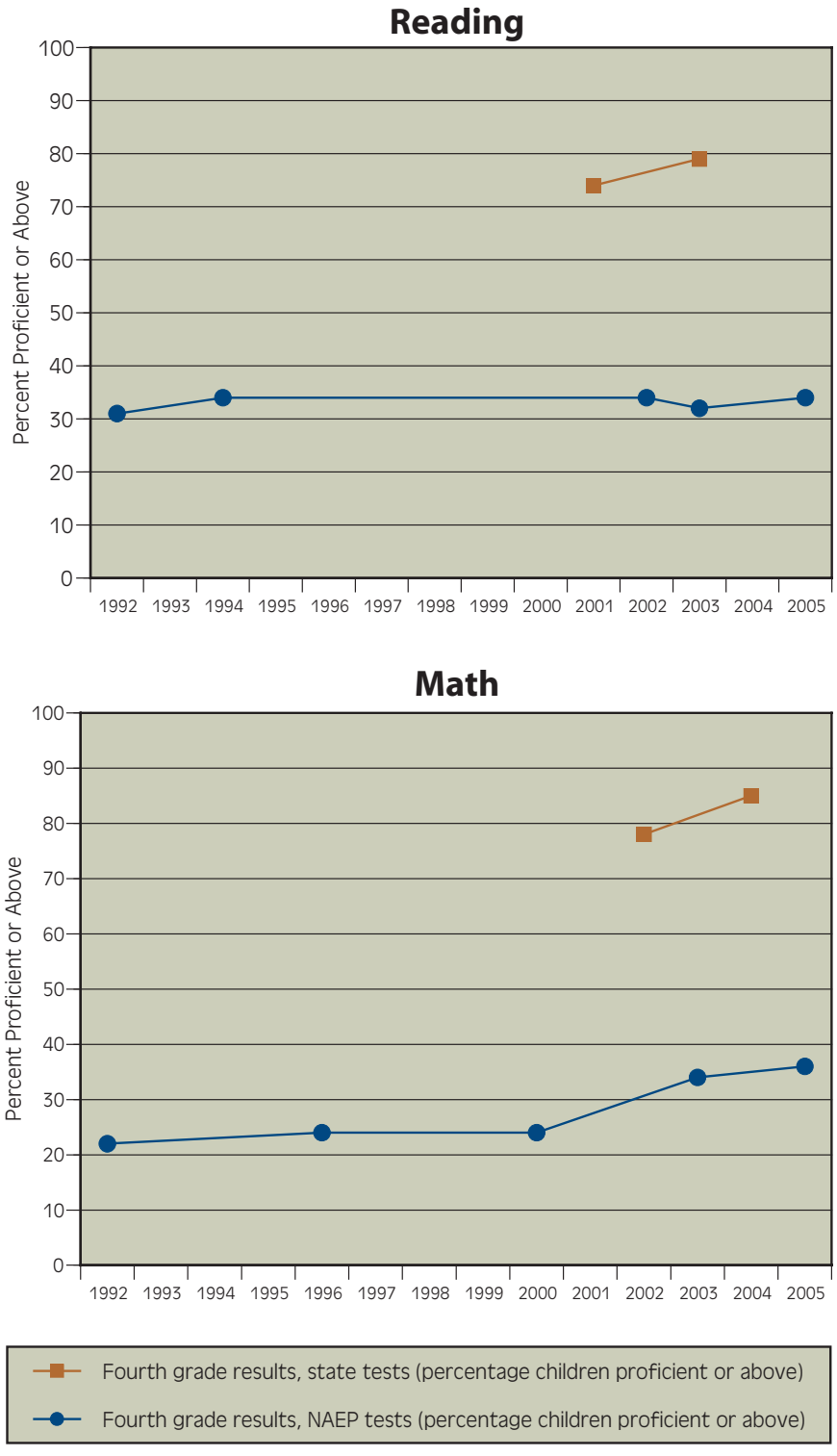


FIGURE A8. New Jersey – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

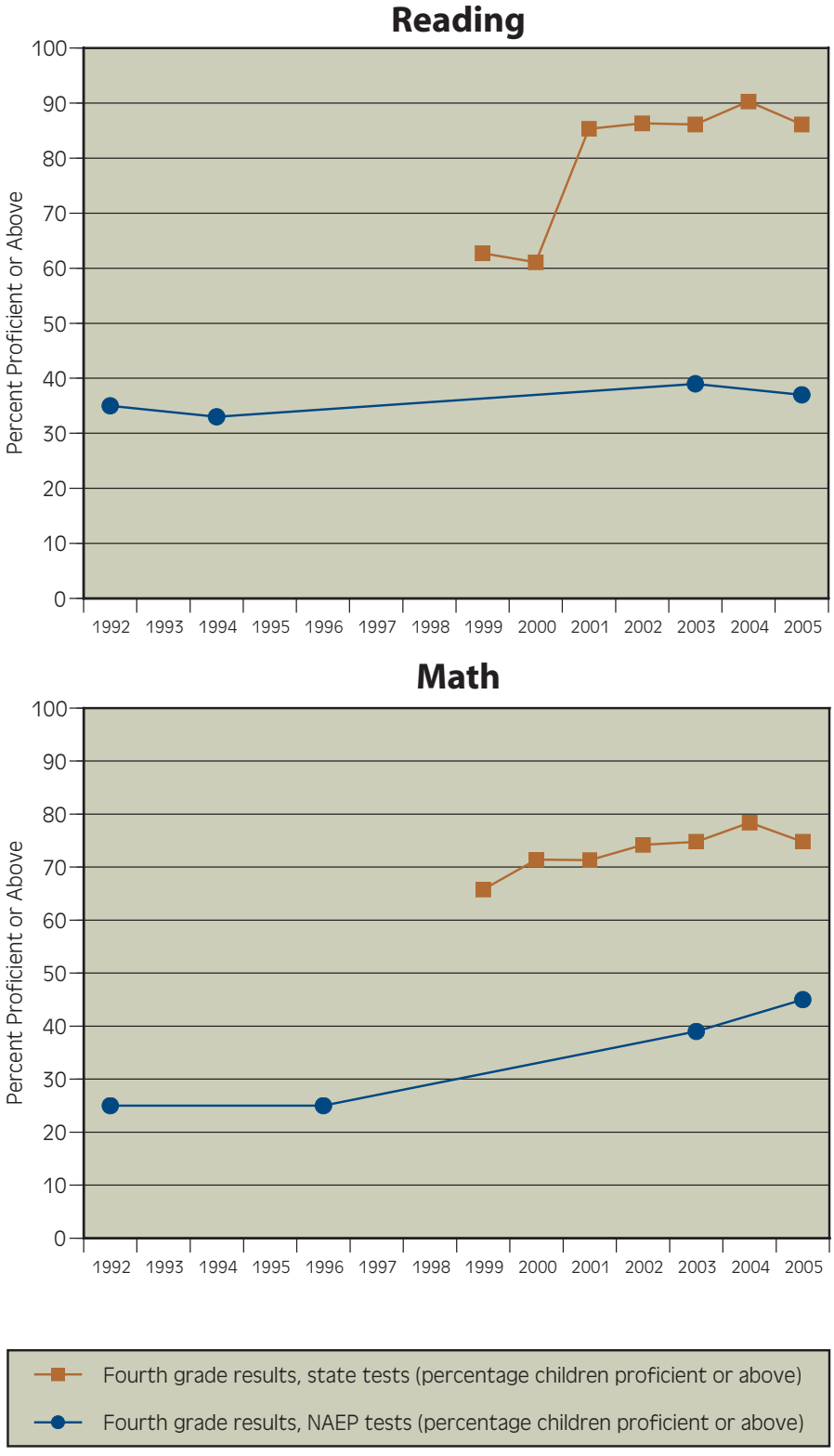


FIGURE A9. North Carolina – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

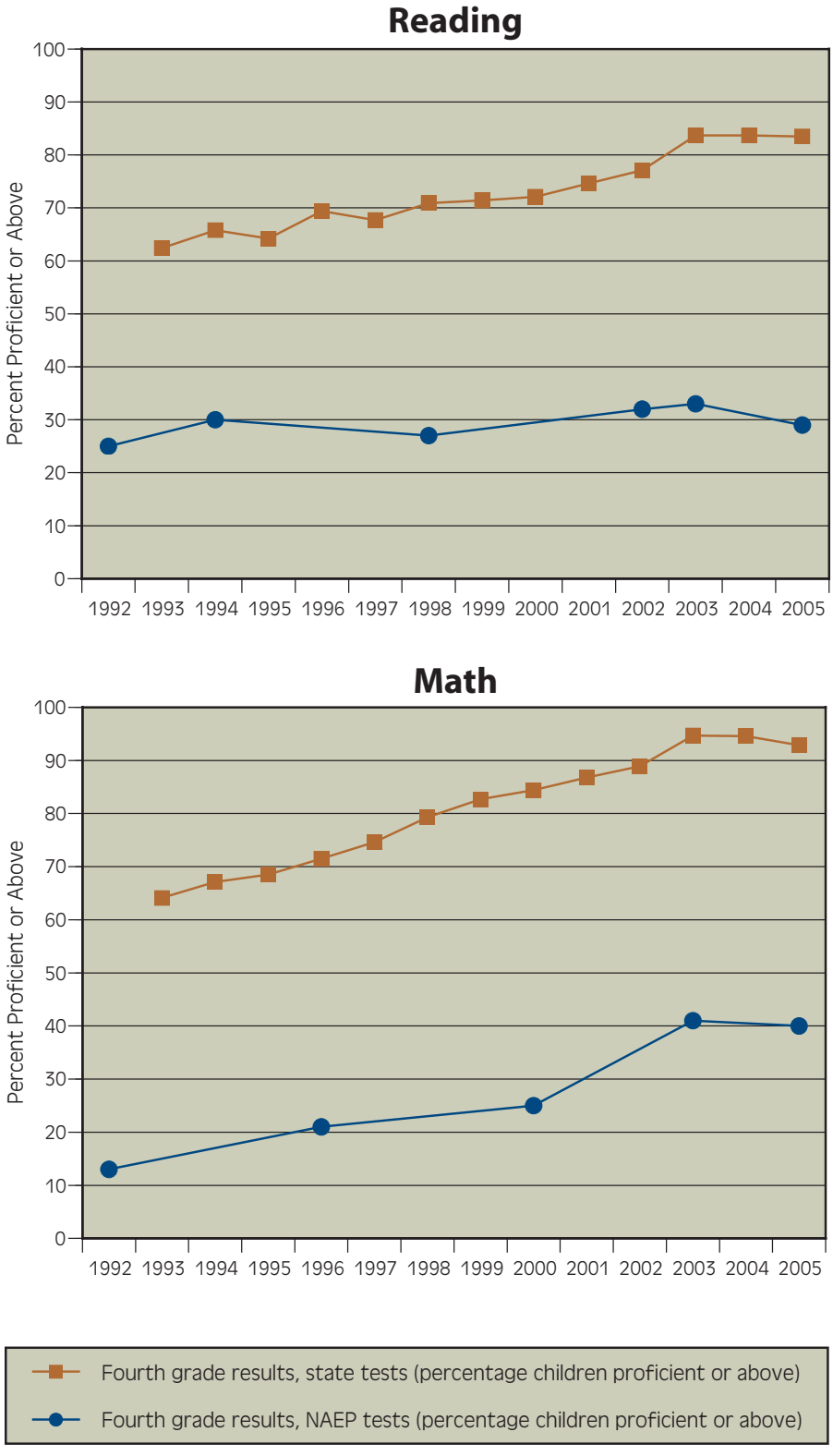


FIGURE A10. Oklahoma – Percentage of fourth- or fifth-graders proficient or above in reading and math, according to state and NAEP testing

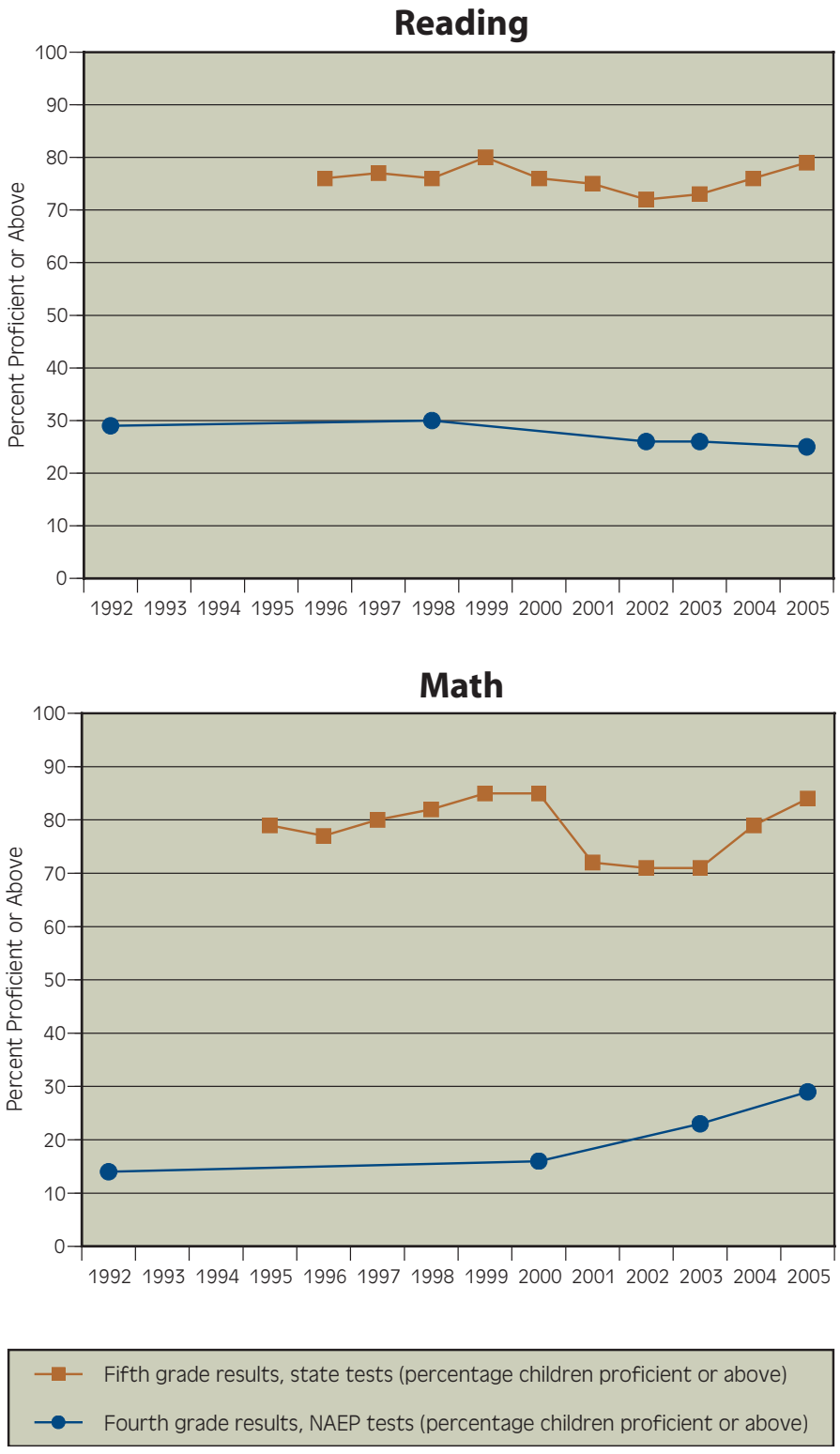


FIGURE A11. Texas – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing

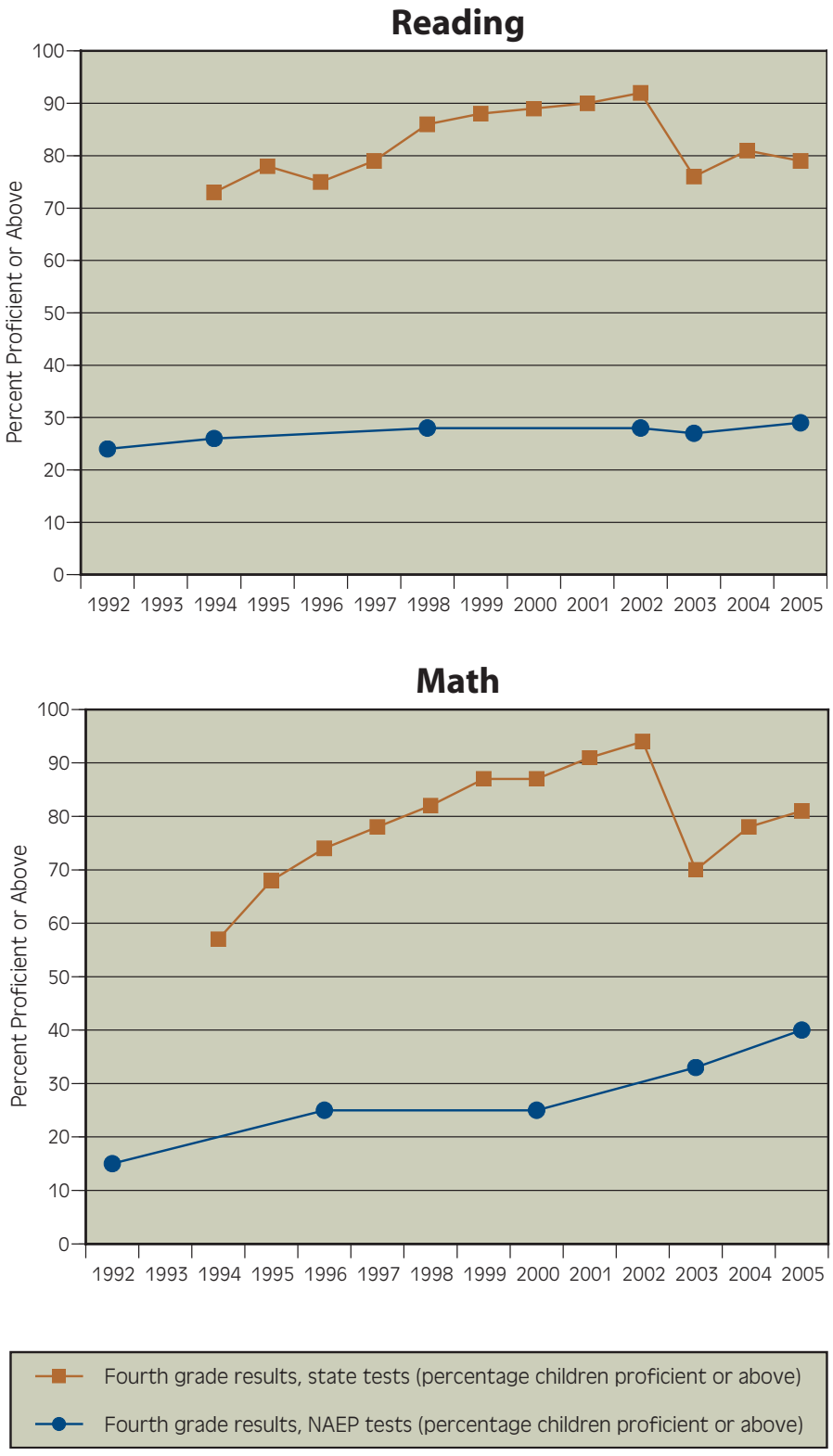
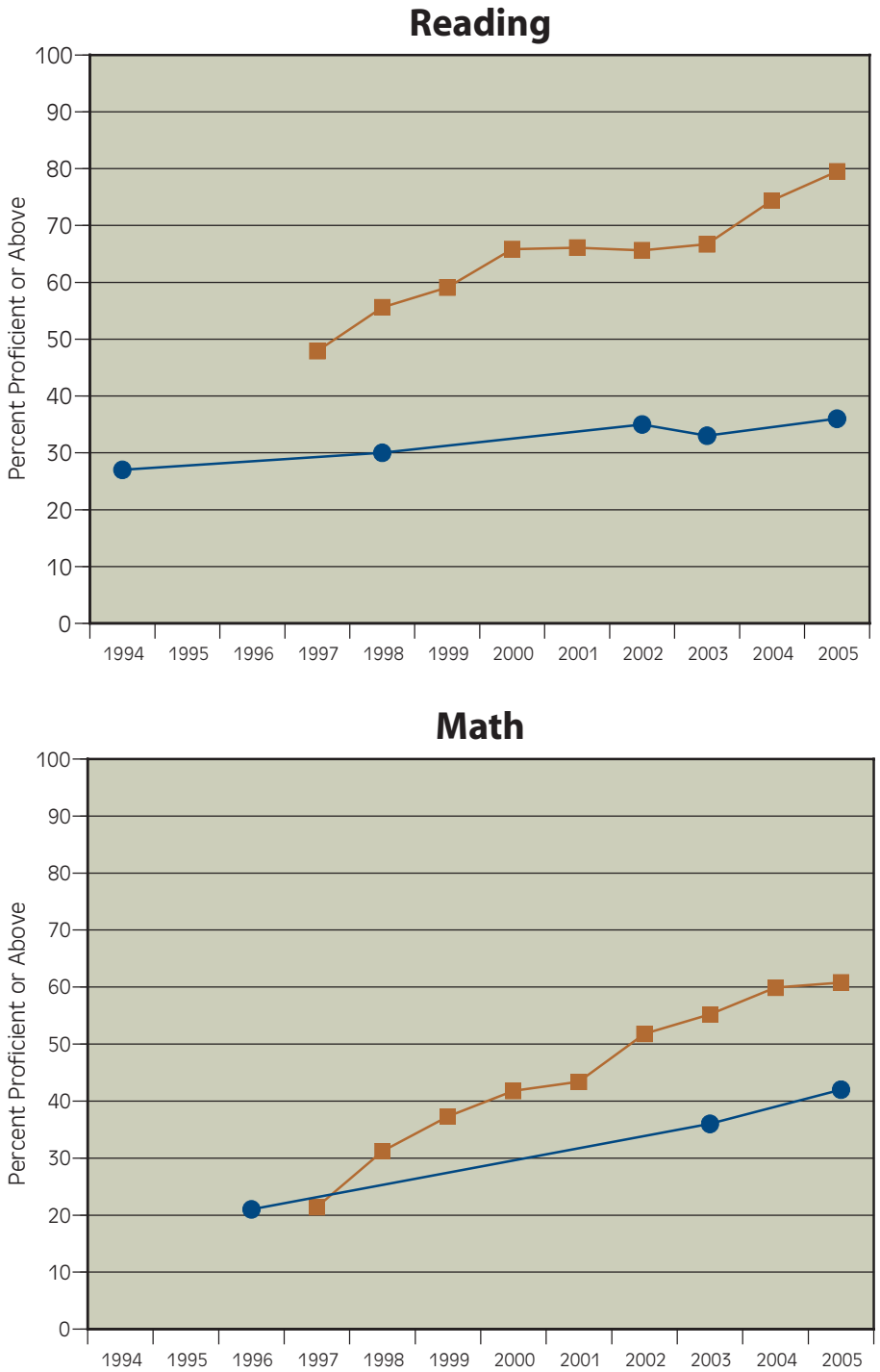


FIGURE A12. Washington – Percentage of fourth-graders proficient or above in reading and math, according to state and NAEP testing



Appendix 2

Sources for test data and state accountability policies, 1992-2005

Arkansas

Arkansas Department of Education (most recently accessed November 2005)
http://arkedu.state.ar.us/whats_new/benchmark_exams.html

Arkansas School Information Site (most recently accessed November 2005)
http://www.as-is.org/reportcard
http://www.as-is.org/indicators/search_actap.php

Arkansas Democrat-Gazette Northwest Edition, Posted on Sunday, October 9, 2005
http://www.nwanews.com/story.php?paper=adg&section=Editorial&storyid=131066

California

California Department of Education (most recently accessed November 2005)
http://www.cde.ca.gov/nr/ne/yr05/documents/star1.pdf (2005)

Colvin, R. (1995). State's reading, math reforms under review as scores fall, *Los Angeles Times,* March 23, p.1.

EdSource (most recently accessed October 2005). State testing prior to 1995: California laws and policies. Palo Alto.

Goertz, M., Floden, R., & O'Day, J. (1996). Systemic reform, archived information: California. Philadelphia: Consortium for Policy Research in Education.

Massachusetts

Massachusetts Department of Education (most recently accessed October 2005)
http://www.doe.mass.edu/mcas/2005/results/summary.pdf

Nebraska

Nebraska Department of Education (most recently accessed September 2005)
http://reportcard20032004.nde.state.ne.us/DownloadFiles/ReportCard20032004_Inside.pdf

New Jersey

New Jersey Department of Education (most recently accessed November 2005)
http://www.state.nj.us/njded/assessment/es/
http://www.state.nj.us/njded/schools/achievement/2005/njask4/
http://www.state.nj.us/njded/assessment/es/sample/reports/StateDissagg.pdf

North Carolina

North Carolina State Board of Education (most recently accessed November 2005)
http://www.ncpublicschools.org/docs/accountability/testing/reports/green/gb0405.pdf

Illinois

Illinois State Board of Education, Standards and Assessment Division
http://www.isbe.net/assessment/isat.htm (most recently accessed September 2005)

Illinois State Board of Education (most recently accessed September 2005)
http://www.isbe.state.il.us/assessment/pdfs/isat_guide_1999.pdf

Illinois State Board of Education (most recently accessed November 2005)
http://www.isbe.net/assessment/pdfs/isat_interpretive_guide_05.pdf

Iowa

Iowa Department of Education (most recently accessed November 2005)
http://www.state.ia.us/educate/fis/pre/coer/index.html

Kentucky

Elmore, Abelmann, & Fuhrman (1996).

Kentucky Department of Education (1999). Gender and race subgroup performance difference in KIRIS accountability cycle 2 and cycle 3 results.

Kentucky Department of Education (most recently accessed November 2005)
http://www.education.ky.gov/KDE/Administrative+Resources/Testing+and+Reporting+/Reports/Kentucky+Performance+Reports/default.htm
http://apps.kde.state.ky.us/cats_reports/index.cfm?action=display_regionstate

Koretz & Barron (1998).

Oklahoma

Oklahoma State Department of Education (most recently accessed November 2005)
http://title3.sde.state.ok.us/studentassessment/2005results/reportcard2005state.pdf

Oklahoma Educational Indicators Program (May 2005)

Texas

Klein, Hamilton, McCaffrey, & Stecher (2000).

Texas Education Agency (most recently accessed April 2005)
http://www.tea.state.tx.us/student.assessment/reporting/results/swresults/august/g4all_au.pdf

Texas Education Agency (most recently accessed October 2005)
http://www.tea.state.tx.us/student.assessment/reporting/results/swresults/taks/2005/gr4_05.pdf

Washington State Report Card, Office of Superintendent of Public Instruction (2005)
http://reportcard.ospi.k12.wa.us/summary.aspx

National and multi-state data reports

Education Trust (2004). Measured progress: Achievement rises and gaps narrow, but too slowly. Washington, D.C.

Koretz (1986).

Massell, Kirst, & Hoppe (1997).

National Center for Education Statistics (most recently accessed October 2005)
http://nces.ed.gov/nationsreportcard/states/profile.asp

PACE, Policy Analysis for California Education (2004)

Appendix 3

State Policy Milestones—School Accountability and Budget Reforms

Arkansas

1983: Revenue boost for schools under Governor Bill Clinton. Standards based education is implemented along with Minimum Performance Test in grades 3, 6 & 8.

1991: Benchmark Exams begin. Outcome Based Education (OBE) implemented.

1992: Shift from Metropolitan nationally-normed test to the SAT-8/9.

1995: Teaching Standards Adopted. Legislation causes a Teacher Licensure Taskforce to be created under the direction of the State Board of Education.

1996-7: "Smart Start" begins.

- All children to meet or exceed grade-level requirements in reading and math by Grade 4.
- "Report cards" up to 14 pages long, with rubrics (sets of assessment criteria) instead of subject titles. Students are evaluated and assigned a "performance level" of "below basic," "basic," "proficient" or "advanced" on each of the rubrics.

California

1990: Governor George Deukmejian vetoes appropriations for the California Assessment Program (CAP).

1991: Governor Pete Wilson enacts the California Learning Assessment System (CLAS); coupled with state curriculum frameworks and mainly focusing on open-ended questions.

1994: CLAS discontinued by Governor Wilson after certain groups claim the literary sections promoted inappropriate values; no testing system in place for the next 4 years.

1997: Legislature authorizes the Standardized Testing and Reporting (STAR) program for English language arts and math in grades 2-11 and in history-social science and science in grades 9-11; State Board of Education designates the *Stanford Achievement Test, Ninth Edition (Stanford 9).*

1998: Testing begins; new set of academic content standards created.

1999: The *Stanford 9* is augmented with *California Standards Test (CST)* to measure the achievement of students on the state content standards.

2000: National ranking for per-pupil expenditures in public elementary and secondary schools falls by 18 points from 1969 to 1999.

2001: The CST's in English-language arts in grades 4 and 7 include a writing assessment.

2003: The CST's in English-language arts for grades 2-11 and in math for grades 2-7 separated from the *Stanford 9*.

Illinois

1988-1998: The norm-referenced Illinois Goal Assessment Program (IGAP) measures achievement of Learning Goals of 1985.

1992: The Academic Watch List enacted into law in Illinois; schools placed on the Academic Watch List are subject to having personnel replaced or the reassignment of students to another school.

1996: The Academic Early Warning List becomes law; schools with low or declining assessment results are identified as not having met state standards and are placed on the state's Academic Early Warning List; continued declining assessment results could result in a school being placed on the Academic Watch List.

1997: Illinois Learning Standards implemented; define what all students in all Illinois public schools should know and be able to do in the seven core areas as a result of their elementary and secondary schooling.

1999: Illinois Standards Achievement Test (ISAT) begins for grades 3, 5, 8, and 10, measuring the extent to which students meet the Illinois Learning Standards; IGAP testing ends.

2003: The Illinois Assessment Frameworks are designed to assist educators, test developers, policy makers, and the public by clearly defining those elements of the Illinois Learning Standards that are suitable for state testing.

2003: State Board of Education announces $6.9 million in grants to regional agencies to provide services to schools in academic difficulty.

2004: Enhanced ISAT assessments, under a new test contractor, to be administered to ALL grades 3-8 in reading and math in 2005-06.

2005: By 2006-07, the State Board of Education shall provide grants to the lowest performing school districts to help these districts improve parental participation.

Iowa

1993: ITBS (Iowa Test of Basic Skills) first implemented.

1997: The "Iowa Model" (House File 2272) passed in response to the 1994 ESEA creates the comprehensive school improvement plan (CSIP), assessment of all students aligned with standards, and the annual reporting requirements (APR). Bill amended in 1998 to give districts more authority in determining performance expectations.

1999: Waterloo school district challenges in court the new evaluation criteria for schools.

2001: Teacher Quality Bill passed. Outlines eight teaching standards that beginning and veteran teachers will have to demonstrate. Authority is given to DOE to develop specific criteria for standards. The Iowa State Board of Education adopts standards and model criteria in May of 2002.

2002: Supreme Court rules against Waterloo and affirms evaluation criteria.

2005: Secondary schools establish the Core Curriculum. The bill mandates the following actions:

- The State Board of Education will develop a model core curriculum, with consideration to the ACT recommended core curriculum.
- School districts are required to track and report data on all students' progress toward the model core curriculum. Reporting goes to students, parents, the community, and the Department.
- All students shall have a four-year plan that includes a sequence of courses and tracking toward the model core curriculum. This plan is reviewed and updated annually through grades 9-12.
- The Department of Education must convene a Data Definitions work group to review current definitions related to data collected by the DOE.

Kentucky

1989: Kentucky Supreme Court rules the financing of the state education system unconstitutional.

- The decision (*Rose v. Council for Better Education*) applied to the statutes creating, implementing and financing the system and to all regulations. The legislature was charged with recreating a new, equitable and adequate education system. Although the court did not specify a remedy, it declared schools shall be substantially uniform throughout the state and "... shall provide equal educational opportunities to all Kentucky children, regardless of place of residence or economic circumstances."

1990: KERA Reform Act passed. The legislation sets out to reconstruct the state system with a focus on school accountability. $300 million focused on accountability efforts (compared to $50 million previously)

1993: KIRIS testing begins.

· KIRIS widely viewed as a major innovation in state assessment programs due to its use of open-ended items, portfolios, and performance events. For a time, it did not use multiple-choice items.

1996: Core Content for Assessment established: curriculum guidance for all grades.

· Updated periodically (1998 version 3.0, 2007 (planned) version 4.0)

1997: KIRIS eliminated.

· Kentucky terminats the contract of *Advanced Systems* (now *Measured Progress*), which had worked with the state from the start to develop KIRIS. A computer error was discovered, producing lower-than-actual scores. Correcting the error caused the state to pay rewards to additional schools. The testing innovations mentioned above in 1993, as well as the rewards and sanctions attached to assessment results, were very controversial and fueled disaffection with KIRIS. (Fairtest Examiner, *Fall 1997*)

1998: KIRIS test replaced by CATS.

· CATS, designed by the State Board, does not depend on a single type of testing. For grade 4 the test includes multiple-choice, open-response, and writing questions and portfolios. .

Massachusetts

1988-1996: Massachusetts Education Assessment Program (MEAP) is instituted for grades 4, 8, 12. The test is designed to measure the effectiveness of the curriculum; no passing score and no individual student scores are provided.

1993: Education Reform Act of 1993 calls for improvements in funding, accountability, and statewide standards.

· Mandated a new statewide testing system, the Massachusetts Comprehensive Assessment System (MCAS), a "high-stakes" test based on the new curriculum standards.

1994: DOE creates "Common Core of Learning" a vision to hold students to high standards of achievement.

1995: State sets curriculum frameworks, establishing what students should know and be able to do.

1996: Board of Education approves a norm-referenced reading test for grade 3 and a GED test for grade 10.

1997: The Iowa Tests of Basic Skills (ITBS) in reading administered to grade 3 and the Iowa Tests of Educational Development are administered to grade 10.

1998: MCAS testing begins.

2000: National ranking for per-pupil expenditures in public elementary and secondary schools rose by 13 places from 1969 to 1999.

2004: MCAS remediation grants are reduced from $50 million in 2003 to $10 million in 2004.

2005: School funding court case, Hancock v. Driscoll, decided. Court rejects findings that the Commonwealth was failing to provide an adequate education to all Massachusetts children, including those from low-income districts, and rejects recommendations to revise the state school funding system.

Nebraska

Early 1990's: Nebraska's Deputy Commissioner of Education, Doug Christensen, releases the "High Performance Learning Model," promoting quality learning, equity, and accountability.

1998: Legislature adopts state content standards, L.E.A.R.N.S.

· L.E.A.R.N.S. (Leading Educational Achievement through Rigorous Nebraska Standards): academic content standards adopted by the State Board of Education.

2000: Legislative Bill 812 requires assessment and reporting of student performance on content standards

· STARS (School-based Teacher-led Assessment Reporting System): *not* a state-wide test but a state *system* of assessments where each school district designs their own measure.

· Legislature requires that the state department of education publish a state report card.

· National ranking for per-pupil expenditures in public elementary and secondary schools fell by 18 places from 1969 to 1999.

2000-01: Assessment of standards in reading, speaking and listening plus a statewide writing assessment pilot for grades 4, 8, and 11 begins.

2001-02: Assessment of standards in math for grades 4, 8, and 11 begins.

2003: All school districts are required by the legislature to adopt academic content standards for reading, writing, math, science, social studies, and history; may adopt the L.E.A.R.N.S. standards set by State Board of Education or local school districts can set their own standards that are equal or more rigorous.

2003: The State Board of Education issues "essential education" guidelines that define what constitutes an adequate education for meeting standards and preparing students for the future.

2004: State Board of Education sets student achievement targets for schools and other related goals including the following:

- Increasing the percentage of children proficient on state standards by 2 percent each year through 2010.
- Providing professional development for all Nebraska teachers over three years, from 2005-2008.
- Assuring that all Nebraska schools develop and use high-quality assessments to measure student learning.
- Assuring that all schools reach the required level of student performance.
- Assuring that all schools have quality school improvement and instructional improvement plans that include increased targets for student achievement.
- Consolidating and streamlining reporting systems.

2004: Twenty-three learning centers open, providing extra help with reading, math and other studies for students attending primarily high-poverty schools. The before-school and after-school programs are offered through the 21st Century Community Learning Centers. The federal funds for the grants are associated with No Child Left Behind and are administered by the Nebraska Department of Education.

2004: State board of education approves preliminary regulations requiring low-performing schools to improve both the quality of the assessments they use to measure learning and to improve student performance on state standards.

New Jersey

1996: Core Curriculum Content Standards (CCCS) are adopted, outlining what all New Jersey students should know and be able to do by the end of the 4th and 8th grades and upon completion of a NJ public school education.

1997: The Elementary School Proficiency Assessment (ESPA) is administered to Grade 4 to provide an early indication of student progress toward achieving the knowledge and skills identified in the Core Curriculum Content Standards (CCCS).

1998: *Abbott v. Burke* decision requires all districts affected by the court decision (poor, urban school districts) to implement a local accountability system.

2003: ESPA is replaced by the New Jersey Assessment of Skills and Knowledge (NJ ASK), a comprehensive, multi-grade assessment program. The results of the elementary-level assessments are intended to be used to identify students who need additional instructional support in order to reach the CCCS.

North Carolina

1992-93: Nationally-normed tests replaced with the "End-of-Grade" testing program; results represent student performance on the set curriculum

1995: The New ABC's of Public Education outline a framework for a dramatic restructuring of the North Carolina's public school system. Goals of framework: provide strong local school accountability, emphasize mastery of basic subjects and as much local decision making as possible. The assessment later became a high-stakes test

1996-97: The new School-based Management and Accountability Program (the ABC's) is put into law; implementation begins for grades K-8.

1999: The State Board of Education develops a student accountability standards policy to be effective 2000-01. Students in grades 3, 5, 8 who do not score a level III or above on the End-of-Grade test are required to have a personalized education plan (PEP) in an attempt to close the achievement gap

2002: State Board of Education approves revised achievement levels for math.

Oklahoma

1990: The Oklahoma Education Reform and Funding Act is passed, which reduces class sizes, mandates budget increases for public schools, and creates the Office of Accountability.

- Common education experienced a 97.8% increase in appropriations from FY'90 to FY'99. The actual funding years of HB 1017 (FY'91 to FY'95) increased common education's budget by approximately $558 million.

1990: Norm-referenced tests begin for grades 3, 5, 7, 9, and 11.

1991: Initial development of core curriculum, *Priority Academic Student Skills (PASS)*.

1992: Legislation mandates the development of criterion-referenced tests (CRT) for grades 5, 8, and 11.

1994-95: Norm-referenced tests are discontinued for grades 5, 9, and 11 and are replaced by criterion-referenced tests.

1995: Implementation of criterion-referenced tests, Oklahoma Core Curriculum Tests (OCCT), to measure the achievement of PASS.

1997: SB 81 requires all persons under the age of eighteen to be proficient in reading at the eighth grade level. HB 2017, The Reading Sufficiency Act, requires that, beginning in the 1998-99, 2nd and 3rd grade students be assessed for grade appropriate reading skills and a prescribed program of instruction.

1998: Reading assessment of students is expanded to kindergarten and 1st grade students in HB 2878 (Boyd B./ Williams) during the 1998 Legislative Session and an appropriation of $4.2 million in the FY-99 budget is made for the implementation of The Reading Sufficiency Act.

1998-99; 2000-01; 2001-02: Vendors supplying the CRT change.

2000: State lawmakers call for an Academic Performance Index (API) to measure success and initiate improvement in school and district performance.

2001: The Oklahoma School Performance Review (OSPR) program, administered by the Office of Accountability, is created to identify ways districts could to hold the line on costs, reduce administrative overhead, streamline operations, and improve educational services.

2003: Implementation of Oklahoma School Testing Program Act (OSTP).

Texas

1979: Texas Assessment of Basic Skills Test (TABS) begins for grades 3, 5, 9 and assesses basic skill competency.

1984: Texas Educational Assessment of Minimum Skills Test (TEAMS) begins. The test seeks to increase the rigor of state assessment and institutes individual student sanctions for performance at the exit level.

1990: Texas Assessment of Academic Skills Test (TAAS), a new criterion-reference program, begins. TAAS shifts focus from minimum skills to academic skills.

1992: TAAS includes grades 3-8 in reading and math.

1993: Creation of a new statewide-integrated accountability system that includes rating of campuses and districts.

1994: Alignment of passing standards at grades 3-8 entitled the Texas Learning Index (TLI).

1999: Development begins for the more rigorous Texas Assessment of Knowledge and Skills (TAKS) test. TAKS to be aligned with the state-mandated curriculum, the Texas Essential Knowledge and Skills, and requires 3rd, 5th and 8th graders to demonstrate proficiency on a state assessment test and achieve passing grades to advance to the next grade.

1999: In April, the deputy superintendent of the Austin school district, which had shown dramatic score improvements, is indicted for tampering with government records. In Houston three teachers and a principal are dismissed for prompting students during test sessions.

2000: National ranking for per-pupil expenditures in public elementary and secondary schools rose by 13 points from 1969 to 1999.

2002: Last administration of the TAAS test; TAKS is field-tested.

2003: TAKS becomes the new statewide assessment program.

Washington

1993: Washington State Legislature passes Education Reform Act (House Bill 1209). Bill establishes common learning goals for all students, academic standards, and assessment.

1993-96: Academic standards are developed in reading, writing, math, social studies, science, arts and health & fitness.

1996-01: The Washington Assessment of Student Learning (WASL) for reading, writing and math begins as a requirement for grades 4, 7 and 10. Teachers and community members oversee development of WASL.

2000: State Board of Education determines that the class of 2008 will be the first to meet new statewide graduation requirements. Requirements include: pass 10th-grade WASL, complete culminating project, create "High School & Beyond Plan" and earn minimum class credits.

2004: State Legislature puts into law the graduation requirements. Bill provides students five opportunities to take the 10th-grade WASL and earn a Certificate of Academic Achievement. It also calls for struggling students to receive individualized academic help and an alternative for students unable to show their skills on the 10th-grade WASL. Certificate of Individual Achievement created for special education students unable to take the WASL.

2004: Academic standards are refined and expanded. Grades K-10 to devise learning expectations for each core subject.

2006-08: Students in the class of 2008 take the WASL as sophomores.

Enacted for 2008: First class to meet new statewide graduation requirements.

References

Anderson, N. (2005). Bush Administration grants leeway on 'No Child' rules, *Washington Post*, November 22, A01.

Blank, R., & Schilder, D. (1991). State policies and state role in curriculum. Pp. 37-62 in *The politics of curriculum and testing: The 1990 yearbook of the politics of education association*, edited by S. Fuhrman & B. Malen. New York: Falmer Press.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, 24, 305-331.

Catterall, J., Mehrens, W., Ryan, J., Flores, E., & Rubin, P. (1998). Kentucky Instructional Results Information System: A technical review. Frankfort, KY: Kentucky Legislative Research Commission.

Cohen, M. (1990). Key issues confronting state policymakers. Pp. 251-288 in *Restructuring schools: The next generation of educational reform*, edited by R. Elmore. San Francisco: Jossey-Bass.

Colvin, R.L. (1995). State's reading, math reforms under review as scores fall, *Los Angeles Times*, March 23. Retrieved from the Web: http://pqasb.pqarchiver.com/latimes/22744700.html?MAC=6fb1729.

Congressional Budget Office (1986). Trends in educational achievement. Washington D.C.

Cronin, J., Kingsbury, G., McCall, M., & Bowe, B. (2005). The impact of the No Child Left Behind Act on student achievement and growth, 2005 edition. Lake Oswego, OR: Northwest Evaluation Association.

Cross, C. (2004). *Political education: National policy comes of age*. New York: Teachers College Press.

Dillon, S. (2005a). Education law gets first test in U.S. schools, *New York Times*, October 20, Web archives, www.nytimes.com/2005/10/20/national/20exam.html?pagewanted=print.

Dillon, S. (2005b). Students ace state tests, but earn D's from U.S., *New York Times*, November 26, 2005, p.1.

Education Trust (2004). Measured progress: Achievement rises and gaps narrow, but too slowly. Washington, DC: Author.

Education Week (2005). No small change: Targeting money toward student performance (*Quality Counts* supplement), table entitled, "student achievement," January 6, p.84.

Elmore, R., Abelman, C., & Fuhrman, S. (1996). The new accountability in state education reform: From process to performance. Pp. 65-98 in *Holding schools accountable: Performance-based reform in education*, edited by H. Ladd. Washington, D.C.: Brookings.

Fuller, B. (2004). Are test scores really rising? School reform and campaign rhetoric, *Education Week*, October 13, pp. 40, 52.

Goertz, M., Duffy, M. with Carlson Le Floch, K. (2001). Assessment and accountability systems in 50 states: 1999-2000. Philadelphia: Consortium for Policy Research in Education (RR-046).

Grissmer, D. & Flanagan, A. (2001*). Improving student achievement: What state NAEP test scores tell us*. Santa Monica: RAND Corp.

Hall, D., & Kennedy, S. (2006). Primary progress, secondary challenge: A state-by-state look at student achievement patterns. Washington D.C.: Education Trust.

Hurst, D., Tan, A., Meek, A., Sellers, J., & McArthur, E. (2003). Overview and inventory of state education reforms: 1990-2000. Washington, D.C.: United States Department of Education, Institute for Education Sciences.

Klein, S., Hamilton, L., McCaffrey, D., & Stecher, B. (2000). What do test scores in Texas tell us? Santa Monica: RAND Corp. (Issue Paper, October 24).

Koretz, D. (in press). Alignment, high stakes, and the inflation of test scores. To appear as Chapter 7 in *Uses and misuses of data in accountability testing*, edited by J. Herman & E. Haertel, Yearbook of the National Society for the Study of Education.

Koretz, D., personal communication, January 22, 2006.

Koretz, D. & Barron, S. (1998). The validity of gains in scores on the Kentucky Instructional Results Information System (KIRIS). Santa Monica: RAND Corp.

Koretz, D., Linn, R., Dunbar, S., & Shepard, L. (1991). The effects of high-stakes testing: Preliminary evidence about the generalization across tests. Paper presented at the American Educational Research Association, Chicago.

Linn, R. (2000). Assessments and accountability. *Educational Researcher*, *29*, 4-16.

Linn, R. (2001). The design and evaluation of educational assessment and accountability systems. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (UCLA, technical report 539).

Loveless, T. (2003). How well are American students learning? Washington, DC: Brookings Institution, Brown Center report on American education, 2003.

Massell, D., Kirst, M., & Hoppe, M. (1997). Persistence and change: Standards-based reform in nine states. Philadelphia: Consortium for Research in Education Policy, University of Pennsylvania.

McDonnell, L. (2005). *Politics, persuasion, and educational testing*. Cambridge, MA: Harvard University Press.

McDonnell, L. & Choisser, C. (1997). Testing and teaching: Local implementation of new state assessments. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

NAEP, National Assessment of Educational Progress (2005). The nation's report card: Reading (and mathematics) results: Executive summary for grades 4 and 8. Washington, D.C.: United States Department of Education. On the Web http://nces.ed.gov/nationsreportcard/.

Olson, L. (2006). A decade of effort, *Education Week – Quality Counts*, January 5, 8-16.

PACE, Policy Analysis for California Education (2004). Children's reading scores stalled in major states. Berkeley: University of California, policy brief.

Paige, R. (2004). Statement in regard to the PACE study. Washington, DC: United States Department of Education, Office of Public Affairs, October 8, p.1.

Perie, M., Grigg, W., & Dion, G. (2005). *The Nation's Report Card: Mathematics 2005* (NCES 2006-453). Washington D.C.: National Center for Education Statistics.

Perie, M., Grigg, W., & Donahue, P. (2005). *The Nation's Report Card: Reading 2005* (NCES 2006-451). Washington D.C.: National Center for Education Statistics. State-level trend data appear on: http://nces.ed.gov/nationsreportcard/states/profile.asp.

Resnick, L. (1987). *Education and learning to think*. Washington, D.C.: National Academy Press.

Rhoten, D., Carnoy, M., Chabron, M., & Elmore, R. (2003). The conditions and characteristics of assessment and accountability. Pp. 13-54 in *The New Accountability: High Schools and High-Stakes Testing*, edited by M. Carnoy, R. Elmore, & L. Siskin. New York: Routledge Falmer.

Romano, L. (2005). Test scores move little in math, reading: Improvement appears slight since No Child Left Behind. *Washington Post*, October 20, Web archives: www.washingtonpost.com/wp-dyn/content/article/2005/10/19/AR2005101900708.

Sigman, D., & Zilbert, E. (2006). Personal communication, comments from the California state department of education, March.

Skinner, R. (2005). State of the states. Pp. 77-80 in *No small change: Targeting money toward student performance, Education Week*, January 6.

Smith, M., & O'Day, J. (1991). Systemic school reform. Pp. 233-267 in *The politics of curriculum and testing: The 1990 yearbook of the politics of education association*, edited by S. Fuhrman & B. Malen. New York: Falmer Press.

Stecher, B., Hamilton, L., & Naftel, S. (2005). Introduction to first-year findings from the Implementing Standards-based Accountability Project. Santa Monica: RAND Corporation (working paper series).

Stetcher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. Pp. 79-100 in *Making Sense of Test-Based Accountability in Education*, edited by L. Hamilton, B. Stetcher, and S. Klein. Santa Monica: RAND Corporation.

Treisman, U. (2005). Building instructional capacity in large urban districts. Lecture presented, November 21, Berkeley.

Zilbert, E. (2006). Confounding of race and ethnicity and English learner status in the reporting of NAEP data: A five state study. Paper presented at the American Educational Research Association, San Francisco, April.

Endnotes

[1] The history of the Texas testing system, including its role within the accountability regime, is detailed by Rhoten, Carnoy, Chabron, & Elmore (2003).

[2] In Missouri the share of fourth-graders labeled proficient in reading based on state exams equaled the NAEP percentage, 34 percent.

[3] Retired congressional staffer, Christopher Cross (2004), details how Cohen and Smith infused the 1994 amendments to Title I with elements of their systemic reform model, including achievement standards, an attempt to outlaw rote instruction of low-level skills in pull-out programs, and reports on student progress.

[4] These analysts included the presence of state curricular standards, the degree of alignment with state tests and the use of short answer or "extended responses" on state tests (*Education Week*, 2005:86-87).

[5] Looking across states, NAEP scores vary significantly, about one-third of a school year between high and low-performing states, controlling on demographic features of the states (Grissmer & Flanagan, 2001).

[6] For these 2003 data the bivariate correlation (r) equals just 0.20, meaning that only four percent of the variance in NAEP reading scores can be accounted for by state scores. The correlation for math is lower, equaling 0.17.

Additional PACE research papers on federal and state-led school accountability appear below and on the web: pace.berkeley.edu

- Tom Timar (2004) Categorical school finance: Who gains, who loses. PACE working paper 04-2.
- Bruce Fuller (2004) Are test scores really rising? Commentary in ***Education Week***, Oct. 13. edweek.org.
- Elisabeth Woody, Jennifer O'Day and colleagues (2004) Assessing California's accountability system: Successes, challenges, and opportunities for improvement.
- Elisabeth Woody and colleagues (2004) Voices from the field: Educators respond to accountability.
- John Novak and Bruce Fuller (2003) Penalizing diverse schools? Similar test scores, but different students, bring federal sanctions.

PACE

Policy Analysis for California Education
University of California, Berkeley and Davis
Stanford University