Gauging Growth: How to Judge No Child Left Behind?

by Bruce Fuller, Joseph Wright, Kathryn Gesicki, and Erin Kang

Many policymakers feel pressure to claim that No Child Left Behind (NCLB) is boosting student performance, as Congress reconsiders the federal government's role in school reform. But how should politicians and activists gauge NCLB's effects? The authors offer evidence on three barometers of student performance, drawing from the National Assessment of Educational Progress (NAEP) and state data spanning the 1992–2006 period. Focusing on the performance of fourth graders, where gains have been strongest since the early 1970s, the authors find that earlier test score growth has largely faded since enactment of NCLB in 2002. Gains in math achievement have persisted in the post-NCLB period, albeit at a slower rate of growth. Performance in many states continues to apparently climb. But the bar defining proficiency is set much lower in most states, compared with the NAEP definition, and the disparity between state and federal results has grown since 2001. Progress seen in the 1990s in narrowing achievement gaps has largely disappeared in the post-NCLB era.

Keywords: accountability policy; No Child Left Behind

Setting aside the spirited debates that now engulf No Child Left Behind (NCLB), most analysts agree on one basic fact: Political leaders feel growing pressure to claim that this bundle of centralized reforms is working, as Congress reviews NCLB's impact and tries to craft a more effective federal role.

Just 2 years after signing NCLB into law, President Bush began claiming that this ambitious initiative was already boosting student achievement. During his weekly radio address, Mr. Bush (2004) said, "We have recently received test results that show America's children are making progress." By early fall, as his reelection campaign was heating up, Bush sounded even more upbeat. "We're making great progress. We're closing the achievement gap," he alleged in a speech delivered in King of Prussia, Pennsylvania (Hernandez, 2004).

But analysts could find just one number backing the president's claim: a 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 observed in 2000. But students had sat for the 2003 exam in the first full year subsequent to the January 2002 enactment of NCLB. And math scores, having begun their ascent back in 1986, were likely buoyed by earlier state-led, not necessarily federal, accountability reforms (Loveless, 2003).

Educational Researcher, Vol. 36, No. 5, pp. 268–278
DOI: 10.3102/0013189X07306556

More sobering news arrived a year later when fresh NAEP scores stemming from 2005 testing were released. They showed how the nation's students had performed over the 3 school years following NCLB's enactment. Reading scores among fourth graders remained flat, with 31% of the nation's children at or above proficient in 2002, 2003, and 2005 (Perie, Grigg, & Donahue, 2005). The share of eighth-grade students proficient or above in reading had fallen 2 percentage points. The percentage of fourth graders proficient in math continued to climb between 2003 and 2005. Math scores at the eighth-grade level had reached a flat plateau.

The Administration highlighted another bright spot—rising performance among Black fourth graders. "It shows there's an achievement gap in America that's closing" (Dillon, 2005). But veteran Washington analyst Jack Jennings, referring to earlier state-led reforms that took root in the 1990s, 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" (Dillon, 2005).

Still, Bush was buoyed by findings earlier this year, detailing how state test scores (not NAEP scores) have climbed in some states since NCLB's enactment (Center for Education Policy, 2007). The president again voiced his upbeat inference during a White House ceremony for young scholars. "No Child Left Behind is working. . . . We are making good progress," said Bush (2007).

This intensifying joisting illustrates the slippery nature of trying to judge the broad effects of NCLB and prompts key empirical questions. First, what is the evidence on change in average (mean) test scores or regarding the possible narrowing of achievement gaps? Second, while NAEP results have been emphasized by many, governors and state school chiefs continue to highlight apparent gains in state test scores. But state trend lines often look like a jagged mountain range, erratically moving up and down as tests are changed and proficiency bars are moved. This article details these patterns, focusing on results at the fourth-grade level, given that 9-year-olds have been assessed by the NAEP since the early 1970s and their progress has been greater historically than for older students.

Third, the issue of timing is crucial. All but one state had in place a school accountability program prior to 2002. Several states registered solid gains during the 1990s. We know that achievement growth was more impressive in states that advanced more aggressive accountability programs (Carnoy & Loeb, 2002; Lee & Wong, 2004). A parallel argument by some in the civil rights community is that NCLB's distinct value added may be modest in states that already experienced gains, but federal action is required to reap similar benefits in southern or midwestern states that have legislated weaker standards-based accountability programs.

Some argue that it is too early to assess the effects of Washington's complex reform effort. Others argue that we must delve into various facets of implementation—operating across federal, state, and local levels—to understand whether this unprecedented set of federal rules and resources is touching the motivation of teachers and students. One study of NCLB implementation, led by RAND researchers, is beginning to illuminate action across organizational layers (Stetcher, Hamilton, & Naftel, 2005). Evaluations are emerging that focus on specific elements or programs advanced under NCLB, such as after-school tutoring or the conversion of failing organizations into charter schools. Yet as these debates over NCLB's overall benefits demonstrate, many will glean signs of progress or regress from state tests and federal NAEP results.

We begin with an historical overview, emphasizing how the interplay between the states and Washington, within a federalist structure of governance, has obscured and sometimes illuminated how achievement levels are changing over time.

Defining and Tracking Student Learning

The recent history of state testing programs unfolded 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, from the immediate postwar period forward, 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 gains on the Iowa Test of Basic Skills (ITBS) through much of the 1970s (Congressional Budget Office, 1986; Wirtz et al., 1977). Even SAT scores had floated upward since the late 1960s, after falling by a third of a standard deviation in the immediate postwar period as the GI Bill propelled a diverse range of students into higher education.

Koretz 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 assessments would have to provide "annual or nearly annual scores," equate scores to make them comparable over time, and test comparable groups of students each year (Congressional Budget Office, 1986, p. 100). The fact that the SAT results were not adjusted to take into account the influx of diverse GIs now taking this exam was an obvious case in point.

The rise of so-called systemic reform out in the states—crafting clearer learning standards and stronger pupil testing—was to help remedy these institutional shortcomings. Reform architects, such as Michael Cohen (1990) at the National Governors Association and Marshall Smith and Jennifer O'Day (1991) at Stanford University, began to articulate a state-led model of organizational change that called on governors and school districts to sharpen what children should learn and then to align state tests to these transparent standards.

The restructured system was to focus on achievement outcomes, measured by state education departments, rather than remaining preoccupied with how to best mix school inputs and classroom practices from afar. These fresh elements of systemic reform were borrowed, in part, from the writing of popular management gurus who reported how centralized managers in leading companies were focusing their efforts on tracking the performance of local units while decentralizing the means by which site managers pursued greater efficiency and innovation (e.g., Peters & Waterman, 1982).

Cohen (1990) emphasized how this streamlined education system would advance public accountability and greater political legitimacy. State spending on education had increased by 26% 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 said (p. 254). Fully 48 states had put in place statewide testing programs by 1999 (Goertz & Duffy, 2001). Detailing between-state variation in the implementation of accountability reforms is beyond the scope of this article.

States commonly adopted the NAEP-style reporting of achievement levels in ways that communicated student levels of mastery (criterion referenced) rather than percentile rankings that compared one state's scores to other states (norm referenced). Several years before NCLB required that all states set cut-points for proficient levels of performance, many states began to report their own test score results in these terms, beyond reporting scale scores or percentile rankings (Elmore, Abelmann, & Fuhrman, 1996).

Analysts began to compare by the late 1990s the share of students deemed proficient under state versus NAEP definitions, at least for the three grade levels included in the federal assessment. Caution is warranted in making such comparisons. First, most states have devised curricular guidelines and instructional materials that are painstakingly aligned with the state's standardized tests. The NAEP exams are not necessarily aligned to any state's intended curriculum.

Second, how states define proficient performance in terms of what knowledge must be demonstrated by students, and how cut-points are set across the distribution of scale scores, varies among the states and between state and NAEP procedures. The National Assessment Governing Board has established a general definition of proficiency: "Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter" (Brown, 2000, p. 15). This definition is then operationalized within specific grade-level NAEP exams; cut-points are set above which students are defined as proficient or not. This same process plays out differently across state testing authorities.

At the same time, the legislative crafters of NCLB assumed that proficiency meant a similar level of student performance across tests and states. The act mandates universal student proficiency by 2014 and requires that all states follow the mastery-oriented conception of basic, proficient, and advanced levels of student performance. We are not the first investigators to compare the corresponding shares of students determined to be proficient under state and NAEP assessment systems (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Lee & Wong, 2004).

Long-Term National Trends

Let us first look at national trends in NAEP scale scores, 1971 to 2004. This draws on the long-term trend series as opposed to the regular NAEP scale scores. The latter time series includes additional observation points, but the former series allows for valid comparison over the full 33-year period.

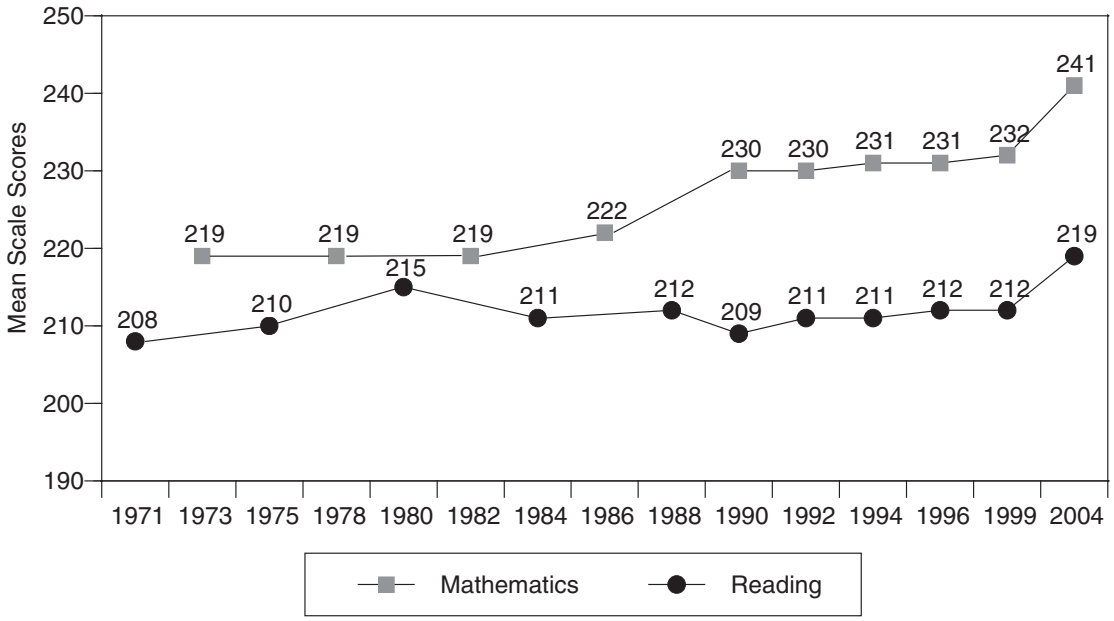

FIGURE 1. *Mean scale scores in fourth-grade reading and mathematics, long-term National Assessment of Educational Progress trends.*

Are Students Learning More?

Figure 1 displays long-term trend data for reading and math in fourth grade. We see that reading scores inched upward from 208 in 1971 to 219 in 2004, equal to about one grade level. Much of this growth occurred between 1999 and 2004, at the tail end of state implementation of their own accountability programs. On the other hand, eighth-grade scale scores for reading (not shown) remained essentially flat over the full period (National Assessment of Educational Progress [NAEP], 2006a; Smith, 2006).

The picture for math performance is more encouraging. Mean scale scores climbed from 219 in 1971 to 241 in 2004, about two grade levels. This rise was strong for eighth graders as well (not shown), moving from 266 in 1971 to 281 in 2004. Both reading and math scale scores remained flat over this entire period among 12th-grade students, dropping slightly in 2006. Overall, fourth graders have shown the most buoyancy in scores across the three grade levels that participate in the NAEP.

Are Achievement Gaps Closing?

Let us turn to results from the regular NAEP because it provides three additional data points since 1999 that are not available from the long-term data. Figure 2 displays fourth-grade reading patterns since 1992 for the nation's three largest ethnic groups (NAEP, 2006b). White students inched up half a grade level between 1992 and 2005, rising from a mean scale score of 224 to 229. Black and Latino subgroups fell between 1992 and 1994, then climbed well over one grade level by 2004.

The long-term trend data show a similar pattern and reveal that Black fourth graders improved two full grades levels, on average, between 1971 and 1988 (not shown), before dipping down in the early 1990s. Between 1971 and 2004, Blacks gained a remarkable 30 scale points, or about three grade levels. During the same period, however, Whites gained over one grade level. So, the long-term data show that the Black-White achievement gap closed from 44 scale points in 1971 to 26 points in 2004. Yet no progress has occurred since 2002 in closing Black–White or Latino–White gaps.

Gains in fourth-grade math have been strong for all three ethnic groups since 1990. Whites have gained over two grade levels, from 219 to 246 between 1990 and 2005. Black fourth graders have climbed by about three grade levels over the same period. In 2005, Black and Latino fourth graders performed at the same level that Whites performed back in 1990. But the Black–White gap remained unchanged, and the Latino–White gap closed slightly. Again we see that mean gains have slowed since 2003, and progress in closing ethnic gaps has stalled.[1]

Fifty State Proficiency Standards

Governors and education officials out in the states rarely talk about NAEP results, even though federal scores and proficiency data are now available for every state. Instead, they focus on year-to-year changes in scores that stem from their own state testing programs.

Setting Cut-Points and Test Inflation

Although NCLB mandates that all children be proficient in basic subjects by 2013, each state defines proficiency in its own unique way. Specific domains of knowledge covered in state tests, for instance, in English language arts or social studies, vary among states and within a given state over time (Catterall, Mehrens, Ryan, Flores, & Rubin, 1998; Linn, 2001). Earlier research has detailed how similar test items are used year after year, encouraging teachers to teach to a narrowing range of domains. At times, entire state tests are quietly circulated among teachers (Stetcher, 2002).

Some of these factors were at work when the reading performance of fourth graders appeared to climb dramatically over the first 2 years (1992–1994) of the Kentucky Instructional Results Information System (KIRIS)—by a stunning three quarters of a standard deviation (Koretz & Barron, 1998). But student

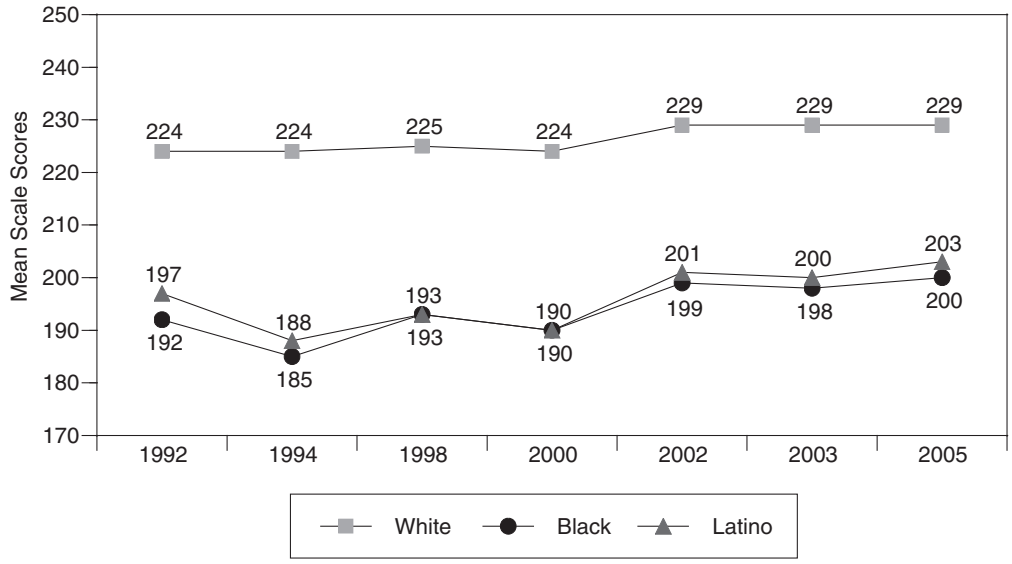

FIGURE 2. *Achievement gaps: mean scale scores in reading, fourth-grade regular National Assessment of Educational Progress trends by ethnic group.*

performance was unchanged for the same grade level on the NAEP assessment. Kentucky varied its test forms year to year, so teaching to the test could not fully explain the 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 achieving at a proficient level, compared with the benchmarks set by the NAEP governing board. Take the case of Alabama, where the state determined that fully 77% of all fourth graders could read proficiently or above in 2003, compared with just 22% as assessed by the NAEP. Our analysis 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.[2]

We do not assume that NAEP-determined conceptions of proficiency are necessarily optimal. They have yet to be validated against the actual knowledge demands pressed by colleges or employers. Linn (2006) notes that not one country participating in the Third International Mathematics and Science Study had more than 75% of its students achieving at the proficient level as defined under the NAEP standards (see also Linn, 2000). Nor has the NAEP governing board shown much interest in tracking achievement in ways that account for the evolving demographics of America's students (Zilbert, 2006). Still, the stable proficiency cut-points and the consistency of NAEP tests yield consistent trend lines, compared with the erratic trends stemming from states' own testing programs, as detailed below.

Tests Sensitive to Low Achievers

Where cut-points are set and the emphasis placed on certain curricular topics can yield tests that are disproportionately sensitive to gains made by low-performing students. This allows large numbers of students (at the low end of the distribution) to clear the basic or proficient hurdle with greater ease relative to the low level of discrimination at the middle or high end of the achievement distribution. This is in sharp contrast to other tests, like the SAT for college-bound students, which discriminates poorly across performance levels of low achievers while spreading out high achievers across high levels of the distribution.

RAND researchers detailed this dynamic when they examined the rise in scores on the Texas Assessment of Academic Skills (TAAS), a study that achieved notoriety after being released 2 months prior to the 2000 presidential election. The results were eye-opening, essentially following the pattern seen in Kentucky (yet in Texas the data were not analyzed at the item level as Koretz and Barron had done in Kentucky). The TAAS results showed that fourth-grade reading scores climbed, between 1994 and 1998, fully 0.31 of a standard deviation 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: 0.13, 0.14, and 0.14 standard deviations, respectively (Klein et al., 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." The team 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)" (Klein et al., 2000, p. 16).[3]

Interrupted Trend Lines

When a state changes its testing regimen, mean scores typically fall as the factors that tend to inflate results are temporarily suspended. Teachers do not know the test items to which they might teach, questions likely align with a new set of curricular domains and constructs, and the new test's novel format often constrains student performance in the short run. Linn (2000) has documented this sawtooth pattern of, at first, steady gains under Test A, followed by a sharp decline after the state shifts to Test B. The pattern then repeats as a third test replaces the second (for case studies, see Koretz, in press; Koretz, Linn, Dunbar, & Shepard, 1991).

Koretz (personal communication, January 22, 2006) emphasizes that fluctuating state testing results, including the lack of

association with NAEP proficiency estimates, is driven by multiple factors. The effect of setting proficiency cut-points low, or designing tests that are sensitive to small gains at the low end, can be set in place at one point in time. Then, inflation of scores may proceed as teachers and students become acquainted with the test. Separating the actual mastery of curricular topics by cohorts of students from inflation, under which sustained learning is illusory, is a slippery analytic task. Some states have tried to include more complex test items to combat inflation, such as the original KIRIS, which included open-ended items and short essays. But they were never used in the accountability system (Koretz & Barron, 1998; Massell, Kirst, & Hoppe, 1997).

Comparing State and NAEP Trends

Given these various sources of noise that may distort state testing results, how do proficiency levels compare between state and NAEP assessment programs? Do state results appear to be valid and useful when trying to understand whether students are learning more? And do state assessments help to inform the question of whether NCLB is truly making a difference in raising achievement levels? These questions have been examined within state-specific studies, as reviewed above, or for particular years (Fuller, Gesicki, Kang, & Wright, 2006; Lee, 2006; Skinner, 2005). This article extends and updates this earlier work by looking across the 1992–2006 period.

We set out to collect state data going back to 1992 that would track trend lines at the fourth-grade level. Building from the earlier work, we then examined whether states tend to claim stronger progress in raising the share of students who reach basic or proficient levels of performance, compared with federal NAEP results.

The NCLB mandate presently requires that all children must be proficient in basic subjects by 2014. Yet some analysts argue that state accountability programs will be more effective in raising students to basic levels, evidenced in part by the earlier gains in fourth-grade performance not observed at higher grade levels. So, we moved beyond earlier studies by examining trends at the basic level, as gauged by the NAEP, asking whether this bar may better match the average state's definition of proficient. This exercise may offer fresh options for how to bring state definitions of adequate student performance in line with federal NAEP standards.

Selecting States

Uncovering state scores going back to the early 1990s requires lots of digging. We discovered a lack of institutional memory and limited capacity in states' ability to even dig up earlier results. The shift from (norm-referenced) percentile scores to (criterion-referenced) percentage-proficient results also constrains states' ability to report long-term trends. Documents and electronic bulletins from state departments typically focus on the current picture or look back only 2 or 3 years.

Thus, a longitudinal comparison of state and NAEP results requires selecting a manageable sample of states. After choosing 12 diverse states, we spent 16 months contacting state education officials, fellow researchers, and education associations, along with searching newspaper archives, to construct time series for comparable test results. This took us back a decade prior to passage of NCLB and 4 years hence. Then we matched state-level NAEP results for reading and math over the same period, 1992 to 2006.

We took into account four criteria in selecting the 12 states. First, we endeavored to sample a diverse range of states that might illustrate differing patterns of fourth graders' performance over time.[4] Among candidate states, we examined population size, demographic diversity, and each state's urban or rural character. Second, we aimed to ensure a geographically dispersed set of states. Third, we selected states in which education departments or other sources could locate time-series data on student achievement, be it mean percentile scores or percentage at or above proficient as defined by the state. (Data sources from each state are available on the Web at pace.berkeley.edu/ER/State_Test_Scores.pdf.)

Fourth, we considered the intensity of state accountability programs during the 1990s. At least three research teams have reviewed the presence and strength of states' accountability policies (Carnoy & Loeb, 2002; Goertz & Duffy, 2001; "No Small Change," 2005, Student Achievement table). An important possibility is that NCLB may yield significant gains for students in states that maintain weak accountability efforts. Two accountability indices proved to be highly related; each included the 12 states that we eventually selected. Carnoy and Loeb (2002), for instance, awarded a zero to Iowa and Nebraska for their weak or nonexistent state accountability programs in the 1990s. The *Education Week* team gave these states an F and D, respectively ("No Small Change," 2005). Kentucky, New Jersey, and North Carolina earned index scores of 4, 5, and 5, respectively, under Carnoy and Loeb, whereas *Education Week* analysts awarded these states grades of A, B, and B.

We cannot formally associate, with a sample of just 12 states, the intensity of accountability systems with test score trends. But we can look at 3 states that displayed weak accountability regimes—Arkansas, Iowa, and Nebraska—to see if their achievement trends differ from other states. We do not attempt to generalize our findings to the nation, nor do we use inferential statistics to move from results in our final sample of 12 to advance claims about all states.

Disparities in State Versus Federal Proficiency Estimates

Table 1 reports on the mean difference over the 1992–2006 period between the share of fourth graders that state education departments deem proficient or above, stemming from their testing programs, versus the share estimated to be proficient or above under the federal NAEP exams administered in each state.

Our results replicate and extend earlier analyses, detailing how state cutoffs that define which students are proficient or above are set lower, often far lower, than the thresholds set by the NAEP governing board. Importantly, these gaps are not new. The chasm between state and federal proficiency estimates was apparent long before NCLB was enacted in 2002. These disparities did not stem solely from NCLB's incentive for states to define proficiency at low levels to ease the mandated pathway toward universal student proficiency. The maximum time period analyzed was 1992 to 2006, but several states began reporting percentage proficient later than 1992 (see Table 2 notes).

In Kentucky, for example, the average share of fourth graders reported at proficient or above in reading equaled 31 percentage points higher (annual average) under state testing, compared with the share derived from the NAEP assessment in Kentucky (columns 1 and 2). The state-NAEP gap in reading has averaged 56 points in Texas and 52 points in math since 1994. State tests in Massachusetts have yielded the closet share of students deemed to be proficient or above, compared with NAEP results in reading: The mean annual

Table 1
Gaps Between the Percentage of Fourth Graders Defined as Proficient According to State Versus National Assessment of Educational Progress (NAEP) Testing Results, 1992–2006

States Sorted by Strength of Accountability Policies[a]	Mean Annual Gap in the Percentage of Fourth Graders Deemed Proficient or Above (State Minus NAEP)	
	Reading	Math
Kentucky (A)	31	18
Massachusetts (A)	10	1
California (B+)	20	25
Oklahoma (B+)	51	60
Illinois (B)	35	47
New Jersey (B)	42	36
North Carolina (B)	43	54
Washington (B)	33	19
Texas (C+)	56	52
Arkansas (C)	27	28
Nebraska (D)	46	49
Iowa (F)	38	45

[a]As determined by *Education Week* (2005).

gap between the two gauges equaled 10 percentage points, and there was just 1 percentage point difference in math.

Differing Growth Rates

One could imagine that states set lower proficiency bars when pegged against federal NAEP standards but that growth rates look similar over time. This proves not to be the case when tracking trends in reading proficiency. For math, the state and NAEP trend lines are more closely parallel. Table 2 reports on the mean annual change in the percentage of children deemed proficient or above in reading under state test results for the most recent continuous time series prior to the 2001–2002 school year, along with the same estimates derived from NAEP results (Table 2, columns 1 and 2). Kentucky's reading scores, for instance, showed a 1.3 yearly point increase, on average, in the percentage of fourth graders deemed proficient or above based on state test scores. But NAEP results showed an annual mean increase of just 0.7 percentage points.

Two states displayed sharp gains in the pre-2002 period, including New Jersey, where state reading scores climbed 7.9 percentage points annually. This includes a remarkable 24.2 percentage point jump in the share 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 proficient or higher in reading between 2001 and 2002, contributing to this sizeable average yearly growth rate.

Annual progress in math performance is more consistently gauged by state exams, relative to NAEP results (Table 2, columns 3 and 4). North Carolina's mean annual gain of 2.8 percentage points tracked well against a mean rise of 2.3 percentage points under NAEP results. Still, the inflation of test results over time is vivid for some states. Washington's state tests yielded a 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. Overall, the inflation of state results does not result only from setting the proficiency bar far below the federal standard at the inception of a state testing program. In addition, gains reflected in state test results often exceed modest improvements (or no change), as determined by NAEP assessments.

We report weighted means for the 12 states, while noting that no attempt is being made to generalize to the population of all states. The means weighted by estimated K–12 enrollment in 2005 (National Center for Education Statistics, 2005, Table 3) are close to the unweighted means with one exception. The mean percentage-point rise in state math scores is even higher when weights are applied, largely due to California's above average rate of annual growth.

Gains Following Enactment of NCLB?

The final four columns of Table 2 report mean yearly changes in the post-NCLB period in the share of fourth graders at or above proficient as defined by the state versus NAEP. For example, the share deemed proficient in reading under Kentucky's state testing program, between 2001–2002 and 2005–2006, climbed 2.2 percentage points annually, on average. Yet the NAEP results showed this post-NCLB growth equaling just 0.3 percentage points per year. Washington state shows the least convergence during this period: The share of fourth graders deemed proficient in reading by state officials rose 4.0 percentage points annually, compared with the NAEP estimate of a 0.3 percentage point improvement each year.

Across the 12 states the (unweighted) mean rate of growth in percentage proficient in reading equaled 1.5 percentage points annually over the 4 school years following NCLB enactment. This compares to 0.2 percentage point average decline each year, based on NAEP results and the federal cut-point defining proficient. In short, the chasm between state and federal estimates of proficiency has grown wider since NCLB was signed into law (detailed in Fuller & Wright, 2007).

The Washington-based Center for Education Policy (2007) interprets these widening disparities as true gains in learning when pegged to each state's respective curricular standards. The center's findings, however, often rely on just 2 years of data pre-2002. The authors excluded breaks in time series when a state reset proficiency cut-points or changed its test—years when proficiency shares typically drop, as we saw above.

In addition, we reviewed how states like Kentucky and Texas set tests that are sensitive to small gains by low-performing students, whereas the NAEP assessment weighs progress more evenly across the distribution of student performance. Furthermore, state scores may rise as teachers teach to the test or after test forms circulate among teachers in subterranean fashion. Still, more work is required to understand what domains of reading or math are being captured by state, but not NAEP, exams.

We do find that change in state test results in mathematics has been more consistent with NAEP results post-NCLB. The mean rate of annual growth in math proficiency equaled 2.4 percentage points between spring 2002 and 2006 for our 12 states. This

Table 2

Comparing Test Score Trends Between State and National Assessment of Educational Progress (NAEP) Results for Fourth Graders, Most Recent Time Series, 1992–2006

States Sorted by Strength of Accountability Policies	Mean Annual Percentage Point Change in Percentage of Fourth Graders Deemed Proficient							
	Pre–NCLB Period				Post–NCLB Period (2002–2006)[a]			
	Reading		Math		Reading		Math	
	State	NAEP	State	NAEP	State	NAEP	State	NAEP
Kentucky (A)[b]	1.3	0.7	2.7	0.7	2.2	0.3	5.2	1.9
Massachusetts (A)[c]	3.0	1.1	1.3	1.5	–1.9	–1.0	0.2	3.8
California (B+)[d]	3.0	0.2	4.0	0.9	3.0	0.0	4.2	2.3
Oklahoma (B+)[e]	–0.7	–0.3	–1.1	0.7	3.0	–0.3	3.2	2.8
Illinois (B)[f]	0.7	—	2.0	4.0	1.2	–1.0	1.8	1.3
New Jersey (B)[g]	7.9	0.3	2.8	1.2	–0.1	–0.4	0.2	2.7
North Carolina (B)[h]	1.6	0.7	2.8	2.3	2.1	–1.0	1.3	1.5
Washington (B)[i]	3.5	1.0	6.1	2.1	4.0	0.3	2.0	2.7
Texas (C+)[j]	2.4	0.4	4.6	1.5	1.7	0.3	3.2	3.2
Arkansas (C)[k]	5.0	0.8	4.5	2.1	–0.1	1.3	1.7	4.0
Nebraska (D)[l]	3.5	0.3	—	0.9	1.5	0.0	3.5	1.8
Iowa (F)[m]	–0.1	–0.1	0.1	0.7	1.1	–0.7	1.9	1.4
Unweighted means	2.6	0.4	2.7	1.5	1.5	–0.2	2.4	2.4
Weighted means[n]	2.7	0.5	3.4	1.6	1.9	–0.2	2.8	2.5

[a]Average annual rates of change use interpolated values for NAEP percentage proficient and above and end with spring 2005 tests. State calculations run through spring 2006. Estimates are simple mean rates of change over the 4 school years after enactment of No Child Left Behind (NCLB).

[b]Kentucky Core Content Test, 2000–2005 (Grade 4 reading and Grade 5 math). The base year for the post-2002 NAEP trend for math is interpolated.

[c]Massachusetts Comprehensive Assessment System, 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.

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

[e]Oklahoma Core Curriculum Test, 1996–2005 (Grade 5 reading), 1995–2005 (Grade 5 math) percentage satisfactory or above; 1999–2002 "traditional students" only, 2003–2005. The base year for the post-2002 NAEP math trend is interpolated.

[f]Illinois Goal Assessment Program, 1992–1998 (Grade 3) percentage meeting or exceeding state goals; Illinois Standards Achievement Test, 1999–2005 (Grade 3) percentage meeting or exceeding Illinois learning standards. NAEP testing in reading did not begin until 2003. The base year for the post-2002 NAEP math trend is interpolated.

[g]Elementary School Proficiency Assessment (ESPA), 1999–2002; New Jersey Assessment of Knowledge and Skills (NJ-ASK), 2003–2004. The state does not distinguish between ESPA and NJ-ASK when reporting trend data, so we followed suit. Values are general scores (i.e., combined minus ESL and special education students) as opposed to total scores, which include all students. Between 2000 and 2001, a 24.2 point gain on the ESPA reading test was reported with no published changes to the cutoff values. The base years for the post-2002 NAEP math and reading trends are interpolated.

[h]End-of-Grade Testing Program, 1993–2005, percentage at achievement levels III and IV. The base year for the post-2002 NAEP math trend is interpolated.

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

[j]Texas Assessment of Academic Skills, 1994–2002, percentage meeting minimum expectations; Texas Assessment of Knowledge and Skills (TAKS), 2003–2005, percentage at state panel's recommended level of 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 math trend is interpolated.

[k]Arkansas Benchmark Exams, 1998–2005. Values are combined scores, including all students, as opposed to general scores (combined minus ESL and special education students). In 2005, new cut-points were set, but scores relative to the previous cut-points were obtained. The base year for the post-2002 NAEP math trend is interpolated. The post-NCLB growth estimate for state test scores is sensitive to the base year (2001–2002). Relative to the prior year (2000–2001), the share of fourth graders proficient in reading climbed 22 percentage points.

[l]School-Based Teacher-Led Assessment Reporting System, 2001 and 2003 (reading), 2002 and 2004 (math), percentage meeting or exceeding state standards. State testing in math did not begin until 2002; thus, data are not available for pre-2002 state math trends. The base year for the post-2002 NAEP math trend is interpolated.

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

[n]Weighted by the state's K–12 student enrollment in 2005 (National Center for Education Statistics, 2005).

is identical to the NAEP-derived estimate. Some states, however, continue to report higher annual rates of growth. Kentucky's testing program alleges that the percentage proficient or above in math has climbed at a 5.2 percentage point clip each year, on average, compared with the NAEP estimate of 1.9 percentage points annually. California officials have reported a rate of change in percentage proficient at 4.2 percentage points annually, compared with the NAEP estimate of 2.3 percentage points.

After weighting means by state enrollment, the basic pattern is more distinct. The yearly percentage-point climb in state-estimated proficiency shares for reading moves up to 1.9 points (from 1.5) and to 2.8 points (from 2.4) for math. Weighting by enrollments does not discernibly move annual NAEP changes.

Differences in Basic Versus Proficient Levels of Achievement

Figure 3 displays two common patterns observed among the 12 states, focusing on trends in reading performance (plots for all 12 states appear on the Web at pace.berkeley.edu/ER/State_Test_Scores.pdf).[5] We display the percentage of fourth graders deemed proficient or above according to state or NAEP testing. In addition, we include data points and trend lines for the share of students deemed at or above the basic level. Some analysts emphasize that state accountability programs and finance reforms focus on boosting the performance of low-achieving students. It may be that accountability efforts are moving more students over the basic hurdle than over the proficiency bar.

In Figure 3, the top panel displays trends for Arkansas, where state officials have reported percentage proficient or above since 1998. We see that state testing results yield an erratic trend line, ranging from 39% proficient or above in reading in 1998, climbing to 77% in 2004, then falling to 51% in 2005. This jagged trend line stems from changes in the Arkansas state testing program and presumably ongoing shifts in the cut-point that defines proficiency. The state standard for proficiency or above better approximates the NAEP standard for basic or above.

Importantly, the share of students who performed at the basic (but not proficient) level did not change in fourth-grade reading between 1992 and 2004, according to NAEP results. State accountability efforts may have paid off more in math, where the share of fourth graders at or above basic rose from 37% to 44% over the same 12-year period (NAEP, 2006c). The share of students at proficient or above did climb from 23% in 1992 to 30% in 2006 but at a rate of growth far slower than that reflected in state testing results.

Kentucky reveals another pattern in the second panel of Figure 3, where state test results track along a smooth trend line, sloping upward between 1998 and 2006 and often parallel to NAEP gains. But here, too, we see that the share of students at proficient or above on Kentucky's own tests roughly equals the share at basic or above according to the NAEP standard. This state showed no improvement on the share of fourth graders only at basic (not clearing the proficient bar) between 1992 and 2006 (not shown). But in math, this percentage increased from 38% in 1992 to 49% in 2006. The share at proficient or above in math doubled, from 13% to 26%, during the same period.

Texas exemplifies how changes in the state testing programs, and raising the cut-point defining proficiency, can lead to jolts in trend lines (not shown). After testing in spring, 2002, Texas officials reported that 92% of their fourth graders were proficient in reading, compared with a 29% estimate stemming from NAEP results. The following year, the percentage proficient fell to 76% in reading. This sawtooth pattern in the trend line can stem from ratcheting up the cut-point or from a shift in the rigor of the test (Linn, 2001).[6]

Caution is urged by Koretz (personal communication, January 22, 2006) and Sigman and Zilbert (personal communication, California State Department of Education, March 2006) when comparing year-to-year changes in student performance and when comparing state and NAEP results, especially if the data series begins at the low or high tail of the distribution of raw scores. When a proficiency cut-point is set near one tail of the distribution, the relative proportions of students who must move to exceed the cut-point can vary considerably. This is one reason that we report the full 14-year time series for most states, although care is required before making strong inferences within states regarding the post-NCLB period because only 4 years of observations are available.

Conclusion

These findings illuminate the challenges in answering the bottom-line question: Is NCLB working? One fact is crystal clear: We should not rely on state testing programs and the jagged trend lines that stem from their results. Instead, it is important to focus on historical patterns informed by the NAEP. Achievement gains, going back to the early 1970s, are most discernible at the fourth-grade level. Mean scale scores in reading—independent of the proficiency-bar debate—climbed by about one grade level between 1971 and 2004, with at least half of this bump coming between 1999 and 2002 (seen in the regular NAEP time series). Some policy mix, rooted in state-led accountability efforts, appears to have worked by the late 1990s. But growth flattened out in fourth grade over the 3 years after enactment of NCLB.

Progress in math achievement has been more buoyant, with fourth graders performing about two grade levels higher in 2004 on the NAEP compared with their counterparts back in 1973. About half of this gain occurred between 1999 and 2004, but the discrete effect of NCLB, beyond the momentum of state-led accountability reforms, is difficult to estimate. Remember that 2003–2004 was the second full school year in which schools lived under NCLB rules and sanctions, and growth in math was slower post-2003 than before enactment of NCLB.

When it comes to narrowing achievement gaps, the historical patterns are similar. For reading, ethnic gaps on the NAEP closed steadily from the early 1970s through 1992, then widened in 1994, and then narrowed through 2002. But no further narrowing has occurred since 2002. For math, the Black–White gap narrowed by over half a grade level between 1992 and 2003, but no further progress was observed in 2005. The Latino–White gap has continued to close, with a bit of progress post-NCLB—the one bright spot on the equity front.

The fact that student performance has generally reached a plateau raises the crucial question as to whether standards-based accountability is sufficient to advance more effective and equitable schools. The very slow rise in reading proficiency over the past 15 years remains worrisome as well, especially when compared with the more robust gains in mathematics, notwithstanding the slowing growth rate post-NCLB.

Recent analyses have sparked debate over whether the states can be trusted to devise reliable gauges of achievement, particularly in how they define proficient levels of achievement. Some reformers are calling for national examinations, presumably pegged to standards set by the NAEP governing board (Mathews, 2006). This article details how state results continue to exaggerate the percentage of fourth graders deemed proficient or above in reading and math when compared with NAEP results. For

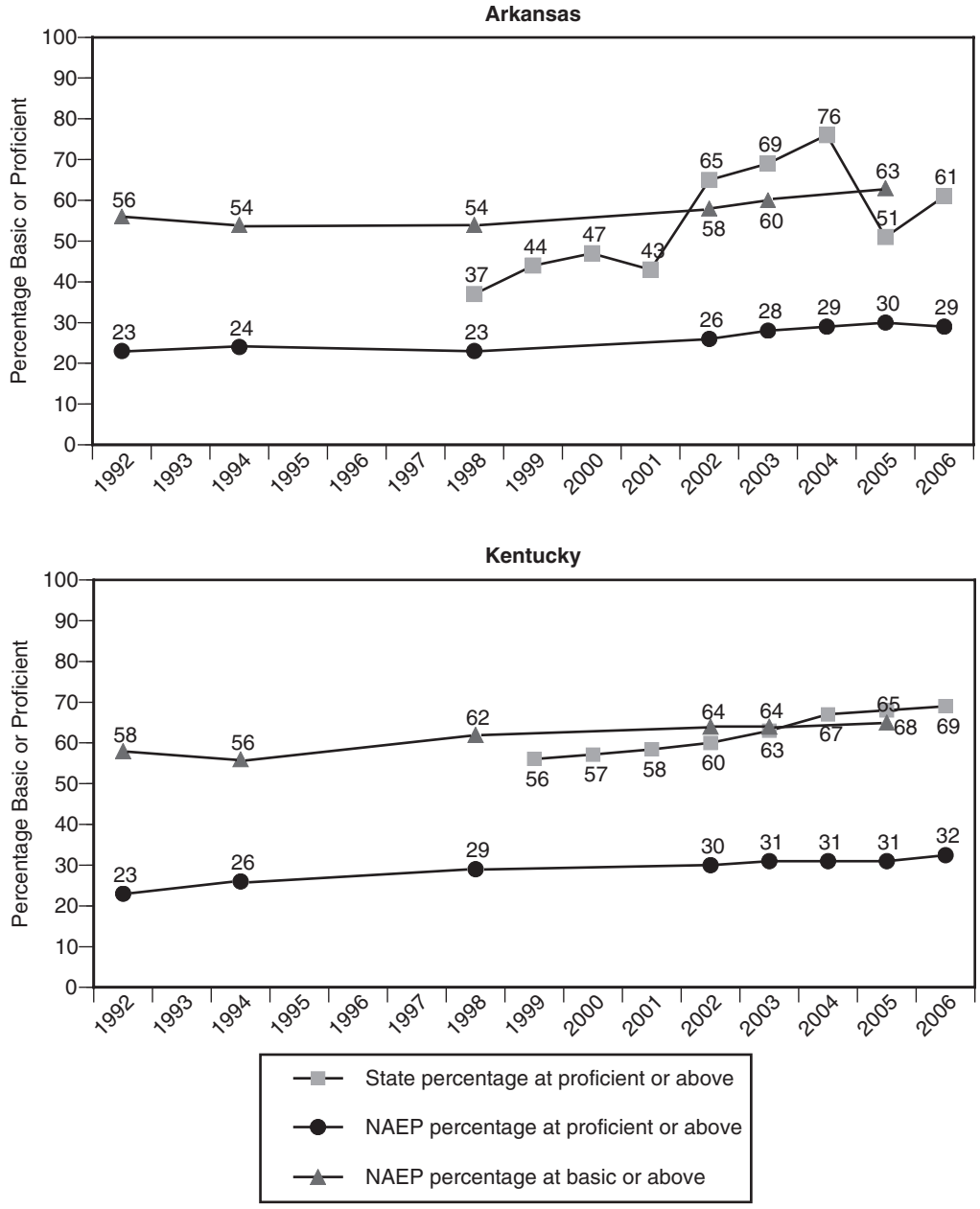

FIGURE 3. *Differing trends in percentage of fourth graders at basic or proficient and above in reading: state versus National Assessment of Educational Progress (NAEP) results. NAEP values for 2006 are simple linear projections.*

reading, this gulf between the dual-testing systems has actually grown wider over time.

Still, state policymakers will undoubtedly stand by their testing regimes given that their assessments are closely aligned to state curricular standards. Any serious move toward national examinations would run counter to the federal structure of public education. Furthermore, the notion that Washington might determine what every child in America should learn, grade by grade, remains highly controversial.

As Congress reviews and struggles to modify NCLB, how might states move toward more legitimate ways of gauging student performance? The fundamental principles of transparency and simplicity might guide state and congressional leaders. For example, even if Washington concludes that current NAEP cut-points for proficiency are too challenging—pegged too far above the states' own cut-points—the labeling of basic and proficient could at least become more consistent between federal and state assessments (Koretz, personal communication, January 22, 2006; Linn, 2000). Otherwise, the credibility of state testing regimes will continue to fade in the eyes of parents and educators.

Another issue pertains to how state education officials are creating tests that are differentially sensitive to student gains at the low end, as revealed earlier in Texas. States should not be discouraged from carefully gauging progress at the low end. But this should not lead to the inflation of estimated progress or regress for the wider spectrum of students. State and NAEP officials could also do more

to inform the public on how student demographics are changing, and achievement trends should be interpreted in this context (Sigman & Zilbert, personal communication, March 2006). It would be questionable to link the rising proportion of English learners or students of color to the anemic progress in reading scores over the past two decades. But even the interpretation of NAEP trends is constrained by our hazy understanding of how achievement is moving, net the prior effects of student and family characteristics.

Washington could first raise confidence in state testing programs—and wider acceptance of NAEP results by state officials—by advancing a consensus as to where the proficiency bar should be set. And Washington should boost the capacity of state education departments to conduct equating exercises to link old and new tests. It is understandable that states 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.

Important studies are emerging that clarify how discrete elements of NCLB may be raising achievement levels for various student subgroups, including the benefits of after-school tutoring, attention to low performers by teachers, and converting failing organizations to charter schools. Over time, we want to learn what specific policy threads, regulations, or bounded programs advanced by NCLB are proving effective for which students and under what conditions.

Still, for all the talk of results-oriented reform, many governors and state school chiefs cannot honestly tell parents whether their schools are getting better and which student subgroups are making progress over time. State officials should engage this fundamental discussion with more candor and with clear ideas for how to improve their assessment programs. At the same time, Washington officials and NCLB proponents should carefully interrogate their claims as to whether NCLB is working and the empirical basis of their pronouncements.

NOTES

We warmly thank Jack Jennings, Dan Koretz, Susanna Loeb, Deb Sigman, Brian Stecher, and Eric Zilbert for their thoughtful comments on earlier drafts. Our studies of accountability policies are generously supported by the Hewlett Foundation. Special thanks to Mike Smith and Kristi Kimball for their steady encouragement and technical feedback. A heartfelt thanks to Ann Bowers Noyce and Amy Gerstein for their support from the Noyce Foundation. Lively discussions with Kati Haycock continue to sharpen our analysis. Much appreciation is expressed to Patricia Gándara and Mike Kirst for their unflagging aid as we worked through the evidence and improved our line of analysis. Any errors of fact or interpretation are solely our responsibility.

[1]These patterns, drawing from the regular National Assessment of Educational Progress (NAEP) time series, appear in Perie, Grigg, and Dion (2005) and Perie, Grigg, and Donahue (2005; state-level trend data appear on http://nces.ed.gov/nationsreportcard/states/profile.asp.).

[2]When *Education Week* ("No Small Change," 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 its bar above the hurdle established by NAEP designers.

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

[4]How to sample states to obtain time-series data has proven controversial. When the pro–No Child Left Behind group Education Trust aimed to track elementary school test results over the prior 3 years, 2003 to 2005, it could locate comparable statewide scores for 32 states (Hall & Kennedy, 2006). The U.S. Department of Education's analysis (Paige, 2004, p. 1) of elementary school and middle school results drew from 25 states, looking back just 1 or 2 years.

[5]Complete time-series data for each state appear in a technical report (Fuller, Gesicki, Kang, & Wright, 2006).

[6]Looking among the 50 states, mean NAEP scores vary significantly, about one third of a school year between high- and low-performing states, controlling on state demographic features (Grissmer & Flanagan, 2001). In Texas, students did experience solid progress in math, based on NAEP results. These gains were substantial 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). Ironically, state test results in Texas accelerated at a more rapid pace under the earlier test, then slowed and followed an erratic pattern after the state switched to the Texas Assessment of Knowledge and Skills exam.

REFERENCES

Brown, W. (2000). Reporting NAEP achievement levels: An analysis of policy and external reviews. In M. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements* (pp. 11–40). Washington, DC: National Assessment Governing Board.

Bush, G. W. (2004, January 3). *President's radio address*. Washington, DC: The White House.

Bush, G. W. (2007). *President Bush congratulates presidential scholars, discusses No Child Left Behind reauthorization*. Retrieved June 28, 2007, from www.whitehouse.gov/news/releases/2007/06/20070625–7.html

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: Kentucky Legislative Research Commission.

Center for Education Policy. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington, DC: Author.

Cohen, M. (1990). Key issues confronting state policymakers. In R. Elmore (Ed.), *Restructuring schools: The next generation of educational reform* (pp. 251–288). San Francisco: Jossey-Bass.

Congressional Budget Office. (1986). *Trends in educational achievement*. Washington, DC: Author.

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

Elmore, R., Abelmann, C., & Fuhrman, S. (1996). The new accountability in state education reform: From process to performance. In H. Ladd (Ed.), *Holding schools accountable: Performance-based reform in education* (pp. 65–98). Washington, DC: Brookings.

Fuller, B., Gesicki, K., Kang, E., & Wright, J. (2006). *Is the No Child Left Behind Act working? The reliability of how states track achievement* (Working Paper No. 06-1). Berkeley: University of California and Stanford University, Policy Analysis for California Education.

Fuller, B., & Wright, J. (2007, April). *Diminishing returns? Gauging the achievement effects of centralized school accountability*. Paper presented at the American Educational Research Association, Chicago.

Goertz, M., & Duffy, M. (with Carlson Le Floch, K.). (2001). *Assessment and accountability systems in 50 states: 1999–2000* (No. RR-046). Philadelphia: Consortium for Policy Research in Education.

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

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

Hernandez, R. (2004, September 23). Bush carries his attack against Kerry to Pennsylvania. *New York Times*, p. 23.

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

Koretz, D. (in press). Alignment, high stakes, and the inflation of test scores. In J. Herman & E. Haertel (Eds.), *Uses and misuses of data in accountability testing* (Yearbook of the National Society of for the Study of Education).

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

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.

Lee, J. (2006). *Tracking achievement gaps and assessing the impact of NCLB on the gaps*. Cambridge, MA: Harvard Civil Rights Project.

Lee, J., & Wong, K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41,* 797–832.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29*(2)*,* 4–16.

Linn, R. (2001). *The design and evaluation of educational assessment and accountability systems* (UCLA Technical Report No. 539). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.

Linn, R. (2006, October-November). *You can't get there from here* (Research and Policy Notes). Washington, DC: American Educational Research Association.

Loveless, T. (2003). *How well are American students learning?* Washington, DC: Brookings Institution.

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.

Mathews, J. (2006, September 3). National school testing urged: Gaps between state and federal assessments fuel call for change. *Washington Post,* p. A1.

National Assessment of Educational Progress. (2006a). *National trends in reading/mathematics by average scale scores*. Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/ltt/results2004/nat-reading-scalescore.asp

National Assessment of Educational Progress. (2006b). *NAEP data explorer* (customized data run). Washington, DC: U.S. Department of Education. Retrieved from http://nces.ed.gov/nationsreportcard/nde.

National Assessment of Educational Progress. (2006c). *The nation's report card* (state test data time series). Retrieved from http://nces.ed.gov/nationsreportcard/states/profile.asp

National Center for Education Statistics. (2005). *Digest of educational statistics, 2005*. Washington, DC: Author.

No small change: Targeting money toward student performance. (2005, January 6). *Education Week* (Quality Counts suppl.), p. 84.

Paige, R. (2004, October 8). *Statement in regard to the PACE study*. Washington, DC: U.S. Department of Education, Office of Public Affairs.

Perie, M., Grigg, W., & Dion, G. (2005). *The nation's report card: Mathematics 2005* (NCES No. 2006–453). Washington, DC: National Center for Education Statistics.

Perie, M., Grigg, W., & Donahue, P. (2005). *The nation's report card: Reading 2005* (NCES No. 2006–451). Washington, DC: National Center for Education Statistics.

Peters, T., & Waterman, R. (1982). *In search of excellence.* New York: HarperCollins.

Rhoten, D., Carnoy, M., Chabrán, M., & Elmore, R. (2003). The conditions and characteristics of assessment and accountability. In M. Carnoy, R. Elmore, & L. Siskin (Eds.), *The new accountability: High schools and high-stakes testing* (pp. 13–54). New York: Routledge Falmer.

Skinner, R. (2005. January 6). State of the states. *Education Week,* pp. 77–80.

Smith, M. (2006, October 10). *Standards-based reform: The next generation* (manuscript). Menlo Park, CA: William and Flora Hewlett Foundation.

Smith, M., & O'Day, J. (1991). Systemic school reform. In S. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing: The 1990 yearbook of the politics of education association* (pp. 233–267). New York: Falmer.

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

Stetcher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. Hamilton, B. Stetcher, & S. Klein (Eds.), *Making sense of test-based accountability in education* (pp. 79–100). Santa Monica, CA: RAND.

Treisman, U. (2005, November 21). *Building instructional capacity in large urban districts*. Lecture presented at the University of California, Berkeley.

Wirtz, W., et al. (1977). *On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline*. New York: College Entrance Examination Board.

Zilbert, E. (2006, April). *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.

AUTHORS

BRUCE FULLER is professor of education and public policy at the University of California, Berkeley, Tolman Hall 3659, Berkeley, CA 94720; *b_fuller@berkeley.edu*. His new book is *Standardized Childhood*, published by Stanford University Press.

JOSEPH WRIGHT recently graduated from Berkeley's Graduate School of Public Policy. He works on education and housing policy from New York City; *josephcoton@yahoo.com*.

KATHRYN GESICKI served as a research assistant at the Berkeley–Stanford Center, Policy Analysis for California Education. She is currently pursuing her master's degree in public health at the University of California, Berkeley, with an emphasis in health policy and management; *kgesicki@gmail.com*.

ERIN KANG served as a research assistant at the Berkeley–Stanford Center, Policy Analysis for California Education, before leaving for a teaching post in South Korea. She currently works on early childhood programs with the First 5 California Children and Families Commission; *erinhkang@gmail.com.*

Manuscript received January 9, 2007
Revision received May 25, 2007
Accepted July 5, 2007