Educational Evaluation and Policy Analysis
Winter 2002, Vol. 24, No. 4, pp. 305–331

Does External Accountability Affect Student Outcomes? A Cross-State Analysis

Martin Carnoy
Susanna Loeb
Stanford University

We developed a zero-to-five index of the strength of accountability in 50 states based on the use of high-stakes testing to sanction and reward schools, and analyzed whether that index is related to student gains on the NAEP mathematics test in 1996–2000. The study also relates the index to changes in student retention in the 9th grade and to changes in high school completion rates over the same period. The results show that students in high-accountability states averaged significantly greater gains on the NAEP 8th-grade math test than students in states with little or no state measures to improve student performance. Furthermore, students in high-accountability states do not have significantly higher retention or lower high school completion rates.

Keywords: *capacity building, external accountability, high school survival rates, high stakes, inclusion/exclusion, retention rates, standards-based reforms, state assessment, student performance, test scores*

THE current wave of assessment-based school accountability reforms combines two traditions in American education—public accountability and student testing. The combination seems to be changing what schools do and how they do it. Strong accountability increases state control, even as other reforms, such as charter schools and choice plans, strive to decentralize educational decision-making. This article examines factors associated with stronger accountability and asks whether stronger statewide approaches to accountability achieve their declared goal of improving student outcomes.

Students' performance on tests is the main measure states use for gauging educational improvement. However, improvements on state tests may not be an accurate measure of educational gains, since schools may substitute for more durable student learning by using strategies to increase performance on the particular testing instrument. Because of this, it is important to use alternative measures to gauge the success of these policies. In this article, we use a variety of measures of average student performance at the state level, including National Association of Educational Progress (NAEP) math test scores, 9th-grade retention rates, and high school survival rates (the proportion of students who reach the 12th grade).

We are interested in five questions:

1. Do states adopting strong accountability measures have different characteristics from states without such measures? In particular, are there factors that influence which states adopt strong

Funding was provided by OERI under a grant to the Consortium for Policy Research in Education (CPRE). The authors would like to thank Paula Razquin and Tiffany Smith for their invaluable help in developing the enrollment data used in the article. We would also like to thank CPRE's Susan Fuhrman and Margaret Goertz, as well as the anonymous *EEPA* reviewers, for their useful comments, and Don McLaughlin of the American Institutes of Research, Palo Alto, CA, and Arnold Goldstein of the National Center of Educational Statistics in Washington, D.C., for their crucial help regarding the NAEP results. The analysis of the data and interpretations expressed in this article are wholly the authors and should not be attributed either to OERI or CPRE.

accountability that also independently affect later student outcomes? This analysis is important for assessing whether our subsequent assessment of the relationship between accountability and student performance may be driven by an underlying factor that influenced implementation, as well as for identifying controls to use in these analyses.

2. Do states with strong accountability systems see larger increases in national assessment test results?

3. Do states with strong accountability see increases in their retention rates in 9th grade relative to states with weak accountability? Once these systems are implemented, they are supposed to improve the likelihood that students finish high school. But according to some analysts, the pressure on schools to do well on high school minimum competency tests pushes administrators and teachers to increase retention rates in the first year of high school (Haney, 2000).

4. Do high school progression rates in states with strong accountability increase or decrease relative to progression rates in states with weak accountability? If accountability raises student performance in the earlier grades, this should ultimately raise the proportion completing high school; if, however, strong accountability increases retention rates in earlier grades, it could well reduce survival rates.

5. Does the relationship between accountability and student outcomes differ between racial/ethnic groups?

Standards-Based Reform and the Role of Testing

In the past, accountability and assessment were only loosely connected. Assessment was used mainly to divide students into academic tracks or for diagnostic purposes, helping school administrators and teachers see whether students were learning loosely defined state curricula. Accountability has traditionally been based in community participation and parent control, as represented by local school boards. Schools have been accountable to district administrators, who, in turn, answer to elected boards.[1] Parents have also been able to influence schools directly. School test results enter into parental decisions on where to live and fuel parent criticisms of school board actions, especially in higher income neighborhoods.[2] The link between traditional local accountability and traditional student assessment has long been important in neighborhoods with high parent participation. In the majority of schools, however, this link has either been indirect, acting through family residential choice, or practically nonexistent. It has been especially weak in low-income communities and large urban school districts.

The use of student testing at the state level with consequences for students and even as a measure of school performance is also not new. New York has used Regent examinations to test students' command of high school curriculum since the 19th century. The Iowa Test of Basic Skills (ITBS) has been given to 8th graders in Iowa since 1935. It was subsequently applied in many other states for students in many grades. However, the purpose of the ITBS was (and in Iowa, continues to be) diagnostic. How well students, classes, or schools performed on ITBS-type tests had few consequences. Schools stored the data, and school results were often published in local newspapers, but few administrators were compelled to take action because of declines in scores or continued low performance. High-stakes testing, such as the Regents, the Scholastic Aptitude Test (SAT), or Advanced Placement (AP) tests, were mainly related to college entrance and focused on individual student performance, not school performance. Even so, John Bishop has tried to develop an empirical case that such individual tests have had positive effects on overall student performance, since students understand the benefits of performing well on these tests and getting students to pass them is part of schools' objectives (Bishop, 1998).

Texas was a pioneer in using a state assessment test in order to measure school performance directly and both to sanction those schools not meeting improvement norms and to reward schools exceeding norms. Other states, such as South Carolina and North Carolina began implementing such systems a few years later.

Based on these new types of state accountability systems, educators developed the notion of *standards-based reform* in the late 1980s. They incorporated two main concepts: alignment and capacity building (Smith, O'Day, & Cohen 1990; O'Day & Smith, 1993). Alignment means that, in order to focus on improving outcomes, school systems need to set clear standards and align curriculum and accountability mechanisms

with those standards. For example, if the community feels that all students should learn to read at a certain level or better by 3rd grade, that level has to be clearly defined for parents and teachers and curriculum has to be designed to meet that standard. The assessment system should be aligned with the curriculum to measure what teachers are supposed to be teaching, and the accountability mechanisms need to provide incentives and sanctions aligned with success or failure in achieving the standard. Testing students on the standards-based curriculum is meant to measure how well the system is meeting standards.

The second concept is capacity building. Standards-based reform does *not* necessarily assume that alignment alone can improve education. Educational reform needs to improve the capacity of teachers and administrators to deliver better education. Coherent organization, built around aligned standards and assessment, can produce the increased capacity to deliver improved education. If teachers know what to teach, and the organization acts to support their efforts to achieve the defined standard, this increases the collective capacity of the school or the district to deliver education. Much of the policy literature on building capacity focuses on organizational change (Talbert, McLaughlin, & Rowan, 1993; Elmore, 1995; Cohen & Ball, 1999). But, capacity is not just embedded in organization.

Standards-based reform considers that an important component of improving education is increasing the "quality" of teachers and educational administrators. "Quality" teachers need to know the subject matter they are teaching and be effective pedagogues. Principals and superintendents need to know how to manage schools and school districts. Such outright skills also constitute capacity. For education in classrooms to improve, these skills also have to be raised (Darling-Hammond, 1997).

In the standards-based reform movement, testing is just one component of a much broader and deeper set of sustained changes that proponents claim are needed for educational improvement to occur. Testing can be used in several ways. It can be an indicator to tell administrators and teachers whether they are reaching the organization's goals and to provide information on which elements of the curriculum are reaching students and which are not. It can be used as a measure of success or failure in an incentive system. It can be used as a gauge to increase standards, to assess curricula, or to provide technical assistance. It can be used as a mechanism to allocate additional resources in order to improve outcomes for groups having difficulty reaching the standard.

Testing in the states that implemented accountability systems has been used in several of these ways. But, in contrast with the 1984 Texas reform, which addressed a number of different facets of standards-based reform—including raising teacher salaries from very low levels to attract better-prepared individuals into teaching—testing has been *the* central element of most recently implemented accountability systems. Between 1980 and 1996, per-pupil spending in Texas rose 60% adjusted for inflation compared to a 37% increase in spending per-pupil nationally. Increased resources have not been linked disproportionately to the implementation of accountability in most states. With the focus on measurable and easily understood results, test scores are rapidly becoming the end-all of state accountability reforms. This has generated considerable controversy regarding their effect on student outcomes.[3]

Which States are Likely to Choose Strong Accountability?

In making average student scores on state tests a main gauge of school performance, state school officials have shifted influence over teacher and principal behavior from local school boards and district offices to the statehouse. The new accountability reforms ratchet up the degree of central state power over schools and reduce local control over school policy. Before assessing the impact of accountability on students, it is important to ask what underlying conditions move politicians and educational policy makers to shift school accountability from local communities to state regulatory agencies. By understanding the causes of implementation of accountability, we can design more accurate models for estimating the relationship between accountability and student outcomes.

Analyzing why some states are more likely to legislate and attempt to implement strong, outcome based state accountability systems is complex because states may vary in their motivation.[4] A number of studies in the past 30 years have attempted to model the diffusion of public sector reforms (see, for example, Walker, 1969;

Berry and Berry, 1990). More recently, Minstrom (2000) has modeled and tested the consideration and adoption of school choice legislation by states in the period 1987–1993. He estimates that the main predictors of school choice adoption are the actions of political actors that have a consummate interest in pushing for greater choice and those, notably teachers' unions, positioned against choice. His model does not include student demographics.[5]

In our analysis, we do not attempt specifically to model the diffusion of standards-based reforms. Rather, our purpose is to develop a recursive model of the effect of strength of accountability on student performance, in which strength of accountability is itself a function of variables that may affect student performance. For our purposes, only political, demographic, and educational variables that could influence *both* the strength of accountability reforms *and* possibly student outcomes are relevant. "Educational entrepreneurs" no doubt also have had an effect on standards-based reforms.[6] However, because the entrepreneurs are unlikely to impact student outcomes independent of the reforms, we do not include these in our model. Other factors may impact both accountability strength and student outcomes. For example, if changing student test scores influence implementation then we might attribute to accountability outcome changes that are simply a continuation of a trend. Thus, it is important to assess the relationship between changes in test scores prior to accountability and the strength of the accountability system.

Demographic characteristics of states may also influence both accountability and student outcomes, as may resources available for schools. Further, states in which power is already more centralized may have less of a political leap to make in implementing strong, state-led, assessment-based accountability reforms. Financing and policy-making in southern states compared to the North and West tended to be somewhat more centralized 30 years ago than today.[7] Case studies of states such as Kentucky, Texas, and Vermont also suggest that litigation challenging the structural inequality of local educational financing—inequality strongly related to racial/ethnic or social class composition of school districts—was associated with reform tying assessment to state administered accountability (Rhoten et. al., 2003). Thus, we include population, racial/ethnic composition, the percent of school revenues coming from the state, and revenue levels, as well as student test scores, in our model of accountability strength. Although there is much to criticize in the way we specify our model, we are limited by a small sample size of only 50 states. Our estimates provide greater explanatory power than do Mintrom's (2000), albeit for a different reform.

Expected Outcomes of "Strong" Accountability

The expressed purpose of the new state accountability systems is to raise student achievement and, more generally, to improve the quality of schooling. By testing pupils, states hope to provide performance benchmarks for schools that would, in President Bush's words, "leave no child behind." School administrators and teachers, exposed to scrutiny by published test scores, are expected to improve educational delivery to avoid "failing" and to gain the rewards of high academic achievement. States with strong accountability systems reward schools that perform well on tests and send negative signals to those that do not. A clear measure of the effect of strong accountability systems is thus whether they have a positive effect on test scores.[8]

The NAEP tests students approximately every four years in mathematics and reading at the 4th and 8th grades. These tests, designed at the federal level and considered a reasonable assessment of student knowledge in these subjects, have been used by many analysts to gauge whether students are learning more or less over time. More recently (since 1990), NAEP scores have also been available by state, although not every state participates in the assessment. Some states participate some years, and not in others. Since the NAEP math test was given in 1996 and 2000, it provides a good measure of whether state accountability systems—many of which came into being in the mid-1990s—are having a significant effect on student learning outcomes.

The stated objective of higher standards (for example, requiring all students to pass 9th-grade algebra or biology to graduate from high school) and statewide assessment, including high school exit tests, is to increase schools' focus on how much students learn. This new focus may have unintended consequences, at least in the short and medium run. Raising the bar on student learn-

ing in high school may make it more difficult for students to pass courses, hence increasing student retention and decreasing graduation rates. Dee (2002), analyzing the introduction of minimum competence testing and course graduation requirements, found reductions in educational attainment, particularly for black students. Jacobs (2001) found similar results for low-achieving students in Chicago; and Lillard and DeCicca (2001) found that graduation requirements could increase dropout rates. In his analysis of Texas data, Haney (2000) found that the implicit retention rates in the 9th grade (the number of 9th grade students that remained classified as 9th graders for a second year) increased steadily after the early 1980s to the mid-1990s for all ethnic groups, but particularly for blacks and Hispanics. He associated this trend with the implementation of statewide assessment and particularly with the Texas Assessment of Academic Skills (TAAS) high school exit testing, first implemented in 1991. We reanalyzed Texas enrollment data and confirmed Haney's 9th-grade retention finding. However, we were not as convinced that increased retention in 9th grade could be associated with the TAAS high school exit test, given in the 10th grade. If there is a link between retention and state policies in Texas it is likely to date back to the implementation of accountability in a more general sense and not to the current policy (Carnoy, Loeb, & Smith, 2001).[9]

Evidence of higher retention rates due to the new focus on assessment is important because, at the individual level, retention is a strong predictor of dropping out. For example, Rumberger (1995) shows that retained students are four times more likely to drop out, even after controlling for a host of background and school measures. In Texas in the 1980s, increased 9th-grade retention rates were clearly associated with declining high school completion rates for all ethnic groups (as measured by the ratio of the number of students graduating to the number of students in the 8th grade four years earlier). Yet, graduation rates stopped declining in the early 1990s, just as the 10th grade TAAS exit test was implemented across the state; by the end of the 1990s, the graduation rate was rising. Thus, a few years after the implementation of the high school exit test, retention rates had leveled off and graduation rates had begun to climb (Carnoy, Loeb, & Smith, 2001).

There are several possible explanations for these Texas trends. One is that the exit test has been easy enough or graded easily enough, so that it has not affected the decision of students who would not have dropped out anyway. A second is that it has taken some years for the positive effects of the accountability system to be felt in Texas high schools, so that the initial impact of stronger accountability through assessment was to increase retention rates, hence increase dropouts, but by the mid-1990s, student performance had improved sufficiently to increase graduation rates. Finally, financial resources may have played a role. As we mentioned earlier, Texas increased real spending per student much more than the national average in 1980–1996. The courts also mandated more equal distribution of spending across districts, and Texas began to implement the court order in the early 1990s. Now that enough other states have implemented accountability systems, we can use cross-state analysis to assess the relationship between these reforms and both progression through high school and performance on an independent measure of achievement, NAEP.

Model

Our model is recursive. First, we estimate accountability implementation as a function of the average level of test scores in the state in the early 1990s, test-score gains in the early 1990s, the percent of Latinos and African Americans in the state, the state population, the percent of school revenues raised at the state level in 1963 and 1995, average per-pupil revenues in 1990, and the yearly change in revenues in the early 1990s.

$$A_i = \beta_0 + \beta_1 T_i + \beta_2 R_i + \beta_3 P_i + S_i\beta_4 + D_i\beta_5 + \varepsilon, \quad (1)$$

where

- A = strength of accountability in state *i* (scale of 0 to 5),
- T = average scale score of 4th grade students in state *i* on the 1992 mathematics NAEP (in alternative specifications, we use the percent basic in reading and the change in percent basic in reading from 1992–1994),
- R = the proportion of African American and Hispanic (public school) students in state *i*,
- P = the state population,

S = the proportion of schools' funds coming from the state rather than local sources in 1963 and 1995, and

D = dollars per pupil revenues in 1990 and the yearly percent change in revenue from 1990 to 1995.

Next, we look at student achievement as a function of accountability, testing whether the percent of 8th graders or 4th graders achieving at the basic skills level or better increased more between 1996 and 2000 in states with "strong" outcome-based accountability than in states with "weak" accountability. We control for the 1996 test score to test whether lower scoring states in 1996 had a significantly higher gain in the next four years independent of the accountability index. We include as controls those variables that were significant predictors of accountability strength in Equation 1. We also include population growth and growth in the percent of Black or Hispanic students.

$$G_i = \phi_0 + \phi_1 A_i + \phi_2 Pr_i + X_i\phi_3 + C_i\phi_4 + \varepsilon, \qquad (2)$$

where

G ≡ change the percent of 8th grade or 4th grade students in state *i* who demonstrated basic skills or better on the mathematics NAEP between 1996 and 2000,[10]

Pr ≡ the 1996 level of the outcome measure,

X ≡ controls from Equation 1, and

C ≡ the change in population and in the percent of Black or Hispanic students.

We ran a number of specification checks on this basic model. One specification focuses on scale scores, but we are more interested in the percent of students passing at different skill levels (basic skills or better and proficient or better) because that allows us to test whether stronger accountability just affects gains in basic skills or also in higher level skills. As described in more detail, we check the estimated coefficient for accountability for possible bias due to exclusion from the NAEP math test of students classified as special education or limited English proficient.

Using a specification similar to Equation 2, but including an additional control for 1996 test score, we test whether 9th-grade retention (the number of students in the 9th grade divided by the number of students in the 8th grade the year before) rose more in the late 1990s in states with strong accountability than in states with weak accountability. We also test whether 10th to 12th grade survival rates and 8th to 12th grade survival rates increased more in states with strong accountability than in states with weak accountability.

$$Rt_i \text{ or } Sr = \phi_0 + \phi_1 A_i + \phi_2 Pr_i + X_i\phi_3 + C_i\phi_4 + \phi_5 E + \varepsilon, \qquad (3)$$

where

Rt ≡ 9th-grade retention rate in state *i*, 1996–2001,

Sr ≡ high school survival rates in state *i*, and

E ≡ 8th grade percent demonstrating at least basic skills in NAEP math in 1996.

We ran a number of specification checks on this basic model as well.

Data

We used four sets of data. For test scores, we used the posted NAEP math results by state from *The Nation's Report Card* available at http://nces.ed.gov/nationsreportcard/states. We used a number of different test scores to estimate Equations 1, 2 and 3.[11] In Equation 1, if most states were making their decisions to implement the new accountability systems in the early to mid-1990s, the most relevant scores are on the 1992 and 1994 reading tests and the 1992 math test. In Equation 2, the outcome measure is the latest one available, the gain in score on the NAEP math test from 1996 to 2000. We estimate the effect of the accountability index and other variables on both the 8th-grade results and the 4th-grade results. In checking for possible bias from exclusion, we use an alternative set of gain scores provided by McLaughlin (2001) as the dependent variable in one set of alternative estimates. In another set of estimates of Equation 2, we use two alternative exclusion adjustments provided to us by the National Center of Educational Statistics (NCES) as control variables. In Equation 3, the most relevant measure of test score is the1996 8th-grade math results, since these directly precede in time the changing 9th-grade retention rates and student survival rates in high school.

For retention rate and survival rate student outcome measures, we used enrollment figures that we gathered from a number of different

sources including state department of education web pages and the National Center for Education Statistics. Using data on 8th, 9th, 10th, and 12th grade enrollment for the years 1992–93 through 2000–2001, we calculate (a) the ratio of students in the 9th grade in year *t* to the number of students in the 8th grade in year *t*–1, (b) the ratio of the number of students in the 12th grade in year *t* to the number in the 10th grade in year *t*–2 and (c) the ratio of the number of students in the 12th grade in year *t* to the number in the 8th grade in year *t*–4.[12] We were unable to obtain sufficient enrollment data for Idaho, North Dakota and Utah.[13]

For accountability strength, we use the database developed by the Consortium for Policy Research in Education (CPRE), available on the CPRE web site http://www.cpre.org/Publications/Publications_Accountability.htm. The database provides information on state testing and accountability policies as of 1999–2000 (the year we used as our benchmark). From the database, we constructed a scale of accountability levels 0 to 5, based on pre-2000 accountability conditions, with states such as Iowa and Nebraska, that do not have any state-level accountability requirements for schools or districts coded 0, and states with "maximum" state-level demands on schools and that require a high school competency exam for graduation, such as Texas, North Carolina, New Jersey, and Florida coded 5. Appendix A reports our index, state by state.

The 0–5 scale captures degrees of state external pressure on schools to improve student achievement according to state-defined performance criteria. States receiving a zero do not test students statewide or do not set any statewide standards for schools or districts. States that require state testing in the elementary and middle grades and the reporting of test results to the state but no school (or district) sanctions or rewards (no or weak external pressure) get a 1. Those states that test at the elementary and middle school levels and have moderate school or district accountability sanctions/rewards *or,* alternatively, a high school exit test (that sanctions students but pressures schools to improve student performance) get a 2. Those states that test at the lower and middle grades, have moderate accountability repercussions for schools and districts, *and* require an exit test in high school, get a 3. Those that test and place *strong* pressure on schools or districts to improve student achievement (threat of reconstitution, principal transfer, loss of students) but do not require a high school exit test get a 4. States receiving a 5 test students in primary and middle grades, strongly sanction and reward schools or districts based on improvement in student test scores, *and* require a high school minimum competency exit test for graduation.[14]

Results

Descriptions of the variables central to the analysis appear in Appendix B. We see that average math test scores rose in the period 1996 to 2000 for all three race/ethnic groups in both the 4th and 8th grades. This was not the case in all states, however (see maximum and minimum score changes). A much lower proportion of Blacks and Hispanics compared to Whites achieve at the basic skills level of proficiency or better and at the proficient level or better. In the worst performing state, Mississippi, only one percent of Blacks score at the proficient level or above. In the states in which Black students perform best, such as New York, about 8–10 percent score at the proficient level or above. In the states where White students perform best, such as Connecticut, more than 40% scored at the proficient level or better on the 2000 test.

We also see that the ratio of 9th graders in 2000–2001 to 8th graders in 1999–2000 averaged approximately 1.19 for Black students across the states. With no population growth, this would indicate an approximately 19% retention of students in the 9th grade. In 1996 the corresponding rate was 18%. There are large differences between Black and White students. In both 1996 and 2001 the retention ratio averaged only six percent for White students. The ratio of 12th grade enrollment in 2001 to 8th grade enrollment in 1997 captures students' progression through high school. We see that on average this ratio is 0.85 for Whites and 0.75 for Blacks. With no population growth this would indicate that 85% (75%) of 8th graders in 1997 progressed to 12th grade four years later.

States' Implementation of Strong Accountability

The results of our estimates of the strength of state accountability equation appear in Table 1. We hypothesize that states with lower student test scores earlier in the 1990s would be more likely to implement strong systems. This appears

TABLE 1
Strength of State Accountability System Related to Demography and Educational Performance

Independent variable	I	II	III	IV	V	VI
NAEP 1992 4th grade- math White scale score	–0.049 (1.00)	–0.178 ** (2.85)				
NAEP 1992 4th-grade math Black scale score	0.011 (0.21)	0.069 (1.62)				
NAEP 1992 4th-grade reading White percent at least basic			–0.053 (0.99)	–0.116 (1.52)		
NAEP 1992 4th-grade reading Black percent at least basic			–0.024 (0.94)	0.033 (1.40)		
Change reading 1992–1994 White					0.044 (0.46)	0.105 (1.38)
Change reading 1992–1994 Black					0.037 (1.36)	0.0028 (0.13)
Percent Black or Hispanic 1995		5.40* (2.10)		5.23* (2.38)		4.71* (2.20)
Population—July 1995		0.000114** (2.89)		0.000066~ (1.68)		0.000080* (2.04)
Percent state finance 1963		–1.999 (–0.80)		1.24 (0.60)		1.61 (0.77)
Percent state finance 1995		1.23 (0.39)		–0.77 (–0.31)		0.177 (0.08)
Average per-pupil revenue 1990		0.000597* (2.34)		0.00053* (2.09)		0.00027 (1.28)
Yearly percent revenue change 1990–1995		8.89 (1.03)		8.41 (1.12)		7.58 (1.00)
Constant	5.98 (1.78)	8.36 (1.69)	7.02 (2.05)	4.39 (0.82)	2.62 (7.84)	–1.59 (–0.85)
R^2	0.04	0.58	0.09	0.56	0.06	0.55
Sample size	31	31	36	36	36	36

Note. Using the alternative index as the dependent variable, the coefficient and *t*–statistic on the test score variables are –0.080 (–1.74) and 0.045 (0.92) in column I; –0.167 (–2.28) and 0.080 (1.60) in column II; –0.059 (–1.12) and –0.022 (0.88) in column III; –0.082 (0.92) and 0.020 (0.74) in column IV; 0.070 (0.73) and 0.039 (1.44) in column V; 0.107 (1.25) and 0.012 (0.48) in column VI. *t*-values in parentheses.
~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

to be the case, but only relative to White students test scores, particularly on the math exam. There is no relationship between early Black student test scores and the strength of accountability. The results for Whites does suggest that in our modeling of accountability effects, we should watch for regression to the mean, in which low-scoring students make the largest gains. Our specification with achievement changes as the outcome and a control for prior test scores is suited to this. More centrally, we find no relationship between prior test score gains and the implementation of accountability. If we remove states that implemented accountability prior to the test-score change—Texas, North Carolina, and South Carolina—the coefficients are even closer to zero. This finding suggests that any estimate of an impact of accountability on test scores is unlikely to be biased by a continuing trend in test-score growth.

We do find large effects of racial composition on accountability strength. As discussed earlier, states with large residentially segregated minority or low-income populations may have developed more central control or regulation over educational resources. The relationship between strong accountability and racial composition may be the result of a positive correlation between the

proportion of minority students in the state and the centralization of school policy, either for earlier historical reasons related to segregation or for more recent historical reasons related to litigation over the distribution of public resources for education. The point estimates on the percentage of minority students imply that states with 20 percentage points more Blacks and Hispanics have accountability systems that are approximately one index number stronger.

To test the centralization hypothesis, we included as an explanatory variable the proportion of public school funding revenues coming from the state in 1963–64 and in 1994–95. The data on the proportion of school revenue coming from the state in the early 1960s predates the *Serrano* decision (1971) that moved many states to more centralized funding formulas to achieve greater revenue equalization among school districts. Most southern states had more centralized school financing systems in 1963. Schools in an average southern state received 58% of their funding from the state, compared with 36% in states outside the South. However, school finance in some southern states, such as Virginia was relatively decentralized. And some Western states such as Utah and New Mexico, whose politics continue to be dominated by rural interests, and Hawaii, with a completely different political history, have even more centralized school finance systems than the South. In any case, when accountability strength in 2000 is regressed alone against the proportion of school revenues coming from the state in 1963, the relationship is statistically significant. But when percent minority is included as an independent variable, centralization of school finance seems to play no significant role in explaining the strength of the accountability system. Measures of financial centralization in the mid-1990s, when many of today's state accountability systems were being organized are much less significant in explaining the strength of the accountability system even in the bivariate model.

More populous states with correspondingly larger absolute numbers of disadvantaged minorities, larger school systems, and larger cities also appear to implement stronger accountability systems. These conditions could imply greater difficulty in implementing local school improvement reforms and greater pressure for state-level controls. States vary in size from Wyoming with 480,000 people to California with 33 million.

Finally, while yearly growth in revenues does not predict accountability implementation, the level of resources in 1990 does. Using the results reported in Table 1 we choose to include the percent Black or Hispanic students in 1995, the population in 1995, and per-pupil revenues in 1990 as controls in our assessment of accountability effects.

Do States with Stronger Accountability See Greater Test-Score Gains?

Table 2 shows that for the percent scoring at the basic skills level or better on the 8th grade math NAEP, the effect of a two-step move in the accountability index (from, say, 1 to 3) implies a considerable increase in gains in the percentage of those students who score at the this level. For White 8th graders, for example, a two-step move means 2.8 percentage points more gain in the proportion scoring at basic skill level or more. With a mean gain of 4.8 percentage points and a standard deviation of 3.6 in average state proportions scoring at or above basic skill levels, the increase in gain from raising the external pressure on schools by the state appears to be substantial. The note to Table 2 gives the results using an alternative index that decreases the strength for California and New York to 2 and increases Maryland and New Mexico to 5. The results using the two indexes are very similar. Figure 1a shows how gains of White 8th graders vary from state to state across accountability levels.

Gains for other racial/ethnic groups from greater emphasis on student outcomes and accountability are even greater, perhaps due to lower rates of basic proficiency in 1996. For African Americans, the potential gains on the 8th grade test from increased outcome-based accountability are approximately five percentage points for every two step increase in accountability, relative to an average gain of 5.7 and a standard deviation of 5.3. For a two-step increase in the accountability index, the gain for Hispanic 8th graders is almost nine percentage points. The mean of the gains is 6.1 percentage points, and the standard deviation of gains among states is 8.5 points, so a two-step increase again makes a large difference. Figures 1b and 1c show how the gains of Black and Hispanic 8th graders vary across states by level of state accountability.

Table 3 shows similar, though somewhat smaller estimates of the relationship between accountability strength and the percent of 4th graders

TABLE 2

Gain in Percent of Students at Basic Skills Level or Better, NAEP 8th-Grade Math, 1996–2000, as Function of 1996 Level and Accountability, Across States, by Race/Ethnicity

	White Gain		Black Gain[a]		Hispanic Gain	
Independent variable	I	II	I	II	I	II
Accountability index	1.134**	1.41*	1.77**	2.57**	3.17**	4.47**
	(2.86)	(2.48)	(3.01)	(3.70)	(3.16)	(3.36)
1996 8th-grade math	–0.088	–0.140	0.211	0.052	–0.017	–0.142
	(1.11)	(1.17)	(1.53)	(0.25)	(0.12)	(0.71)
Percent Black and Hispanic		–0.046		–14.53		–31.44~
		(0.01)		(1.52)		(1.96)
Population—July 1995		–0.00014		–0.00017		0.000068
		(–1.06)		(1.21)		(0.25)
Average per-pupil revenue 1990		0.00083		0.00091		0.00102
		(1.25)		(1.04)		(0.81)
Yearly growth in percent Black or Hispanic		2.39		–64.37		–54.49
		(0.07)		(1.16)		(0.74)
Yearly population growth		–75.09		161.19		15.66
		(–0.75)		(1.02)		(0.06)
Constant term	8.37	8.97	–4.20	–1.71	–0.685	4.54
	(1.43)	(1.27)	(1.06)	(0.32)	(0.10)	(0.43)
R^2	0.25	0.34	0.3626	0.5546	0.27	0.41
Sample size	37	34	25	25	33	30

Note. With the alternative index, the coefficient and *t*-statistic on the index in column II for Whites, Blacks, and Hispanics are 1.26 (2.45), 2.43 (3.52) and 4.84 (4.25) respectively.

[a] We omitted Nebraska from the models for Blacks because the Black scores were so low in 2000 and the number of test takers was low. With Nebraska included the coefficients on accountability for Blacks increase to 2.41 (3.88), 3.10 (4.29). *t*-values in parentheses. ~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

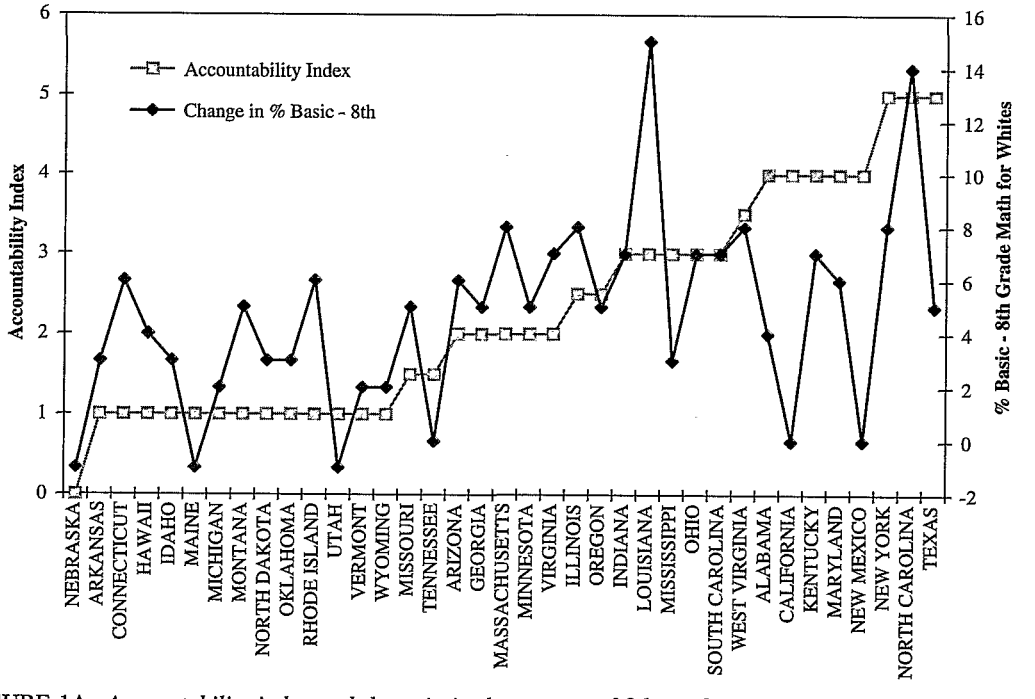

FIGURE 1A. *Accountability index and the gain in the percent of 8th graders reaching the basic level on the NAEP math exam from 1996–2000, Whites.*

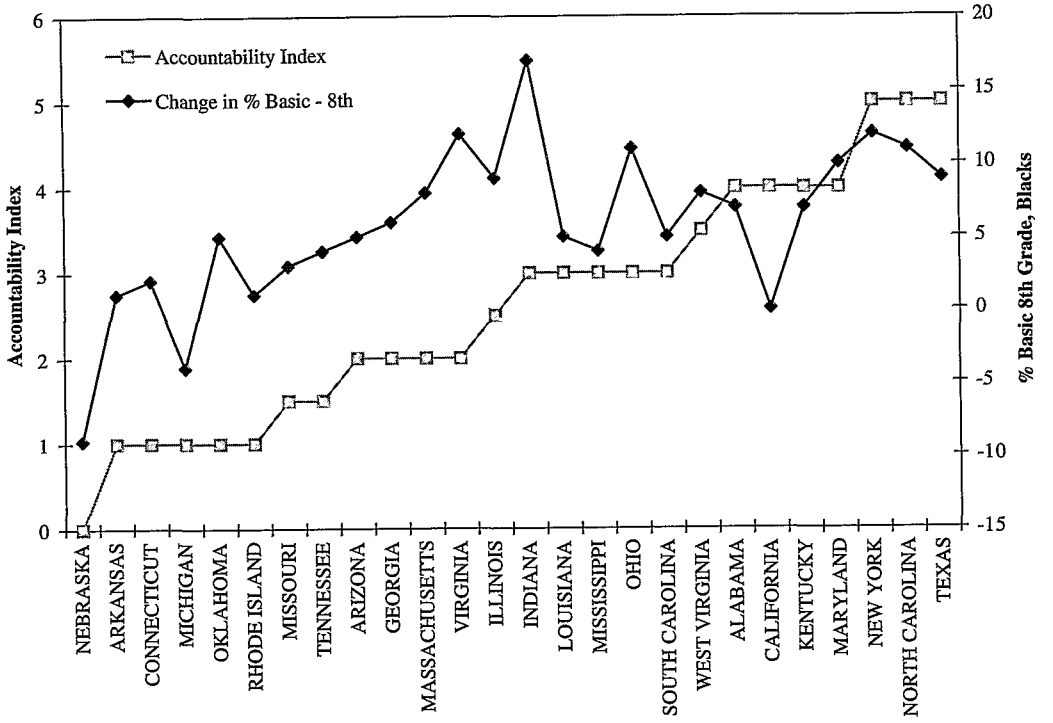

FIGURE 1B. *Accountability index and the gain in the percent of 8th graders reaching the basic level on the NAEP math exam from 1996–2000, Blacks.*

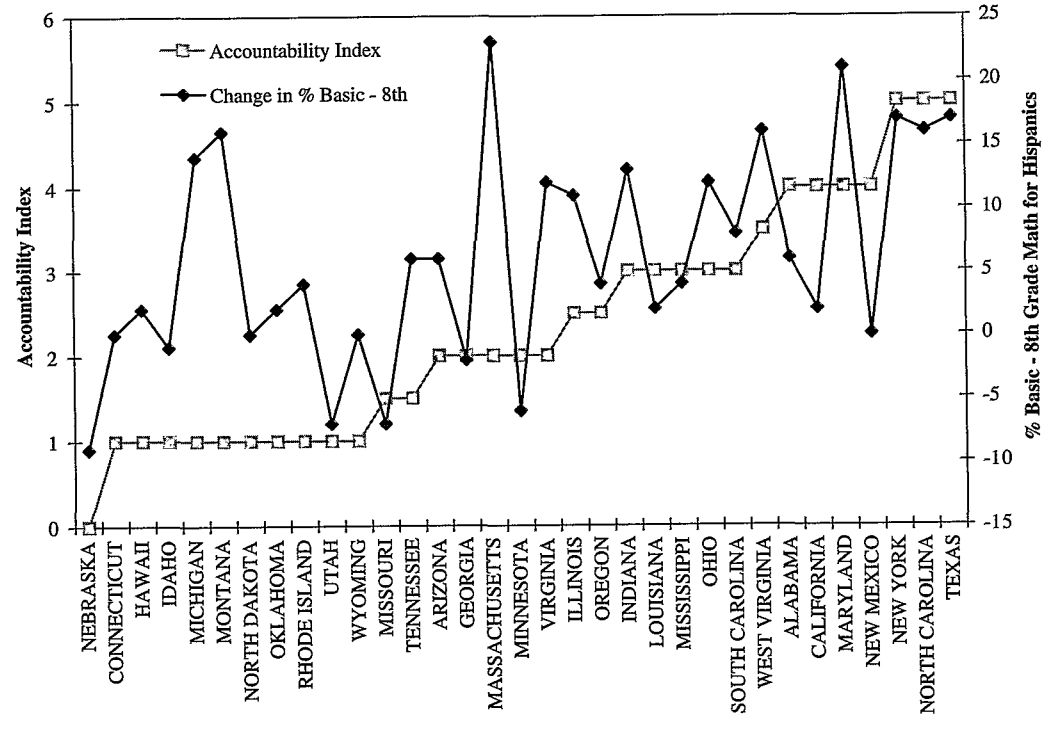

FIGURE 1C. *Accountability index and the gain in the percent of 8th graders reaching the basic level on the NAEP math exam from 1996–2000, Hispanics.*

TABLE 3
Gain in Percent of Students at Basic Skills Level or Better, NAEP 4th-Grade Math, 1996–2000, as Function of 1996 Level and Accountability, Across States, by Race/Ethnicity

	White Gain		Black Gain[a]		Hispanic Gain	
Independent variable	I	II	I	II	I	II
Accountability index	0.766~	.194	1.80*	2.54**	1.91**	1.70~
	(1.68)	(0.29)	(2.35)	(2.94)	(2.95)	(1.70)
1996 8th-grade math	−0.159	−0.268*	−0.091	−1.59	−0.270**	−0.337*
	(1.55)	(2.05)	(0.59)	(0.95)	(2.92)	(2.39)
Percent Black and Hispanic		6.03		−2.32		−10.70
		(0.94)		(0.24)		(0.81)
Population—July 1995		0.000050		−0.00016		0.00023
		(0.36)		(0.97)		(1.10)
Average per-pupil revenue 1990		0.0010		.0011		0.00043
		(1.48)		(1.11)		(0.47)
Yearly growth in percent Black or Hispanic		−8.21		61.23		−46.39
		(0.22)		(1.09)		(0.84)
Yearly population growth		−78.19		434.45*		25.19
		(0.69)		(2.38)		(0.15)
Constant term	14.81	18.28	6.53	−1.22	12.58	19.91
	(1.97)	(2.13)	(1.29)	(0.20)	(2.79)	(2.16)
R^2	0.14	0.28	0.20	0.48	0.41	0.43
Sample size	36	33	25	25	32	309

Note. With the alternative index, the coefficient and *t*-statistic on the index in Column II for Whites, Blacks, and Hispanics are .403 (0.67), 2.43 (3.52), and 2.01 (2.18) respectively.
[a]We omitted Nebraska from the models for Blacks because the Black scores were so low in 2000 and the number of test takers was low. With Nebraska included the coefficients on accountability for Blacks increase to 2.51 (3.16), 3.18 (3.15). *t*-values in parentheses. ~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

achieving at least the basic level. The coefficient of the accountability index is not significantly different from zero in the estimated equation for White 4th graders. Since Blacks and Hispanics start out at lower levels of basic skills proficiency than Whites, it may be easier to raise their low basic skills in the primary grades. This is partially borne out by our estimates. The estimated increase in the proportion of Hispanic 4th graders scoring at the basic skill level or higher corresponding to accountability strength is only marginally significant. The point estimate in the fully specified model suggests that a two-stem increase in accountability would increase the percent achieving basic by 3.4 percentage points (just over half of a standard deviation in the change score). For Black students the impact of accountability is significant and suggests a 5.1% increase in basic skills with a two-step increase in accountability.

The effect of strong accountability systems at higher skill proficiency levels on the NAEP test might be expected to be less, given the relatively "basic" nature of the test that most states use for accountability. We do, however, find a significant relationship between proficiency and the strength of the accountability system for all racial ethnic groups (Table 4). A two-step increase in the accountability index implies a 2.4 percentage point gain in the percent of White students and Black students scoring at proficient or better on the test, and 3.8 percentage point gain in the percent of Hispanic students scoring at proficient or better on the test.

Adjusting Results for Differential Exclusion and Inclusion Rates

A potentially serious bias in the NAEP math gains may arise because some students are eligible for exclusion from the test because they are designated as special education (SD) or limited English proficient (LEP). The proportion of SD plus LEP varies greatly among states. All states have some of these students take the standard NAEP test without accommodation and exclude others. A potential bias in gains arises because the

TABLE 4
Gain in Percent of Students at Proficient Level or Better, NAEP 8th-Grade Math, 1996–2000, as Function of 1996 Level and Accountability, Across States, by Race/Ethnicity

	White Gain		Black Gain		Hispanic Gain	
Independent variable	I	II	I	II	I	II
Accountability index	0.773*	1.23*	0.861**	1.21***	0.787	1.92**
	(2.14)	(2.39)	(3.41)	(6.10)	(1.50)	(2.71)
1996 8th-grade math	–0.0091	–0.093	–0.178	–.606**	–0.145	–0.285
	(0.12)	(–0.84)	(0.92)	(3.66)	(0.72)	(1.05)
Percent Black and Hispanic		–3.96		–10.56***		–19.95
		(0.80)		(4.40)		(2.51)
Population—July 1995		–0.00011		–6.35		0.000072
		(0.99)		(0.02)		(0.53)
Average per-pupil revenue 1990		0.00092		0.0006768**		0.00024
		(1.52)		(2.77)		(0.36)
Yearly growth in percent Black or Hispanic		–25.34		14.94		0.121
		(0.86)		(0.92)		(0.00)
Yearly population growth		25.50		62.69		75.50
		(0.28)		(1.45)		(0.66)
Constant term	2.47	0.79	–0.035	–0.34	1.88	3.05
	(1.08)	(0.25)	(0.03)	(0.24)	(0.82)	(0.64)
R^2	0.12	0.28	0.41	0.80	0.10	0.25
Sample size	37	34	26	26	33	29

Note. With the alternative index, the coefficient and *t*-statistic on the index in Column II for Whites, Blacks, and Hispanics are 1.23 (2.69), 0.966 (4.34), and 1.97 (2.91) respectively. *t*-values in parentheses.
$^{-}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

proportion of students designated SD plus LEP increased in most states administering the math NAEP in 1996–2000, and some states increased the percentage assessed whereas others decreased it. Some analyses have claimed that changing exclusion rates can account for a substantial proportion of NAEP math and reading gains in the late 1990s in states with strong accountability systems (Amrein & Berliner, 2002).

Since our analysis focuses on the relationship between the accountability index and math NAEP gains, it is important to adjust for the possible exclusion bias in the gains. There are several possible ways to do this given available data. Don McLaughlin of the American Institutes for Research has estimated an imputed set of 4th and 8th grade math NAEP scale scores for 1996 and 2000 by state, assuming that all excluded students had taken the test without accommodation (McLaughlin, 2001). He was kind enough to make a special calculation for us of scale scores by race/ethnic group and state. We use his imputed math scores to re-estimate the regression equations.

Since McLaughlin's estimates rely on imputed scores, they may over or under correct for changing exclusion rates. We made our own adjustments to the estimate regression equations that include control variables measuring the change in inclusion and assessment rates by state in 1996–2000. These rates are unpublished, but were provided to us by NCES (for the 8th grade percentages, see Appendix C and Appendix D). One variable we use as a control is the ratio of the percent of identified SD and LEP students who took the NAEP math test without accommodations in 2000 and in 1996 in each state. Some states included a higher percentage of identified SD and LEP students in 2000 than in 1996 and others, a lower percentage. If states with a stronger accountability system have had a propensity to exclude a greater proportion of their identified SD/LEP students from the NAEP test in 1996–2000, this adjustment should reduce the estimated coefficient of accountability in the test-score gain equation. However, the change in the percent included does not capture the fact that the percent identified might also have risen substantially between 1996 and 2000.

TABLE 5

Coefficients of State Accountability Index from Regression Estimates Using Unadjusted NAEP Test Score Gains, 1996–2000, and Various Adjustments for Students Excluded from Taking the NAEP Test

Group/grade and dependent variable	Coefficient of accountability index using unadjusted gains on NAEP	Coefficient using McLaughlin Scale scores as dependent variable	Coefficient using growth in percent included as control variable	Coefficient using absolute change in percent assessed as control variable
8th-grade, basic skills				
White	1.41*	n.a.	1.03	1.37*
Black	2.57**	n.a.	3.48***	3.65***
Hispanic	4.47***	n.a.	2.99*	4.06**
8th-grade, proficient				
White	1.23*	n.a.	0.949	1.27*
Black	1.21***	n.a.	1.28***	1.24***
Hispanic	1.92**	n.a.	1.66~	1.60*
8th-grade scale score				
White	1.01~	.957~	0.964	0.84
Black	2.21**	2.25**	2.71**	2.57**
Hispanic	2.12~	2.68*	2.36~	2.44*
4th-grade basic skills				
White	0.193	n.a.	−0.058	−0.19
Black	2.54**	n.a.	2.23**	3.42**
Hispanic	1.70~	n.a.	1.67	1.69
4th-grade scale scores				
White	−0.104	−0.372	−0.160	−0.27
Black	0.170	0.065	0.97	1.01
Hispanic	0.717	0.743	0.74	0.74

Note. Coefficient of accountability index reported for regression estimates that include the corresponding 1996 NAEP math score, percent Black and/or Hispanic in 1995, average yearly growth in percent Black and/or Hispanic, 1995 population, average yearly growth in population, and 1990 revenues per pupil.

~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$. n.a. = not applicable.

Thus, we estimate a second control variable: the absolute difference in percent of identified SD plus LEP students assessed between 1996 and 2000.

The results in Table 5 suggest that the positive relationship between test-score gains and the strength of a state's accountability system hold up across race/ethnic groups at the 8th grade even when adjusted for changing inclusion rates. The relationship of accountability to gains is generally not statistically significant for White or Hispanic students in the 4th grade, but is for Black 4th graders.

Do States with Stronger Accountability See Lower Retention Rates and/or Improved Progression Rates Through High School?

We do not find a strong relationship between the accountability system in a state in the late 1990s and changes in the retention rate. Table 6 gives these results. Accountability is correlated with Hispanic student retention in the univariate model, but this relationship goes away once prior retention is included. None of the other estimates of the accountability coefficient are even marginally significant and the point estimates are close to zero. We run three specification checks on this model. First, we adjust the 8th grade enrollment in t-1 by the growth in 8th-grade enrollment between t and t-1; thus, retention is measured by 9th-grade enrollment divided by 8th-grade enrollment in the same year. Second, we use the alternative coding of the accountability index; and third we omit 1996 test score in order to increase the sample size. The results are reported in the note to Table 6. We see no relationship between accountability and retention in any of the models for white or black students. However, when adjusted retention is used there is a strong positive relationship between accountability and retention for Hispanic students. One

TABLE 6
Ninth Grade Retention, 2000–2001, as Function of Retention in 1995–1996 and Accountability, Across States, by Race

	White Retention Ratio		Black Retention Ratio		Hispanic Retention Ratio	
Independent variable	I	II	I	II	I	II
Accountability index	0.0047	0.010	0.0067	0.0015	0.023~	–0.0086
	(0.83)	(1.30)	(0.60)	(0.10)	(1.75)	(–0.39)
Prior retention rate	0.850***	0.704**	0.424***	0.369~	0.499***	0.645***
	(4.15)	(2.87)	3.76	(2.00)	(5.77)	(4.96)
1996 8th-grade basic NAEP		0.0019		0.0075		0.0052
		(1.00)		(1.63)		(1.27)
Percent Black and Hispanic		–0.126		–0.108		0.432
		(1.62)		(0.50)		(1.65)
Population—July 1995		6.09		1.31		–2.40
		(0.33)		(0.39)		(0.52)
Average per-pupil revenue 1990		–1.26		1.98		1.921
		(1.24)		(0.39)		(0.88)
Yearly growth in percent Black and Hispanic		–1.20*		0.776		–1.73
		(2.47)		(0.95)		(1.43)
Yearly population growth		1.14		3.05		–2.90
		(0.78)		(0.78)		(0.69)
Constant	0.155	0.245	0.674	0.448	0.587	0.122
	(0.72)	(0.80)	(5.25)	(1.85)	(5.43)	(2.96)
R^2	0.34	0.54	0.29	0.59	0.50	0.74
Sample size	46	33	46	26	46	29

Note. When the adjusted retention rates are used, the coefficient on the accountability index in the six models are 0.0038 (0.45), 0.0107 (0.76), 0.0125 (1.05), 0.0018 (0.11), 0.063 (3.14), 0.069 (2.10). With the alternative accountability index the coefficients on accountability are 0.0049 (0.85), 0.0090 (1.23), 0.0027 (0.24), –0.0020 (0.14), 0.023 (1.73), –0.0055 (0.27). When 8th-grade test scores are omitted from Model II the sample sizes increase to 45 but the coefficients on accountability remain approximately the same: 0.0056 (0.81), 0.0074 (0.52) and 0.022 (1.14). *t*-values in parentheses.
~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

note of caution is that our measures of retention do not do a good job distinguishing among migration, demographic population bulges, and retention. This shortcoming may be particularly problematic for Hispanic students. Yet, we have not ruled out the possibility that strong accountability is associated with increased 9th-grade retention for Hispanic students.

Table 7 gives the results for progression from 10th to 12th grade. None of the estimated effects of accountability are significant in any of the specifications. All of the point estimates for White students are close to zero, but again the point estimates for Hispanic students (and some estimates for Black students) are large enough not to allow us to rule out a negative relationship between accountability and 10th to 12th grade progression. The final set of analyses combines 9th grade retention with 10th to 12th grade progression by looking at 8th to 12th grade progression as a function of accountability strength. The results, reported in Table 8, again show no significant impact of accountability on these survival rates. In all specifications the point estimates for White and Black students are close to zero; however, again, for Hispanic students our estimates are imprecise enough that we cannot rule out a negative relationship between accountability strength and progression through high school.

Conclusion

This analysis set out to assess changes in student outcomes associated with the implementation of state-level accountability systems in the 1990s. We start by developing a 0–5 scale of accountability strength based on testing requirements and repercussions of test results for schools and districts. We next explore the characteristics of states that have implemented accountability systems of different strengths in order to inform

TABLE 7
Tenth to Twelfth Grade Survival, 2000–2001, as Function of Survival in 1995–1996 and Accountability, Across States, by Race

	White Survival Rate		Black Survival Rate		Hispanic Survival Rate	
Independent variable	I	II	I	II	I	II
Accountability index	−0.0013	−0.011	−0.0068	−0.027	−0.039	−0.083
	(0.07)	(0.38)	(0.39)	(0.98)	(0.89)	(0.99)
Prior survival rate	0.717	1.03	0.556~	0.774	0.468	−0.312
	(1.44)	(0.87)	(1.91)	(1.42)	(1.41)	(0.36)
1996 8th-grade basic NAEP		0.0011		0.011		−0.0093
		(0.15)		(1.30)		(0.70)
Percent Black or Hispanic		0.194		0.262		−0.133
		(0.659)		(0.67)		(0.139)
Population—July 1995		2.56		1.15		1.14
		(0.39)		(0.18)		(0.623)
Per-pupil revenue 1990		2.95		2.33		2.05
		(0.08)		(0.06)		(0.02)
Yearly growth in percent Black or Hispanic		1.108		0.442		1.59
		(0.66)		(0.18)		(0.35)
Yearly population growth		10.68~		14.33*		43.11**
		(1.92)		(2.06)		(2.72)
Constant	0.258	−0.252	0.368	0.231	0.553	1.159
	(0.58)	(0.31)	(1.60)	(0.49)	(1.69)	(0.85)
R^2	0.07	0.24	0.10	0.55	0.08	0.36
Sample size	45	33	45	26	44	29

Note. When the adjusted survival rates are used, the coefficient on the accountability index in the six models are −0.0031 (0.19), −0.0025 (0.09), −0.010 (0.68), −0.018 (0.66), −0.013 (0.33), −0.040 (0.49). With the alternative accountability index the coefficients on accountability are −0.0026 (0.14), −0.017 (0.67), −0.0033 (0.20), −0.030 (1.13), −0.033 (0.76), −0.067 (0.82). When 8th-grade test scores are omitted from Model II the sample sizes increase to 43 but the coefficients on accountability remain approximately the same: −0.0064 (0.28), −0.022 (0.97) and −0.048 (0.79). *t*-values in parentheses.
~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

our analysis of the relationship between accountability and outcomes. Doing this reduces the potential that, in our assessment of the relationship between accountability and student outcomes, our index of accountability would simply be a proxy for an underlying factor that affects student performance. We find that states with a higher proportion of minority students and with larger populations are more likely to implement strong accountability. In addition, while we find no relationship between test-score gains prior to implementation and the strength of accountability, the results suggest that states with lower achieving White students are more likely to implement strong systems. Following these analyses, we estimate the relationship across states between accountability and both student test performance and student progression through high school. We ran a number of specification checks to test the robustness of our findings. Our results indicate a positive and significant relationship between the strength of states' accountability systems and math achievement gains at the 8th-grade level across racial/ethnic groups. Surprisingly, students' achievement at higher levels of math skills is also related significantly to stronger state accountability, suggesting that focusing on higher standards and how well schools do on tests may also improve higher level skills. This may result because schools with high-achieving students also feel the pressure to improve their students' performance. Indeed, there is some evidence that better performing schools have greater capacity to respond to external accountability pressures (Carnoy et al., in press). The 8th-grade achievement gains associated with stronger accountability are large. A two-step increase in the accountability scale corresponds to approximately a one half a standard deviation higher gain in the

TABLE 8
Eighth to Twelfth Grade Survival, 2000–2001, as Function of Survival in 1996–1997 and Accountability, Across States, by Race

	White Survival Rate		Black Survival Rate		Hispanic Survival Rate	
Independent variable	I	II	I	II	I	II
Accountability index	0.0069	0.0036	0.017	–0.0031	–0.040~	–.057
	(1.29)	(0.45)	(0.96)	(0.36)	(1.96)	(1.48)
Prior survival rate	0.917***	0.639**	0.825**	1.09*	0.146*	0.058
	(9.35)	(3.16)	(3.20)	(6.99)	(2.61)	(0.77)
1996 8th-grade basic NAEP		0.0036*		–0.0035		0.0048
		(2.07)		(1.39)		(1.01)
Percent Black or Hispanic		0.122		–0.251~		0.654
		(1.60)		(2.03)		(1.55)
Population—July 1995		5.62		5.07**		–2.84
		(0.31)		(2.98)		(0.46)
Per-pupil revenue 1990		–5.38		1.58		1.389
		(0.53)		(0.15)		(0.43)
Yearly growth in percent Black or Hispanic		1.025*		–1.47~		.396
		(2.33)		(1.94)		(0.23)
Yearly population growth		–1.31		1.30		2.07
		(0.93)		(0.64)		(0.35)
Constant	0.049	0.027	0.079	0.073	0.80	0.62
	(0.55)	(0.17)	(0.37)	(0.747)	(10.29)	(2.09)
R^2	0.71	0.73	0.21	0.83	0.20	0.26
Sample size	44	31	43	23	44	27

Note. When the adjusted survival rates are used, the coefficient on the accountability index in the six models are 0.0012 (0.20), .011 (1.04), 0.012 (1.38), .0056 (0.56), –0.0077 (0.32), –0.065 (1.18). With the alternative accountability index the coefficients on accountability are 0.0071 (1.28), .0051 (0.72), 0.019 (1.11), 0.0030 (0.36), –0.030 (1.44), –.045 (1.23). When 8th-grade test scores are omitted from Model II the sample sizes increase to 43 but the coefficients on accountability remain approximately the same: –0.0021 (0.31), –0.00043 (0.02) and –0.029 (0. 87). *t*-values in parentheses.
~$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

percent of students that achieve at least the basic level; and the effect sizes for gains at the proficiency level are even higher.

Another surprise is that 4th-grade test gains are generally not as strongly associated with accountability as 8th-grade gains. We do find that states with stronger accountability saw significantly greater gains in the percent of 4th-grade Black students that achieved at least the basic level on the math NAEP (more than a third of a standard deviation increase associated with a two-step increase in accountability); and marginally significant greater gains in the percent of 4th-grade Hispanic students that achieved at least the basic level on the math NAEP (approximately a quarter of a standard deviation increase associated with a two-step increase in accountability). We had expected the 4th-grade results to be stronger because many states have made greater gains on their own state test in the earlier grades and because elementary school teachers may have more flexibility to organize their time. Alternatively, 8th-grade students may have had more practice on tests and more exposure to accountability since many states start testing in 4th grade. Similarly, 8th-grade students may better understand the benefits of improved test performance and the ramifications of accountability for their academic success. Both the 4th-and 8th-grade results hold up to numerous specification checks, though despite positive effects on math achievement of stronger accountability, we observe considerable variation among states with similarly weak or strong accountability systems.

The long-term effects of stronger accountability are less clear. Our measures of progression through high school are not as reliable as we would like. Because they are based on state-by-grade enrollment in each year we cannot distinguish well among progression, migration and enrollment

changes due to demographic population bulges. We find no evidence of a relationship between accountability and 9th-grade retention, progression from 10th to 12th grade or progression from 8th to 12th for Black or White students. However we cannot rule out the possibility that accountability is associated with increased retention and decreased progression for Hispanic students. Of the many specifications, only a few show a significant relationship between accountability and these outcomes, but the point estimates are not accurate.

Certainly the results show no evidence of a positive affect of accountability on student progression through high school. Why might we find positive test score effects but not positive attainment effects? There are a number of possible explanations. First, while the NAEP results suggest that students in states with strong accountability programs are learning more than simply how to score well on their own state tests, these programs may be improving test taking skills but not changing factors that influence educational attainment and other outcomes of significance. An alternative explanation is that despite the positive effects of high-stakes accountability on math test scores, it may simply be too early to assess the long-term implications of this relatively new policy initiative on attainment outcomes. We may see attainment effects as the students who have spent more of their education under accountability systems move through high school. Our finding that states with stronger accountability systems have higher math gains on the 8th-grade NAEP may mean that students in those states will be more likely to do well in their 9th-grade courses and be more likely to graduate. On average, states with higher math scores do have lower 9th-grade retention rates, but this relationship is much weaker for African American students. African American students' average achievement may be sufficiently low in 8th grade that marginal increases in performance are not enough to improve high school course pass rates significantly, especially if standards for passing these courses are being raised. A third possibility is that outcomes for younger children are more easily influenced than those for high school students. Even though the current 4th and 8th grade students are performing better on the NAEP, other factors may affect them in high school and reverse the impact of accountability even on test performance. A final possible explanation is that higher scores on the NAEP math test may not measure "real" learning. A somewhat higher test score on the NAEP may not measure the learning that converts into better grades in math, English, and social studies courses in high school, enabling students to complete high school with their cohort. We cannot distinguish among these possibilities with this analysis.

In summary, this article provides evidence that states that implemented stronger accountability systems in the 1990s saw larger gains in student performance on the National Assessment of Education Progress mathematics exam between 1996 and 2000. These results are robust to numerous specification checks including controls for exclusion of students from the test due to classification into special education or limited English proficiency. A positive relationship is evident at both the basic level and at the proficient level of achievement and for both 8th grade and 4th grade, though the 8th grade results are stronger. This positive relationship is evident for Black, White and Hispanic students.

Notes

[1] The definition of community has changed over time, particularly in urban areas, and in many urban communities beginning a century ago, business interests became more influential in school policy than parent groups (Tyack, 1974).

[2] Evidence suggests that in choosing schools, parents are as conscious of the socio-economic background of students attending a school as they are of student test scores (Wells & Crain, 1992).

[3] Grissmer and Flanagan (1998) made headlines in 1998 when they released a study showing that Texas and North Carolina, two states that had implemented "strong" accountability systems early (Texas in the mid-1980s and North Carolina in the early 1990s), made much larger gains than other states in the math portion of the National Assessment of Educational Progress (NAEP) between 1992 and 1996. The study supported claims by the Texas Education Agency (TEA) that minority students made the largest gains, at least in primary school mathematics. These results suggested that state accountability systems could help lift academic achievement substantially and that low-performing students could be the primary beneficiaries of the new accountability reforms. Later studies were more critical though they did not contradict Grissmer and Flanagan. For example, Rand's Stephen Klein and his colleagues claimed that Texas' NAEP reading scores made only average increases. Their main argument was that the Texas assessment instrument (TAAS) scores rose over the 1994–1998 period, but were not

reflected in as great a gain in NAEP scores (Klein, Hamilton, McCaffrey, & Stecher, 2000).

[4] Strong accountability is represented by a system of state student testing with consequences for schools based on their students' test score improvement and consequences for students based on their passing high school exit exams (see Goertz & Duffy, 2001, and Table 1 in this article).

[5] Mintrom (2000) showed that the presence of "educational entrepreneurs" pushing school choice in a state was one of only a few significant variables explaining the adoption of school choice legislation, two others were opposition of teachers' unions to school choice and relatively low student test score gains in the preceding period.

[6] For example, Ross Perot in Texas and Rick Mills, the current Education Commissioner in New York and former Commissioner in Vermont, represent influential actors who successfully pushed for school accountability.

[7] Some analysts have claimed that southern states were more likely to centralize the financing and administration of public services in the 19th century because of white competition for resources with blacks in economies decimated by the Civil War. Horace Mann Bond (1939) argued, for example, that in Alabama lowland whites centralized control over scarce educational resources as a way to cut blacks out of public services and regain political power after Reconstruction. Desegregation in the South in the 1970s may have fortified state government influence over local schools.

[8] In cross-state and cross-national comparisons, Bishop et al. (2001) did not find a relationship between minimum competency exams and student test performance.

[9] Texas has tested students since the early 1980s. The Texas Assessment of Basic Skills was administered from 1980 through 1985 in grades 5 and 9 and from 1981 through 1985 in grade 3. Students were required to retake the 9th grade test in grades 10, 11 and 12 if they had not passed it. The Texas Educational Assessment Management System (TEAMS) was administered in 1986 through 1989 in grades 1, 3, 5, 7, and 9 (math, reading and writing) and in grades 11 and 12 (math and English language arts).

[10] Results using levels as the outcome, instead of changes, give similar results.

[11] For the 1992 NAEP reading exam, the following states do not have data: Alaska, Idaho, Illinois, Indiana, Kansas, Montana, Nevada, Ohio, Oklahoma, Oregon, South Dakota, Vermont, and Washington. For the 1994 NAEP reading exam, the same states are missing with the exception of Montana and Washington. For the 1992 NAEP mathematics exam Alaska, Idaho, Illinois, Kansas, Montana, Nevada, Oregon, South Dakota, Vermont, and Washington are missing for White students and the same states plus Hawaii, Iowa, Minnesota, New Mexico, Utah, and Wyoming are missing for Black students. Thirteen states; Alaska, Colorado, Delaware, Florida, Iowa, Kansas, Nevada, New Hampshire, New Jersey, Pennsylvania, South Dakota, Washington, and Wisconsin do not have either the 2000 NAEP math data or an earlier comparison year (1996, 1992, or 1990) for 8th graders. Of the 37 states with 2000 NAEP scores, all but Illinois, Ohio, and Oklahoma had 1996 scores. For those three states, we interpolated between 1992 and 2000 to get a 1996 score for Illinois, and between 1990 and 2000 to get a 1996 score for Ohio and Oklahoma. The situation was identical for the 4th-grade NAEP math exam except for Illinois, which had no 1992 score, so it was left out of the 4th-grade regression estimates. Of the 37 states, Arkansas, Kentucky, Maine, and Vermont were not in the estimates for Hispanic students. Of the 37 states, Hawaii, Idaho, Maine, Minnesota, Montana, New Mexico, North Dakota, Oregon, Utah, Vermont, and Wyoming are not in the estimates for Black students.

[12] Our outcome measures are correlated. Generally 4th-and 8th-grade scores are highly positively correlated (for basic proficiency, the correlations are 0.86, 0.70, and 0.75 for Whites, Blacks and Hispanics, respectively). Test scores are not strongly correlated with retention rates for Whites but are for Hispanics (–0.043, 0.18, and –0.37 with 8th grade basic for Whites, Blacks and Hispanics). These scores are positively correlated with 10th to 12th grade progression (0.31, 0.10, and 0.07). Ninth grade retention is consistently negatively correlated with 10th to 12th grade progression (–0.24, –0.23, and –0.25).

[13] The drawback of these measures is that they confound migration into the state with progression through high school. We create an alternative measure which adjust the base (i.e., the t-1 8th grade enrollment) by the percent increase in 8th grade enrollment during the progression time. The problem with this alternative measure is that it confounds demographic changes not due to migration with the progression rate. The results for the two measures are similar.

[14] Our estimates of state external pressure on local educational authorities are, by their nature, somewhat interpretive, especially in the "middle range" of 2s and 3s. Margaret Goertz, the co-author of the database, checked this index and, in all but a few cases, agreed with the values we assigned. She argued, for example, that in New York and California, the accountability system had not been in place long enough to count as strong accountability. We included her values as a check on the robustness of our results. The change made little change in the regression estimates.

Appendix A

TABLE A1
Accountability Index, by State, 1999–2000

State	Grades with state testing in 1999–2000	School accountability 1999–2000	Repercussion for schools 1999–2000	Strength of repercussion for schools 1999–2000	HS exit test in 2000	Grade HS test first given	First grad class	Index
Alabama	3–11	School report cards	Ratings, intervention	Strong	Yes	10	2001	4
Alaska	4–7	None	None	None	Yes	10	2002	1
Arizona	3,5,8,10	Report cards	'Public shame'	Weak	Yes	10	2002	2
Arkansas	4,6	None	None	None	No			1
California	2–11	Report cards	Ratings, awards, intervention	Strong	No	10	2004	4 (2)*
Colorado	3, literacy	None	None	None	No			1
Connecticut	4,6,8,10	Reporting scores to state	Identify schools with needs	Weak	No			1
Delaware	3,5,8,10,11	None	None	None	No	10	2004	1
Florida	4,5,8,10	Report cards	Ratings, subject to vouchers	Strong	Yes	10	1988	5
Georgia	3,4,5,8,11	School reports	None	None	Yes	11	1995	2
Hawaii	3,5,8,10	None	None	None	No			1
Idaho	ITBS, 3–8	None	None	None	No			1
Illinois	3,4,5,8,10	Academic improvement	Watch lists, warnings, intervention	Moderate	No			2.5
Indiana	3,6,8,10	Performance assessment	Accreditation	Moderate	Yes	10	1999	3
Iowa	None	None	None	None	No			0
Kansas	3,4,5,8,10	School reports	Accreditation	Weak	No			1
Kentucky	4,5,7,8,10–12	Meeting state improvement goals	Monetary rewards, intervention	Strong	No			4
Louisiana	LEAP,4,8	Report cards, growth targets	Intervention	Moderate	Yes	10	1991	3
Maine	4,8,11	None	None	None	No			1
Maryland	3,5,8	School performance index	Monetary rewards, reconstitution	Strong	Yes	10,11, 12	2001	4 (5)
Massachusetts	4,8,10	Students only	Student promotions	Implicit only	Yes	10	2003	2
Michigan	4,5,7,8	School rating	Accreditation	Weak	No			1
Minnesota	3,5,8,10	School reports	None	None	Yes	8,10		2
Mississippi	2–8	Only districts accountable, based on test scores	Public recognition, loss of accreditation	Moderate to strong at district level	Yes	11	1994	3

TABLE A1 *(Continued)*

State	Grades with state testing in 1999–2000	School accountability 1999–2000	Repercussion for schools 1999–2000	Strength of repercussion for schools 1999–2000	HS exit test in 2000	Grade HS test first given	First grad class	Index
Missouri	3–11	School can be deemed academically deficient	Possible audit	Weak	No			1.5
Montana	4,8,11	None	None	None	No			1
Nebraska	None	None	None	None	No			0
Nevada	4,8,10	School reports	None	Weak	Yes	11	1999	1.5
New Hampshire	3,6,10	None	None	None	No			1
New Jersey	4,5,11	Mostly district level, 75% pass rate	Audits, possible state takeover	Strong	Yes	11		5
New Mexico	1–9	School ratings and district rankings	Some money rewards, probation	Moderate to strong	Yes	10	1990	4 (5)
New York	4,5,8,11	State review of school performance	Freeze on pupil registration	Strong	Yes	10	1998	5 (2)
North Carolina	3–8	School ratings	Money rewards, intervention	Strong	Yes	9	1994	5
North Dakota	4,8,12	Improve student learning	Accreditation	Weak	No			1
Ohio	4,6,9,12	Report cards, but mainly district level	Money for schools, sanctions for districts	Moderate	Yes	9		3
Oklahoma	5,8	Reports to state	Accreditation	Weak	No			1
Oregon	3,5,8,10	School performance ratings	Write school improvement plans	Weak to moderate	Yes	10	1991	2.5
Pennsylvania	5,6,8,9,11	High schools have ratings	Money for HS improvement	Weak	No			1
Rhode Island	3,4,7,8,10	Yearly progress on test results	Reconstitution	Weak implementation	No			1
South Carolina	3–8,10	District only	District defined as impaired	Moderate	Yes	10	1990	3
South Dakota	2,4,5,8,9,11	Test reports	None	None	No			1
Tennessee	3–8,9	Test reports	Accreditation	Weak	Yes	9		1.5
Texas	3–8,10	Report cards	School ratings, interventions	Strong	Yes	10	1991	5
Utah	3,5,8,11	None	Accreditation	Weak	No	10	2007	1
Vermont	2,4,8,10	School reports	Identify schools for assistance	Weak	No			1

(continued on next page)

TABLE A1 *(Continued)*

State	Grades with state testing in 1999–2000	School accountability 1999–2000	Repercussion for schools 1999–2000	Strength of repercussion for schools 1999–2000	HS exit test in 2000	Grade HS test first given	First grad class	Index
Virginia	3,4,5,6,8,9	Report tests, other data	Standards of Accreditation	Weak to moderate	No			2
Washington	2–10	School reports	Accreditation	Weak	No	10	2008	1
West Virginia	3–8	Performance audits	Intervention	Strong	No			3.5
Wisconsin	3,4,8,10	Continuous progress indicator	Ratings of schools	Weak to moderate	No	11	2004	2
Wyoming	4,8,11	Only district	Accreditation	Weak	No		2001	1

Note. *Alternative specification of index, as per Margaret Goertz, in parentheses.

Appendix B

TABLE B1
Descriptive Statistics for Analysis Variables

Variable	Sample Size	*M*	*SD*	Min.	Max.
Accountability Index	50	2.12	1.44	0	5
NAEP Scores					
1996 4th grade % basic - Whites	36	71.92	6.25	63	86
1996 4th grade % basic - Blacks	27	30.93	6.63	18	47
1996 4th grade % basic - Hispanics	32	42.66	9.87	24	66
2000 4th grade % basic - Whites	36	77.00	6.48	66	89
2000 4th grade % basic - Blacks	27	38.67	9.04	21	60
2000 4th grade % basic - Hispanics	32	48.22	8.77	30	68
1996 8th grade % basic - Whites	37	70.49	6.95	56	80
1996 8th grade % basic - Blacks	26	27.50	6.13	16	40
1996 8th grade % basic - Hispanics	33	37.36	9.41	11	55
2000 8th grade % basic - Whites	37	75.24	6.95	59	86
2000 8th grade % basic - Blacks	26	33.23	8.22	18	48
2000 8th grade % basic - Hispanics	33	43.48	11.51	15	68
1996 8th grade % proficient - Whites	37	26.46	6.81	12	37
1996 8th grade % proficient - Blacks	26	3.81	1.86	1	8
1996 8th grade % proficient - Hispanics	33	8.18	3.60	2	19
2000 8th grade % proficient - Whites	37	30.46	7.27	14	44
2000 8th grade % proficient - Blacks	26	5.35	2.30	1	10
2000 8th grade % proficient - Hispanics	33	10.73	5.00	1	23
Change in NAEP Score 1996–2000					
4th grade % basic - Whites	36	5.08	3.98	–2	13
4th grade % basic - Blacks	27	7.74	6.60	–11	21
4th grade % basic - Hispanics	32	5.56	6.32	–13	19
4th grade scale score - Whites	35	3.16	2.96	–3.62	8.39
4th grade scale score - Blacks	35	4.48	5.00	–3.77	14.57
4th grade scale score - Hispanics	35	2.88	4.73	–7.33	15.41
8th grade % basic - Whites	37	4.76	3.62	–1	15
8th grade % basic - Blacks	26	5.73	5.42	–9	17
8th grade % basic - Hispanics	33	6.12	8.50	–9	23
8th grade % proficient - Whites	37	4.00	3.09	–2	13
8th grade % proficient - Blacks	26	1.54	2.12	–3	6
8th grade % proficient - Hispanics	33	2.55	4.12	–6	11
8th grade scale score - Whites	33	2.26	3.25	–3.36	11.17
8th grade scale score - Blacks	33	3.63	6.73	–9.61	23.05
8th Grade Scale Score - Hispanics	33	3.33	7.95	–13.98	21.73
Progression Rates					
9th grade 1996 / 8th grade 1995 - Whites	47	1.06	0.04	1.00	1.18
9th grade 1996 / 8th grade 1995 - Blacks	47	1.18	0.14	0.85	1.50
9th grade 1996 / 8th grade 1995 - Hispanics	47	1.26	0.21	0.88	2.35
9th grade 2001 / 8th grade 2000 - Whites	48	1.06	0.06	0.84	1.23
9th grade 2001 / 8th grade 2000 - Blacks	48	1.19	0.12	0.89	1.46
9th grade 2001 / 8th grade 2000 - Hispanics	48	1.26	0.17	0.83	1.95
12th grade 1996 / 10th grade 1994 - Whites	47	0.86	0.05	0.72	0.97
12th grade 1996 / 10th grade 1994 - Blacks	47	0.74	0.09	0.48	0.93
12th grade 1996 / 10th grade 1994 - Hispanics	47	0.87	0.18	0.44	1.43
12th grade 2001 / 10th grade 1999 - Whites	47	0.87	0.14	0.67	1.75
12th grade 2001 / 10th grade 1999 - Blacks	47	0.77	0.16	0.49	1.62
12th grade 2001 / 10th grade 1999 - Hispanics	46	0.88	0.40	0.46	3.16

(continued on next page)

TABLE B1 *(Continued)*

Variable	Sample Size	*M*	*SD*	Min.	Max.
12th grade 1996 / 8th grade 1992 - Whites	44	0.85	0.08	0.68	1.04
12th grade 1996 / 8th grade 1992 - Blacks	43	0.76	0.10	0.56	0.99
12th grade 1996 / 8th grade 1992 - Hispanics	44	1.05	0.51	0.57	3.00
12th grade 2001 / 8th grade 1997 - Whites	48	0.85	0.08	0.62	1.02
12th grade 2001 / 8th grade 1997 - Blacks	48	0.75	0.16	0.13	1.05
12th grade 2001 / 8th grade 1997 - Hispanics	48	0.87	0.20	0.55	1.70
Controls					
Percent Black or Hispanic Students - 1996 (8th grade)	50	0.20	0.15	0.00	0.51
Yearly growth in % Black or Hispanic 1996–2001	46	0.02	0.03	–0.03	0.11
Population 1995 (thousands)	50	5244	5759	480	31589
Yearly population growth 1995–2000	50	0.01	0.01	0.00011	0.045
Percent state funding - 1963	50	0.41	0.18	0.06	0.8
Percent state funding - 1995	50	0.53	0.16	0.07	0.97
Average per-pupil revenues 1990	50	4932	1300	3023	9249
Yearly growth in revenues 1990–1995	50	0.06	0.02	–0.0012	0.12

Appendix C

TABLE C1

NAEP 2000 8th-Grade Math Test: Proportion Students Identified as SD and LEP and Proportion Excluded and Included in the NAEP Test, by Race/Ethnic Group and State in Percent

	Whites, Non-Hispanic			African Americans			Hispanics		
State	Percent identified	Percent included	Percent assessed	Percent identified	Percent included	Percent assessed	Percent identified	Percent excluded	Percent assessed
AL	11.65	3.80	7.85	16.36	7.79	8.57	22.73	0.00	22.73
AZ	10.31	6.48	3.83	10.85	8.08	2.77	31.39	13.47	17.92
AR	10.89	6.84	4.05	15.59	11.31	4.29	38.29	12.59	25.70
CA	13.03	6.72	6.31	11.30	7.37	3.92	38.97	10.90	28.08
CT	13.26	8.25	5.02	17.56	14.72	2.84	25.87	15.77	10.10
GA	9.59	5.75	3.85	11.11	9.11	2.00	15.00	11.68	3.32
HI	20.11	5.05	15.06	16.61	9.68	6.93	26.00	3.00	23.00
ID	10.14	4.47	5.68	0.00	0.00	0.00	42.63	8.89	33.74
IL	13.01	6.42	6.58	15.06	12.05	3.01	23.21	9.17	14.04
IN	11.66	7.37	4.29	9.64	7.17	2.46	26.53	7.43	19.10
IA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KS	10.10	4.93	5.17	11.35	2.30	9.04	40.50	17.52	22.99
KY	12.50	9.21	3.30	16.90	12.61	4.28	0.00	0.00	0.00
LA	13.38	5.55	7.83	12.46	6.71	5.75	16.80	1.37	15.43
ME	14.43	8.72	5.71	0.00	0.00	0.00	0.00	0.00	0.00
MD	12.27	9.68	2.59	13.35	10.15	3.20	19.18	16.55	2.63
MA	16.11	9.83	6.28	20.32	6.26	14.06	36.70	26.60	10.10
MI	9.78	6.33	3.45	11.32	6.44	4.88	20.02	8.84	11.18
MN	11.87	4.54	7.33	0.00	0.00	0.00	44.50	18.89	25.62
MS	9.33	6.21	3.12	12.07	9.57	2.51	8.77	1.64	7.13
MO	13.84	8.14	5.70	17.93	12.98	4.95	22.43	5.52	16.90
MT	9.16	4.41	4.75	0.00	0.00	0.00	7.47	7.47	0.00
NE	10.89	2.30	8.59	20.20	16.73	3.47	28.87	6.32	22.55
NV	11.83	7.35	4.49	21.40	18.75	2.65	24.52	13.70	10.82
NM	15.44	6.40	9.04	0.00	0.00	0.00	27.97	15.54	12.43
NY	12.99	11.20	1.79	10.57	10.30	0.27	25.57	15.77	9.80
NC	13.14	11.38	1.76	18.74	17.13	1.61	29.53	23.72	5.80
ND	10.98	3.63	7.35	0.00	0.00	0.00	16.03	0.00	16.03
OH	11.75	9.00	2.75	13.34	11.05	2.30	5.73	0.00	5.73
OK	12.99	7.99	5.00	12.98	9.60	3.38	30.17	12.58	17.59
OR	13.35	4.05	9.30	19.78	14.36	5.42	37.17	18.80	18.37
RI	18.19	9.40	8.79	17.40	12.82	4.58	26.98	22.44	4.54
SC	11.75	4.93	6.82	14.86	10.43	4.43	12.26	0.00	12.26
TN	11.46	3.24	8.21	13.65	8.41	5.24	24.59	9.42	15.17
TX	13.75	7.34	6.41	16.85	7.94	8.91	28.31	12.24	16.07
UT	9.97	4.36	5.60	0.00	0.00	0.00	37.61	14.75	22.85
VT	17.07	9.90	7.17	0.00	0.00	0.00	0.00	0.00	0.00
VA	12.60	8.80	3.80	19.06	13.68	5.38	22.46	11.47	10.99
WV	14.43	11.18	3.25	20.05	15.29	4.76	11.44	2.39	9.05
WI	14.61	8.48	6.13	22.76	19.39	3.37	27.15	11.10	16.05
WY	11.19	3.97	7.21	0.00	0.00	0.00	25.74	4.38	21.35

Source. National Center of Educational Statistics.

Appendix D

TABLE D1

NAEP 1996 8th-Grade Math Test: Proportion Students Identified as SD and LEP and Proportion Excluded and Included in the NAEP Test, by Race/Ethnic Group and State in Percent

	Whites, Non-Hispanic			African Americans			Hispanics		
State	Percent identified	Percent included	Percent assessed	Percent identified	Percent included	Percent assessed	Percent identified	Percent excluded	Percent assessed
AL	12.88	6.63	6.26	13.19	8.38	4.80	13.93	1.71	12.22
AZ	9.09	4.62	4.47	20.68	19.21	1.47	26.43	13.90	12.53
AR	9.73	5.44	4.29	14.91	12.04	2.88	0.00	0.00	0.00
CA	9.37	5.08	4.28	22.99	13.41	9.58	30.82	15.01	15.81
CT	12.96	5.78	7.18	15.76	12.98	2.78	22.81	17.42	5.39
GA	9.34	5.81	3.53	9.15	7.09	2.06	20.69	11.96	8.73
HI	5.79	1.89	3.89	0.00	0.00	0.00	13.60	2.00	11.59
ID	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IL	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
IN	11.29	5.04	6.26	16.13	9.92	6.21	16.70	2.42	14.29
IA	12.26	4.78	7.48	18.08	14.45	3.63	12.23	8.62	3.61
KS	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
KY	8.64	4.41	4.23	11.94	5.80	6.14	0.00	0.00	0.00
LA	8.75	5.18	3.56	10.28	6.85	3.43	10.87	1.62	9.26
ME	11.57	4.96	6.61	0.00	0.00	0.00	0.00	0.00	0.00
MD	10.03	5.13	4.91	14.21	8.84	5.37	22.42	8.13	14.29
MA	14.39	5.76	8.63	25.82	14.90	10.93	29.05	22.20	6.86
MI	8.63	4.98	3.65	6.46	5.41	1.05	13.38	1.78	11.59
MN	10.33	3.15	7.18	14.46	2.40	12.05	24.11	3.53	20.58
MS	10.38	5.74	4.65	11.81	8.39	3.42	6.51	0.00	6.51
MO	10.71	6.62	4.08	14.01	8.54	5.47	16.17	9.45	6.71
MT	8.93	3.14	5.79	0.00	0.00	0.00	12.39	0.00	12.39
NE	10.49	3.14	7.35	17.18	13.33	3.85	27.88	12.57	15.31
NV	10.84	5.65	5.19	0.00	0.00	0.00	27.16	11.01	16.15
NM	11.88	3.81	8.07	0.00	0.00	0.00	21.07	10.34	10.74
NY	10.69	6.62	4.06	10.95	6.75	4.19	24.52	11.51	13.01
NC	7.62	3.42	4.20	10.58	5.44	5.14	29.43	13.16	16.27
ND	8.94	3.24	5.70	0.00	0.00	0.00	14.17	1.46	12.71
OH	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OK	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
OR	10.44	3.13	7.31	0.00	0.00	0.00	23.14	11.87	11.27
RI	14.04	5.84	8.19	20.44	6.14	14.30	27.22	11.52	15.69
SC	7.86	4.01	3.85	11.82	8.68	3.14	14.68	1.58	13.11
TN	11.20	4.15	7.05	9.13	5.74	3.39	25.76	2.43	23.32
TX	9.21	5.20	4.01	17.55	9.55	8.00	27.34	12.74	14.60
UT	9.17	4.66	4.51	0.00	0.00	0.00	25.45	16.72	8.73
VT	11.61	4.18	7.43	0.00	0.00	0.00	0.00	0.00	0.00
VA	12.32	5.96	6.36	15.28	11.46	3.82	14.74	6.45	8.29
WV	13.04	8.81	4.23	11.22	11.22	0.00	7.20	0.00	7.20
WI	10.19	5.66	4.54	17.99	16.60	1.39	17.94	5.31	12.63
WY	9.57	1.68	7.88	0.00	0.00	0.00	10.32	0.80	9.52

Source. National Center of Educational Statistics.

References

Amrein, A. & Berliner, D. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved February 10, 2003 from http://epaa.asu.edu/epaa/v10n18.

Berry, F. S., & Berry, W. D. (1990). State lottery adoptions as policy innovations: An event history analysis. *American Political Science Review 84,* 395–415.

Bishop, J. (1998). The effect of curriculum-based external exit systems on student achievement. *Journal of Economic Education, 29*(2), 171–82.

Bishop, J. F. M., Bishop, M., & Moriarity, J. (2001). The role of end-of-course exams and minimal competency exams in standards-based reforms. In

D. Ravitch (Ed.), *Brooking Papers in Educational Policy 2001,* (pp. 267–345). Washington, DC: Brookings Institution.

Bond, H. M. (1939). *Negro education in Alabama: A study in cotton and steel.* Washington, DC: The Associated Publishers Inc.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher state test scores in Texas make for better high school outcomes?* CPRE Research Report No. RR-047. Philadelphia, PA: Consortium for Policy Research in Education.

Carnoy, M. et al. (in press) *The new accountability: high schools and high stakes testing.* New York: Routledge.

Cohen, D. K., & Ball, D. L. (1999). *Instruction, capacity and improvement.* CPRE Research Report No. RR-043. Philadelphia, PA: Consortium for Policy Research in Education.

Darling-Hammond, L. (1997). The quality of teachers matters most. *The Journal of Staff Development 18*(1).

Dee, T. S. (2002). *Standards and student outcomes: Lessons from the "first wave" of education reforms.* Working paper PEPG 02-08. Swarthmore College.

Elmore, R. (1995). Teaching, learning, and school organization: Principles of practice and the regularities of schooling. *Educational Administration Quarterly, 31*(3), 355–374.

Goertz, M. E., & Duffy, M. C. (2001). Assessment and accountability systems in 50 states: 1999–2000. CPRE Research Report No. RR-046. Philadelphia, PA: Consortium for Policy Research in Education.

Grissmer, D., & Flanagan, A. (1998). *Exploring rapid achievement gains in North Carolina and Texas.* Washington, DC: National Education Goals Panel.

Haney, W. 2000. *Report for testimony in GI forum v. Texas Education Agency.* Boston, MA: Boston College, School of Education.

Jacobs, B. A. (2001). Getting tough? The impact of high school graduation exams. *EEPA* 23 (2): 99-122.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* RAND Issue Paper. Santa Monica, CA: RAND.

Lillard, D. R., & DeCicca, P. P. (2001). Higher standards, more dropouts? Evidence within and across time. *Economics of Education Review 20,* 459–473.

McLaughlin, D. (2001). *Exclusions and accommodations affect state NAEP gain statistics: Mathematics, 1996 to 2000.* Palo Alto, CA: American Institutes for Research.

Mintrom, M. (2000). *Policy entrepreneurs and school choice.* Washington, DC: Georgetown University Press.

National Center for Educational Statistics. (2001). *The nation's report card: State mathematics 2000 report.* Report No. 2001519. Washington, DC: Author.

O'Day, J. A., & Smith, M. S. (1993). School reform and equal opportunity: An introduction to the education symposium, *Stanford Law and Policy Review 4,* 15–20.

Rhoten, D., Carnoy, M., Elmore, R., & Chabran, M. (2003). The conditions and characteristics of assessment and accountability: The case of four states. In M. Carnoy et al. (Eds.), *High schools and the new accountability: A schools-eye view of standards-based reforms.* Manuscript submitted for publication. New York: Routledge.

Rumberger, R. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*(3), 583–625.

Serrano v. Priest, California State Supreme Court (1971).

Smith, M. S., O'Day, J. A., & Cohen, D. K. (1990). National curriculum, American style: What might it look like? *American Educator 14*(4), 10–17, 40–43.

Talbert, J. E., McLaughlin, M., & Rowan, B. (1993). Understanding context effects on secondary school teaching. *Teachers College Record, 95*(1), 45–68.

Tyack, D. (1974). *The one best system: A history of American urban education.* Cambridge, MA: Harvard University Press.

Walker, L. 1969. The diffusion of innovations among the American states. *American Political Science Review 63,* 880–899.

Wells, A. S., & Crain, R. L. (1992). Do parents choose school quality or school status? A sociological theory of free market education. In P. W. Cookson (Ed.), *The choice controversy* (pp. 65–81). Newbury Park, CA: Corwin Press.

Authors

MARTIN CARNOY is Professor of Education, School of Education, Stanford University, Stanford, CA 94305; carnoy@stanford.edu. His area of specialization is the economics of education.

SUSANNA LOEB is Assistant Professor, School of Education, Stanford University, 224 CEBS 520 Galvez Mall, Stanford, CA 94305; sloeb@stanford.edu. Her area of specialization is the economics of education.

Manuscript Received June 3, 2002
Revision Received January 2, 2003
Accepted January 23, 2003