Does School Accountability Lead to Improved Student Performance?

Eric A. Hanushek
Margaret E. Raymond

Abstract

The leading school reform policy in the United States revolves around strong accountability of schools with consequences for performance. The federal government's involvement through the No Child Left Behind Act of 2001 reinforces the prior movement of many states toward policies based on measured student achievement. Analysis of state achievement growth as measured by the National Assessment of Educational progress shows that accountability systems introduced during the 1990s had a clear positive impact on student achievement. This single policy instrument did not, however, also lead to any narrowing in the Black-White achievement gap (though it did narrow the Hispanic-White achievement gap). Moreover, the Black-White gap appears to have been adversely impacted over the decade by increasing minority concentrations in the schools. An additional issue surrounding stronger accountability has been a concern about unintended outcomes related to such things as higher exclusion rates from testing, increased dropout rates, and the like. Our analysis of special education placement rates, a frequently identified area of concern, does not show any responsiveness to the introduction of accountability systems.© 2005 by the Association for Public Policy Analysis and Management

The cornerstone of current federal educational policy has been expansion of school accountability based on measured student test performance. Although many states had already installed accountability systems by 2000, a central campaign theme of George W. Bush was to expand this to all states, something that became a reality with the No Child Left Behind Act of 2001 (NCLB). The policy has been controversial for a variety of reasons, leading to assertions that it has distorted schools in undesirable ways, that it has led to gaming and unintended outcomes, and that it has not and will not accomplish its objectives of improving student achievement. This paper provides evidence on the expected effects of NCLB not only on student performance but also on other potential outcomes.

The landmark NCLB codified a developing policy view that standards, testing, and accountability were the path to improved performance. Much of earlier educational policy, both at the federal and state level, concentrated on providing greater

Manuscript received May 2004; review complete August 2004; revision complete September 2004; accepted September 2004.

Journal of Policy Analysis and Management, Vol. 24, No. 2, 297–327 (2005)

Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/pam.20091

resources—especially for the education of disadvantaged students. But student outcomes proved noticeably impervious to these policy initiatives. As a result, federal policy made a distinct shift in focus to emphasizing performance objectives and outcomes rather than school inputs.[1]

It is nonetheless not possible to investigate the impact of NCLB directly. First, and most importantly, the majority of states had already instituted some sort of accountability system by the time the federal law took effect. Although only 12 states had accountability systems at the school level in 1996, 39 states did so by 2000. Thus, there is no ready comparison group that can indicate what might have happened without any law. Second, the law has many facets, making it hard to isolate the effects of any single one. Finally, the common pace of NCLB implementation across the states eliminates any status quo alternatives for comparison.

Isolating the impact of state accountability policies is inherently difficult. Because accountability invariably applies to entire states at an instant in time, the variation across schools within a state cannot be employed to identify the impacts of accountability; it is necessary to rely on state-level variation in student outcomes. Yet, states differ not only in their accountability policies but also in a variety other ways involving both population characteristics and other school policies. If these are not accounted for, they are likely to contaminate the estimates of the states' accountability systems.

Our approach uses information about state differences in mathematics and reading performance as identified by the National Assessment of Educational Progress (NAEP). We pursue a number of strategies designed to isolate the effects of school accountability on performance. First, we look at growth in performance between fourth and eighth grades to eliminate fixed differences in circumstances and policies of each state. Second, we include explicit measures for major categories of time varying inputs: parental education, school spending, and racial exposure in the schools. Third, we estimate the growth models with state fixed effects to eliminate any other policies that lead to trends up or down in student performance in each state. Finally, to identify differences by race or ethnicity, we disaggregate the state results for Whites, Blacks, and Hispanics.

We find that the introduction of accountability systems into a state tends to lead to larger achievement growth than would have occurred without accountability. The analysis, however, indicates that just reporting results has minimal impact on student performance and that the force of accountability comes from attaching consequences such as monetary awards or takeover threats to school performance. This finding supports the contested provisions of NCLB that impose sanctions on failing schools.

Much of the explicit interest in accountability and the federal legislation, however, focuses on low achievers. And, given the generally lower achievement by minority groups, an implicit assumption is that accountability—as revealed through mandatory disaggregation of performance for racial and ethnic groups—will simultaneously close the large achievement racial/ethnic gaps along with improving all performance. When we look specifically at the performance of subgroups, we find that Hispanic students gain most from accountability while Blacks gain least.

Since the widespread introduction of accountability, a parallel interest has been whether more rigorous and consequential accountability also leads to other,

[1] This switch to concentration on outcomes is often labeled the "standards movement." See Smith and O'Day (1990) for an early discussion of the precepts.

less desirable impacts. For example, does accountability lead to increased cheating, more classifications of students as "special education," or undesirable narrowing of teaching? To address a subset of these issues, we analyze the rate of placement into special education across states but find no evidence of reaction in this dimension.

RELEVANT STRANDS OF LITERATURE

Any consideration of state accountability systems must recognize the multitude of potential influences on student outcomes. The scientific challenge lies in separating the influence of accountability from these other factors.

The vast production function literature on variations in student performance provides a general backdrop for the analysis of achievement. This literature, dating from the Coleman Report (Coleman et al., 1966) and still being developed today, suggests significant differences in student achievement based on both family background and on schools (Hanushek, 2002).[2] A variety of controversies exists, particularly about the impact of various school resources (see Hanushek, 2003a), but without going into detail about these it is sufficient to conclude that there is a lack of consensus that any specific measures of school characteristics adequately capture the relevant factors determining student performance. Similar ambiguities exist when considering the measurement of family influences, even if there is strong consensus that families are very important in determining achievement. This lack of consensus on the appropriate specification of the determinants of student achievement motivates the analytical approach described below.

Throughout the study of schools and achievement, considerable attention has gone to the distribution of outcomes, and especially racial aspects of schooling. As famously highlighted more than 50 years ago by *Brown v. Board of Education,* the racial composition of schools may be relevant to achievement. The Coleman Report itself was legislatively mandated in the Civil Rights Act of 1964 and spawned attention to the racial composition of schools (U.S. Commission on Civil Rights, 1967). Although most of the subsequent analysis flowing from *Brown* has related directly to the desegregation of schools (for example, Armor, 1995; Rossell, Armor, & Walberg, 2002), recent attention has turned more to issues related to the composition of schools.

Separating the effects of the racial composition of schools from other factors is clearly difficult, in large part because measurement errors for other school and family factors are likely to be correlated with racial composition. The analysis of Hanushek, Kain, and Rivkin (2002) approaches the problem through a generalized peer analysis that controls for family, school, and neighborhood effects through exploiting the rich longitudinal data from stacked panel data on student performance in Texas. That analysis suggests that an increased Black concentration in schools has a detrimental effect on Black achievement, although racial composition does not seem to affect either Whites or Hispanics. This consideration is particularly important given recent concern that racial concentration in the schools has been rising. Partly because court supervision over school racial patterns is ending but more importantly because White attendance in large urban systems has decreased, minority concentration has grown throughout the 1990s (Orfield & Eaton, 1996; Clotfelter,

[2] Much of this literature is reviewed elsewhere. Here we merely identify sources both of basic analysis and of extended bibliographies on the relevant issues.

2004).[3] Thus, racial composition of schools may interact with efforts to improve schools in ways that policy designers have not yet understood.

Each of these influences is embedded within school systems across the states that are pursuing a variety of policy reforms. The difficulty is that these other reforms are neither well specified nor readily measured, leading to considerable difficulty in adequately differentiating the relevant components (Hanushek, 2002). Moreover, as we look forward to an analysis of state level data, we know the potential damage of missing key ingredients to performance is amplified with aggregate data (Hanushek, Rivkin, & Taylor, 1996).

The final strand of relevant literature pertains to accountability itself. Although a recent policy effort, policies related to accountability have already become quite controversial—rising to the level of front page stories in the *New York Times* (Winter, 2002). Much of the work is very new and has not appeared in journals yet. The available studies generally support the view that accountability has had a positive effect on student outcomes, although the limited observations introduce some uncertainty (Carnoy & Loeb, 2002; Hanushek & Raymond, 2003b; Jacob, 2003; Peterson & West, 2003).[4] The existing analyses of accountability and state differences in performance (Carnoy & Loeb, 2002; Hanushek & Raymond, 2003b), which are closely related to our analysis here, rely on more limited NAEP samples (with both stopping in 2000). The available data constrain the possible analyses, leading to serious questions about the strategies employed to isolate separate effects. The analysis of Carnoy and Loeb (2002) attempts to identify selection effects in the introduction of accountability through a series of separate cross-sectional regressions for each ethnic group. It identifies accountability effects by relating an index of different components of accountability systems to changes of math scores of eighth-graders (or fourth-graders) between 1996 and 2000, but the small sample sizes (25–37 states) limit the ability to control for other possible influences on achievement. Hanushek and Raymond (2003b) evaluate overall state math performance but consider growth in scores between fourth and eighth grade for specific cohorts. They employ larger samples by combining data from the different testing periods of the 1990s (see below) and introduce information about how long accountability had been in place. Nonetheless, both analyses are subject to bias from other omitted state changes or state educational policies and stand in contrast to the work here that identifies effects from the changes within states that occur from the introduction of accountability. Our extended analysis reported below expands the sample with newly available testing data, introduces state fixed effects to deal with unmeasured inputs and policies, and follows achievement over time for Whites, Blacks, and Hispanics separately. These innovations permit much clearer identification of accountability impacts along with providing details about impacts on the different ethnic groups.

A larger body of work has concentrated on whether or not accountability has produced gaming and subsequent unintended outcomes. This available work, reviewed in Hanushek and Raymond (2003b), tends to suggest some immediate reactions to accountability in terms of focusing teaching on relevant subjects or even relevant students near performance cutoffs; of increased exclusions from tests; of explicit cheating on tests; and of like attempts to improve scores in ways other than improv-

[3] The increased racial concentrations in schools also occurs at a time when residential segregation has generally declined; see Iceland and Weinberg (2002).

[4] Some variation also comes from analytical methods; see Amrein and Berliner (2002) and the analysis in Raymond and Hanushek (2003).

ing student learning. Nonetheless, as we return to below, little analysis provides information on the longer run outcomes of this nature.

STRATEGIES FOR DEALING WITH THE ANALYTICAL DIFFICULTIES

Analyzing the effects of accountability on student performance is difficult. Because accountability systems are introduced across entire states, all local school districts in a state face a common incentive structure. Thus, the only possible variation comes from interstate differences in accountability, but, as noted above, states also differ in ways other than accountability and ways in which past research has not been very informative. The difficulty is that, with little progress having been made in describing explicitly the different policies, regulations, and incentives that might be important in determining student performance, statistical estimates of accountability will be biased.

Fundamental educational policy is made at the state level and involves a wide range of factors, including financial structure, collective bargaining rules and laws, explicit regulations on educational processes, curricular specification, and so forth. The analytical complications are immediately apparent.

Consider a simple model of achievement such as:

$$O_{st} = f(X_{st}, R_{st}, \rho_s) \tag{1}$$

where O is the level of student outcomes in state s at time t, X is a vector of family and nonschool inputs, R is a vector of school resources, and ρ captures the policies of the state.[5] It is not possible to understand the impact of newly introduced accountability systems without considering the range of other factors influencing achievement.

A linearized version of this model is simply:

$$O_{st} = \beta_0 + \beta_X X_{st} + \beta_R R_{st} + (\rho_s + \varepsilon_{st}) \tag{2}$$

where the β's are unknown parameters of the educational process.[6] If, however, ρ is not observed and the β's are estimated with just information on X and R, correlations with ρ obviously lead to bias in the estimation. When background factors (X) and/or school resources (R) are correlated with state policies (ρ), these variables will partially proxy for the other policies—leading to incorrect inferences about what would happen if just X or R changed.

Now consider just adding A, a measure of whether or not accountability affects incentives and thus student performance.

$$O_{st} = \beta_0 + \beta_X X_{st} + \beta_R R_{st} + \gamma A_{st} + (\rho_s + \varepsilon_{st}) \tag{3}$$

[5] It does not matter for this discussion that we begin with aggregate outcomes for a state instead of building up from the individual student level (where the outcomes are presumably generated). The more general situation is discussed and developed in Hanushek, Rivkin, and Taylor (1996). Where the aggregation is important, we discuss the implications.

[6] The linear form is not particularly crucial but simply makes the exposition easier. An alternative model where policies act as an efficiency parameter affecting the impact of resources is developed in Hanushek and Somers (2001). Within the limited data for this study, however, it is virtually impossible to distinguish between the alternative models. The results of estimating the alternative form, discussed below, are qualitatively very close to the included estimates.

The objective is to understand γ, but under almost all circumstances γ will also be biased by omission of relevant other state policies, through either their direct correlation with accountability or with the other inputs into achievement.

Moreover, Hanushek, Rivkin, and Taylor (1996) demonstrate that the bias in any estimation will generally increase with the level of aggregation in situations like this. Specifically, when the omitted variable is relevant at the state level, estimation of the model across states will have the most bias. Note that this does not say anything about the direction of any bias, only that aggregation worsens the bias. In the case of measures of school resources, all evidence indicates that there is an upward bias from omitting state policies (Hanushek, 2003a; Hanushek, Rivkin, & Taylor, 1996). It does not, however, give much indication of how any estimation of partial models of accountability would bias analyses of γ.

If, however, the relevant state policies other than accountability are constant over our observation period, a variety of estimation approaches becomes possible. In the simplest form, simply looking at outcome changes over time eliminates any state differences that are constant over the period *t* to *t**:

$$\underset{t,t^*}{\Delta} O_s = \beta_X \Delta X_s + \beta_R \Delta R_s + \gamma \Delta A_s + \Delta \varepsilon_s \tag{4}$$

The key element is that effects of accountability systems are identified from differential changes in accountability across states during the sample period. Specifically, if all states introduced new accountability systems at the same time, ΔA would be constant, and γ would not be separately identified. This estimation relies on the variation in introduction of accountability systems over the period during which student achievement gains are observed.

But states do a variety of things to try to improve their schools—not just relying on accountability (or the absence of accountability). In order to allow for other policies that are evolving over time, we add a state fixed effect (δ_s) to the estimation, as in equation 5:

$$\underset{t,t^*}{\Delta} O_s = \beta_X \Delta X_s + \beta_R \Delta R_s + \gamma \Delta A_s + \delta_s + \Delta \varepsilon_s \tag{5}$$

Such a model can be estimated when there are multiple observations of achievement growth for each state. With multiple observations for states, achievement growth during periods of accountability can be contrasted with achievement growth when the state had no accountability.

This formulation provides much better control for other factors influencing performance growth, because the formulation effectively adds a linear trend in performance that is specific to each state. The growth formulation incorporates any state differences in policies, student and family characteristics, or other things that exert a constant influence on states performance over the relevant observation period. Adding the state fixed effect permits states to have policies that lead to trend differences in their student performance. (And, of course, the other policies of each state may or may not be effective in raising achievement, and no presumption is made about how they influence achievement.) Now estimates of the effects of accountability are identified and estimated entirely on the basis of the introduction of accountability systems within each state. In essence, the estimation relies on a state-specific prediction of performance gains and then considers how the addition of an accountability system affects outcomes.

One final issue is relevant for the estimation. The objective is to generalize about what would happen when accountability is introduced to all states. But, the analysis relies on observed student performance data, and the sample of students may not be representative of the entire population.

A school can respond to disappointing assessments in two ways. First, it can adjust teachers, curriculum, and programs in an attempt to improve the teaching that occurs. This is, however, a difficult long-run proposition, made even more difficult in schools with high rates of staff turnover. A second, shorter-run strategy may result: to become more selective about the student scores that are incorporated into the school scores. The second approach could supplement or possibly replace the first. By weeding out students who are poor performers, the school score can appear to be improving even if nothing different is being done.

The formal version of this, selection bias through testing rules, can be considered simply by looking in more detail at equation 5. The estimation of the effects of accountability (γ) depends on ΔA being uncorrelated with Δε. If in fact states and schools differentially influence who will take the tests that enter into the performance calculation, this condition will be violated.[7]

The main issue, which we explicitly consider below, is that some individual states may be more lenient in the exclusion of students for reason of language or special education, and this may correlated with the introduction of accountability. Two approaches are suggested. First, in the spirit of Heckman (1979), one can simply estimate:

$$\underset{t,t^*}{\Delta} O_s = \beta_X \Delta X_s + \beta_R \Delta R_s + \gamma \Delta A_s + \Delta p(t) + \Delta \varepsilon_s \tag{6}$$

where Δp(t) is the observed change in probability of taking the test over the observation period.[8]

Second, it is possible to estimate directly the exclusion probabilities:

$$p(t) = f(X, R, A) \tag{7}$$

This second approach, which we follow in a secondary analysis, provides direct information about the unintended outcomes of accountability systems.

Our estimation of the direct effects of accountability relies on variants of Equation 6. The essential question throughout is whether the introduction of accountability into a state alters the achievement that would be expected due to parents, school characteristics, and other policies that have also been put in place. Below we return to the estimation of whether accountability also leads to changes in the tested population in addition to any potential impacts on student performance.

DATA ON STATE SCHOOL PERFORMANCE

The primary assessment of student performance for our analysis is the National Assessment of Educational Progress. This testing, often referred to as the “nation’s report card,” provides a consistent measure of student performance that allows comparisons of students across time and across states. Although states have long had test-

[7] The selection of students, particularly special education and Limited English Proficiency students, has been frequently identified as an outcome of high-stakes testing (see Hanushek & Raymond, 2003b). Although NAEP is not a high stakes test, different exclusion practices arising from specific state rules could be related to the state accountability policies either because of purposeful actions by schools or because of simple coincidence.

[8] Note that, if the probability of exclusion from the testing is constant, this term will drop out from the growth calculations. Only changes in test taking *rates* will be relevant.

ing programs—even without an integrated accountability system—the tests differ across states and frequently change over time for any given state, thus precluding any common comparisons of states under differing accountability regimes. The focus throughout the NAEP testing (which began in 1969) has been developing assessment information for a representative sample of students at different age and grade levels.

The estimation of accountability effects uses two elements of the NAEP testing information. First, since the introduction of state level testing in 1990, NAEP has tracked performance over time for participating states. This testing provides directly useful data for two tests (mathematics and reading). The sampling/testing design of NAEP is particularly helpful because it has a basic four-year testing cycle that involves testing fourth- and eighth-graders. Thus, for example, fourth-grade tests in math in 1992 can be paired with eighth-grade math tests in 1996. Even though not the same students, this approach allows tracking the same cohort in each state, and thus holds constant common experiences of the cohort. Two cohort observations for math growth (1992–96 and 1996–2000) and two for reading growth (1994–98 and 1998–2002) make it possible to create a panel of achievement growth in each subject—thus permitting estimation that removes individual state fixed effects.[9]

Second, throughout this analysis we also disaggregate by race and ethnicity. The consistent performance data separated by Whites, Blacks, and Hispanics permits a direct investigation of relative performance gains. Note, however, that the availability of disaggregated data for Blacks or Hispanics within a state depends upon having a sufficiently large population to support separate reporting of test information. Thus, there are fewer state observations of Black and Hispanic achievement than of White achievement.

We measure student performance in a state by the average scale score on NAEP. The current accountability policies are more aimed at performance at the bottom end of the achievement distribution, suggesting that an alternative would be analyzing performance at other points in the distribution. Nonetheless, at the state level there is little information beyond the mean of the distribution. The NAEP reporting of results also provides information about students scoring at different levels. The "basic" level is a minimal degree of knowledge that roughly corresponds to what is labeled proficient in most current accountability schemes and in NCLB. For 2003 (when all states participated in testing), the correlation of eighth-grade performance in both reading and math of the state average and the percentage scoring below basic was above 0.97. (For the NAEP "proficient" level—a much higher standard—the correlation with the mean is just as high.)

The sample of student performance for the estimation thus depends both on the availability of disaggregated achievement data and on participation of the state in testing during both of the relevant testing years (for example, eighth-grade math testing in 1996 and fourth-grade math testing in 1992). Appendix Table A1 shows the relevant testing and racial/ethnic breakdowns of state observations for each of the sample periods for the separate tests. A total of 348 observations of state gains on the tests is available.[10] This sample has more state observations for performance of Whites, with fewer Black observations and even fewer Hispanic observations. Note, however, that there are more distinct states (42) than appear for any of the

[9] Note that the analysis relies on state aggregates and not individual level scores. Even though NAEP provides some disaggregated data, the testing scheme does not permit analysis of individual level performance. Pooling the data presumes that other state factors equally affect growth in both math and reading.

[10] Because of missing data on exclusions from testing, the analytical samples are reduced to 348 observations from the 351 state observations with matched fourth- and eighth-grade testing for specific cohorts.

time period-test breakdowns; a varying group of states participates in each of the tested grades and subjects for the different years.

Measured attributes of state education inputs include three primary factors: demographics, school resources, and school racial and ethnic composition. The key demographic factor we employ is the education of the adult population.[11] Although we have various measures of the education distribution, we concentrate on the percentage of the population 25 years old or older that has at least a high school education, which we calculate separately for each population subgroup and for the relevant years of testing.[12] Not surprisingly, there are significant differences in average attainment for each of the groups nationwide: Whites, 82 percent; Blacks, 74 percent; and Hispanics, 60 percent. Substantial differences in these aggregate patterns also exist across states.

School resources are measured by the average state expenditure per pupil in real terms over the relevant period. This measure cumulates the spending over the growth period being studied (that is, each relevant four-year period on which achievement growth is defined) and varies by state and time but not by subgroup.[13]

To investigate the impact of racial concentration and trends over time, we include summary data on the racial and ethnic composition across the schools in each state. Specifically, for Whites, Blacks, and Hispanics, we calculate exposure to minority students in each school of the state (using the Common Core of Data of the U.S. Department of Education). The exposure measure indicates the proportion of schoolmates who are minority for the average White, Black, and Hispanic student in the state in each year. These exposure rates are again averaged over the relevant test growth periods. The pattern of concentration of minorities by school yields disparate results for the degree of minority exposure for each group. Whites attended schools that on average over the period have 16 percent minority students, while the comparable percentages for Blacks and Hispanics are 48 and 38 percent, respectively. Exposure rates have also changed substantially both over time and across states, suggesting a potentially important element of ethnic differences in performance.[14]

Finally, although the NAEP testing provides a consistent sample of performance for the states, some variations might arise simply because of differences in the test taking procedures in the states. Specifically, over the period a variety of students

[11] The precise specification of family background factors has received relatively little attention. The Coleman Report (Coleman et al., 1966) first emphasized the importance of family background, but provided little guidance except for finding some composite measure of socio-economic status. Other analyses have emphasized particular facets such as family size and permanent income (Hanushek, 1992), but Mayer (1997) cautions against causal interpretations of income.

[12] The analysis interpolates data from the decennial censuses in 1990 and 2000 for each state and race/ethnic group to get the appropriate annual data for each state. We use the percentage of high school or more adults at the midpoint for each testing period.

[13] As we discuss below, after standardizing for intertemporal differences in school spending, geographic differences still exist. No agreed-upon cost index exists, although a variety of suggestions have been made. See the general review in Fowler and Monk (2001) and the specific data series in Chambers and Fowler (1995), Chambers (1998), Goldhaber (1999), Friar, Leonard, and Walder (n.d.), and Nelson (1991). In the context of our estimation with state fixed effects, the cross-sectional variation is inconsequential.

[14] Changes in exposure, frequently linked to school desegregation efforts, have been analyzed by Welch and Light (1987), Orfield and Eaton (1996), Orfield and Yun (1999), and Clotfelter (2004). The implications of this for school performance is, however, more controversial. Although Hanushek (2001) attributes much of the narrowing of the Black-White achievement gap in the 1980s to school desegregation (in part based on the estimated impact of racial composition in Hanushek, Kain, and Rivkin (2002), other authors suggest more uncertainty about the impacts of racial composition (compare Armor [1995] and Schofield [1995]).

could be excluded from the testing because of special conditions, including, most importantly, being identified as either a special education or Limited English Proficient student. The common presumption is that, since these students usually fall near the bottom of the achievement distribution, excluding them will artificially raise average scores of the tested population (Grissmer et al., 2000). Fortunately, NAEP provides information on test exclusions by test and year. Over the relevant time, special education placements rose for the nation as a whole and for the separate states—going from 11.4 percent in 1990 to 13.3 percent in 2001. Over that same period, test exclusions also rose, but by amounts that exceed the overall growth in the special education population. The pattern, however, differs dramatically by state, with some states actually reducing the NAEP exclusion rate while others saw very large increases. These data on NAEP exclusions permit us to adjust for whether exclusion rates increased or decreased across separate testing periods in each state (which we do in a regression framework).

THE INTRODUCTION OF ACCOUNTABILITY

States began experimenting with school accountability systems during the 1980s, but the decade of 1990s began the age of accountability. States generally worked on developing standards for what should be learned in each grade and subject, and these standards were linked to tests of student performance. Finally, states began to link tests to individual schools and to develop rating systems for performance.

Our ability to analyze the impacts of accountability comes directly from the uneven introduction of systems over time. Data on accountability come from a survey and analysis of all states by CREDO (Fletcher & Raymond, 2002). For each state, information was collected on when a state introduced an accountability system for schools. For these purposes, an accountability system is defined as publishing outcome information on standardized tests for each school along with providing a way to aggregate and interpret the school performance.[15] States are classified by whether or not they both report results and attach consequences to school performance ("consequential" states) or simply stop at providing a public report ("report card" states). As we discuss below, states have employed heterogeneous consequences ranging from monetary rewards for individual schools and school personnel to potential state takeover of schools to providing students in failed schools the opportunity to go to different schools. Most states did not, however, actually impose the consequences, particularly any sanctions, during the introductory phase. Thus, they are really potential consequences in most cases. Finally, data were also collected on when a state began disaggregating test information by subgroups of the population.

The estimation relies on the varying timing of introduction of accountability systems into the different states. Figure 1 displays the overall cumulative pattern of accountability across all states. The data are broken up into states that attach consequences to their systems and states that simply report on school achievement. To understand the estimation strategy better, the set of NAEP testing dates for eighth-grade math and reading performance is superimposed on the pattern of accounta-

[15] The survey further collected information on the method by which schools aggregated scores. The alternative approaches are discussed in Hanushek and Raymond (2003b). Note that these accountability measures pertain just to accountability for schools and do not mix in accountability for students that may have been introduced at a different time. Carnoy and Loeb (2002) employ an index of intensity of accountability that covers both school and student accountability measures but do not consider differential times of introduction. Although we concentrate entirely on school accountability, others have emphasized student accountability. See, particularly, Bishop et al. (2001) and Bishop and Mane (2004).

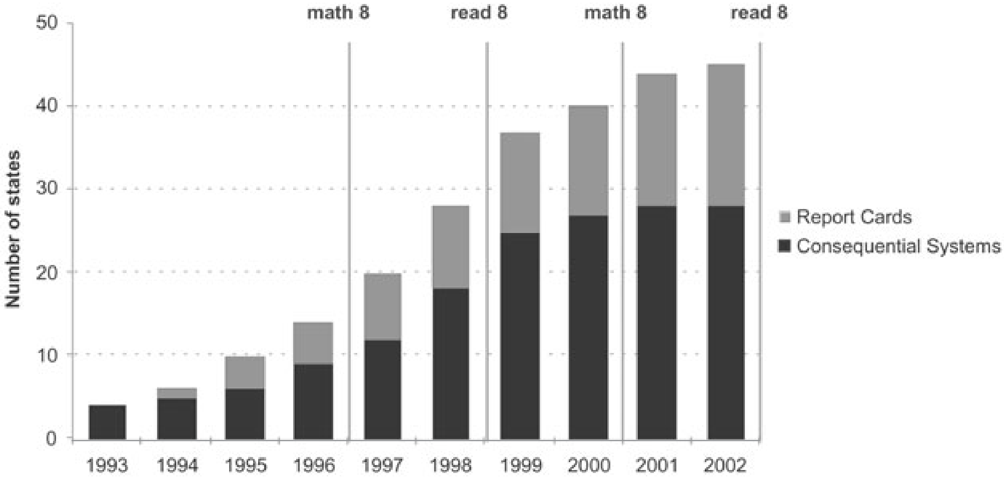

Figure 1. State accountability over time (with NAEP testing dates).

bility. The phased introduction across time and across the different testing periods permits disentangling the impact of accountability.

The validity of our estimation and the inferences that can be drawn depend crucially on the nature of the process by which accountability has been introduced across the states. At the outset, it is important that accountability is not simply a proxy for other characteristics and policies of the states. Although the empirical structure is designed to guard against other systematic influences on achievement (whether measured or unmeasured), having the timing of the introduction of accountability be related to special characteristics of states and their school system would be worrisome.

Table 1 displays the pattern of introduction of consequential and report card accountability systems for the states used in our analysis of NAEP performance.[16] The introduction appears to be geographically dispersed, but more detail is required to ascertain whether patterns exist in terms of attributes of the states.

Table 2 presents simple evidence of whether or not early implementers (1996 or earlier) differ systematically from late implementers (1997–2002).[17] This table compares a series of general characteristics of the state population, political preferences, school system characteristics, and initial student performance on NAEP. The remarkable feature about these comparisons is how little variation exists by time of adoption. None of the average differences between early and late adopters for the items shown in the table are close to being statistically significant at the 5 percent

[16] Throughout the analysis, we treat the District of Columbia as a separate state. Nine states are not included in the analysis, because we lack the requisite NAEP data on cohort scores or exclusion rates: Alaska, Iowa, Idaho, Illinois, New Hampshire, New Jersey, Ohio, Pennsylvania, and South Dakota.

[17] The dates refer to when the systems became operational as opposed to when they were established in legislation. After 2002, NCLB required all states to establish an accountability system and attached a variety of consequences to performance. Thus, all states are now effectively consequential accountability states (at least as soon as they phase in NCLB).

Table 1. States included in analytical sample by date of introduction and type of accountability systems.

Start Year	Consequential Accountability	Report Card Accountability
1993 or earlier	CT, NC, WI	
1994	TX	MS
1995	KY	IN, KS
1996	NV, OK, TN	MN
1997	AL, RI, WV	DC, MO
1998	DE, MA, MI, NY, VA	MT, WA
1999	AR, CA, FL, LA, MD, SC, VT	ME, WY
2000	GA, OR	AZ
2001		HI, NE
2002		CO
2003 or later	ND, NM, UT	

Note: Only states with NAEP scores that are employed in the analysis are included. States not included: AK, IA, ID, IL, NH, NJ, OH, PA, and SD.

Table 2. Comparison of early accountability states with later accountability states[a].

	Early Implementers (1996 or Before)	Late Implementers (1997–2002)
General economic[b]		
% ≥ High school (pop > age 25), 1990	74.0	76.0
% White students, 1994	73.1	69.3
% Black students, 1994	16.2	17.8
% Hispanic students, 1994	6.8	6.8
% Poverty, 1997	12.7	13.2
% Child poverty (age < 18), 1997	18.3	19.5
% State and local taxes, 1995[c]	10.3	10.5
Politics (presidential votes)		
% Bill Clinton vote, 1992	40.1	44.2
% George H.W. Bush vote, 1992	40.1	36.0
Schools[b]		
% State share revenues, 1994	50.6	47.9
State spending per pupil, 1994[d]	5,494	5,978
% Graduate high school (pop 18–24), 1993–95	87.2	88.0
NAEP performance[e]		
Reading, 4th grade, 1992	215.1	212.4
Reading, 4th grade, 1994	214.6	209.7
Mathematics, 4th grade, 1992	218.1	215.5
Mathematics, 4th grade, 1996	223.7	218.6

Note: a. Data pertain to states included in the NAEP analysis; see Table 1.
b. Data on the state population characteristics and school characteristics come from U.S. Census Bureau (2002) and U.S. Department of Education (2003).
c. Tax rates by states come from the web site of Tax Foundation (1995).
d. School spending is deflated by the geographic cost of living index of Friar, Leonard, and Walder (n.d.).
e. NAEP data are found on the Web site of National Assessment of Educational Progress.

level. The states have essentially the same average of education, income, racial composition, and state and local tax rates. In terms of politics, the late-adopting states were a little more likely to vote for Bill Clinton and a little less likely to vote for George H.W. Bush (implying they were more likely to vote for Ross Perot in the 1992 election). And the late adopters tend to spend a little more per student, although the state share of school revenues is very similar. Finally, although the late adopters tend to have slightly lower scores on the fourth-grade NAEP tests, the differences, as with the other elements of the table, are not statistically significant.[18]

The diffuse pattern of introduction of accountability systems lends credence to the empirical strategy set out here. Obtaining estimates of the causal influence of accountability with the aggregate data used here is clearly difficult, but the analytical design coupled with the observed, nonsystematic pattern of adoption strengthen the causal interpretation.

STATE ACCOUNTABILITY AND STUDENT ACHIEVEMENT

Our accountability analysis concentrates on the effect of state accountability on NAEP performance in the eighth grade across the three race/ethnic groups: Whites, Blacks, and Hispanics. The basic estimation pools observations across periods and tests but includes indicator variables for both. The regression estimates predict eighth-grade performance based on fourth-grade performance of students in the state four years prior. These models follow the formulation in Equation 6 into cohorts with the exception that the coefficient on prior achievement scores is not constrained to be equal to one.[19] Specific variable definitions along with descriptive statistics are found in Appendix Table A2. All estimates include individual state fixed effects. The accountability measure used throughout captures the share of the period of study when a state had accountability (that is, it ranges from 0.25 for accountability being in place for one year of the growth period for performance to 1.0 for accountability being in place for all four years). Although the impact of accountability may not be linear as we model it, the limited samples preclude any investigation of nonlinearities. The data collection was designed to measure when the accountability system became effective, not when it was legislatively passed (Fletcher & Raymond, 2002).[20]

Basic Results on the Impact of Accountability

From Table 3 we find consistent evidence that introduction of state accountability had a positive impact on student performance during the 1990s. Specifically,

[18] Carnoy and Loeb (2002) analyze differences in the intensity of state accountability systems in 2000 and suggest that school spending and racial composition of states is positively related to the number of separate components of accountability (for example, test reporting and student exit exams).

[19] Different formulations of the basic achievement model have been used in prior work, including the simple gain formulation and formulations that put the prior achievement on the right-hand side (see Hanushek, 1979). The latter formulation permits the prior achievement to have a different scale and allows for coefficients different from one, but it potentially introduces problems with errors in variables. Although the effects of this have not received much prior attention, the one correction for measurement error at the individual level found little effect (Hanushek, 1992).

[20] Nonetheless, potential state-to-state differences in the phase in of accountability systems could effectively introduce measurement error into the accountability variable. An alternative approach is simply to measure whether or not the accountability system was in effective during the period, that is, taking on the values 0 or 1. Pursuing this estimation yields qualitatively similar results, although a variety of the effects are not as precisely estimated (Hanushek & Raymond, 2004).

states that introduced consequential accountability systems early displayed more rapid gains in NAEP performance, holding other inputs and policies constant.[21] This is consistent with our prior estimates of the effects of accountability for aggregations of all students in each state (Hanushek & Raymond, 2003a, 2003b).[22] The state accountability systems diverge considerably in the types of consequences attached to performance, including both carrots (for example, bonuses to teachers) and sticks (for example, giving vouchers to students to permit shopping for a different public school or even a private school). We are unable, however, to investigate the relative power of alternative rewards or sanctions and, indeed, most states delayed employing the consequences during the initial phase of accountability. So we cannot directly observe the impacts of actually using any of the consequences.

Interestingly, we find that report cards do not have a significant influence on performance. The point estimates, although positive, are not significantly different from zero. Thus, it seems important that policies include direct incentives rather than rely on indirect forces operating through just information.

The large differences in spending per pupil never influence scores. Consistent with past evidence on the impacts of resources, the pattern of NAEP scores across states is not explained by spending. The impact of aggregate state spending is consistently small and statistically insignificant. The models in Table 3 (and in alternative specifications below) indicate that effects of parental education are imprecisely estimated and insignificant.[23] This insignificant effect of parental education and family circumstances appears to reflect the fact that relatively little variation exists within each state over the sample period.

Test exclusions always have the expected effect on tests: More exclusions from a test for special education or language increase the average growth in test score. The introduction of exclusions, however, does not affect the estimates of accountability—chiefly because the introduction of accountability was not associated with large increases in exclusions. In fact, when states introduce accountability measures, they tend simultaneously to reduce their exclusion rates by a small amount.

The remainder of Table 3 concentrates on the basic differences in performance by race. With disaggregation of performance by race (compared to aggregate state effects presented in Hanushek & Raymond, 2003a, 2003b), we see distinct differences in gains by Blacks and Hispanics. Other things equal, these subgroups show growth that is 6–10 points lower than for Whites on NAEP between fourth and eighth grade. This spread overshadows the 3.5-point gain that came with accountability. Thus, even though accountability provides a positive gain on average, that dividend is not sufficient to override the prevailing differential in performance when students are broken out by race/ethnicity. This finding of lower Black and Hispanic growth is particularly interesting in light of the narrowing of the achieve-

[21] Throughout the analysis, we report robust standard errors (Huber-White). Our analysis is essentially a difference-in-differences approach, and thus the standard errors are potentially influenced by serial correlation (Bertrand, Duflo, & Mullainathan, 2004). To allow for this possibility, we also allow for clustering by states.

[22] The prior aggregate estimates, however, did not find a statistically different impact of report card systems versus consequential systems. In the estimates here, equality of consequences and reporting is rejected at the 10 percent level or better.

[23] If these models do not include state fixed effects so that between-state information is also used in the estimation, parental education is uniformly positive and highly significant. Nonetheless, a random effects model is inappropriate in the face of unmeasured influences on achievement that bias the estimates. Moreover, the insignificant spending effects are independent of the estimation approach.

Table 3. Determinants of state growth in NAEP reading and mathematics performance (4th to 8th grade), 1992–2002.

	Basic Results	Minority Exposure
Consequential accountability	3.24 (3.0)**	3.46 (3.3)**
Report card system	0.55 (0.5)	0.76 (0.6)
% Pop (age 25+) ≥ high school	0.08 (1.1)	0.03 (0.4)
School spending, $/ADM ($1,000)	–1.44 (0.8)	–1.02 (0.5)
Change in exclusion rates	0.52 (3.7)**	0.51 (3.6)**
Black	–10.86 (4.6)**	–7.37 (3.2)**
Hispanic	–9.75 (4.0)**	–10.07 (3.4)**
Minority exposure × White		1.82 (0.3)
Minority exposure × Black		–8.48 (1.9)+
Minority exposure × Hispanic		–3.29 (0.8)
Observations	348	348
Number of states	42	42
R-squared	0.943	0.953

+ Significant at 10%; * significant at 5%; ** significant at 1%.
Note: All models estimated with state fixed effects. Models include NAEP 4th grade scores for reading and for math (lagged four years) and indicator variables for test and period. Absolute value of robust t statistics (with clustering by state) in parentheses.

ment gap that occurred in the 1980s and the subsequent explanations for this improvement.[24] The analysis of state details here that controls for state policy, family backgrounds, and testing exclusions shows a clear reversal of the trend in the prior decade.

The second column of Table 3 focuses directly on the potential influence of changing concentrations of minorities.[25] Spurred by *Brown* v. *Board of Education,*

[24] During the 1980s, Black-White achievement gaps closed by over one-half standard deviation, even though they remained 0.75–1.0 standard deviations in 2000. An influential set of papers in Jencks and Phillips (1998) examined various aspects of the pattern of test score narrowing in the 1980s, but there are very few suggestions that this narrowing would stop or reverse.

[25] Earlier discussion of the lack of progress in closing the Black-White gap in the 1990s speculated that changing patterns in school composition due to school desegregation patterns influenced the aggregate time series pattern of scores (Hanushek, 2001).

policies related to race have arguably been the most important overall school policy of the past 50 years—suggesting both that racial composition of schools may independently influence achievement and that accountability policies may be related to state-specific racial factors. In the second column, we introduce measures of exposure rates of White, Hispanics, and Blacks to minorities (Hispanics and Blacks) across the schools in each state, which are permitted to vary by subgroup.[26] Higher minority concentrations have a statistically significant negative impact on Blacks but do not significantly affect either Whites or Hispanics. This finding is generally consistent with the analysis of racial composition in Texas by Hanushek, Kain, and Rivkin (2002). In that work, Blacks were quite sensitive to school composition—specifically the proportion of Blacks in the school negatively affected Blacks, but Whites and Hispanics were unaffected by student body composition.[27]

The models discussed so far (and represented in Table 3) consider the effects of accountability to be the same across the separate groups. For a variety of reasons, the effects may not be uniform. Thus, we estimate the same basic models but permit the effects of accountability to differ by race and ethnicity (Table 4). The first column is directly comparable to the previous table, but it now indicates distinct differences by subgroup. Specifically, we see in the first column that Hispanics seem significantly more affected than Whites do by having consequential accountability, while Blacks appear significantly less affected than Whites do. In fact, although the net impact of accountability on Blacks is positive, it is not statistically significant when tested against zero.

When states introduced accountability systems, they may or may not have chosen to disaggregate the test results by racial group immediately (as now required by NCLB). Because that variation might explain differences in results by race, the second column considers the differential impact of accountability for systems with subgroup disaggregation. The similarity of results with the earlier model reflects the fact that most states in fact have disaggregated data since the beginning of their accountability programs.

In these more detailed models, we again find the strong indications that the racial composition of the schools is important for Blacks. With the substantial negative impact of increased minority exposure, Blacks do worse when attending less-integrated schools.

It is useful to understand the magnitudes of both the accountability effects and the racial differences. Figure 2 displays the expected gains for states without consequential accountability and for states with consequential accountability. These gains are based on the disaggregations in column 1 of Table 4. As can be seen, the introduction of consequential accountability leads to improved growth in NAEP performance for each of the groups. To put the gains in perspective, on average, the White improvement is 0.22 standard deviations.[28]

[26] These exposure rates are calculated on an individual school basis within each state. The variable for minority exposure in column 2 calculates exposure relative to each subgroup in the pooled sample; that is, the variable is the exposure of White students to minorities for the White subset of the sample and the exposure of Blacks to minorities for the Black subset.

[27] To test the effect of ethnic influences, a further refinement of these models (not shown) considered Black exposures to Blacks instead of to minorities (Blacks plus Hispanics). It is very difficult within these data to distinguish between the two alternative specifications. Using Black exposure for Blacks produced slightly less precise estimates ($t = 2.0$) but did not alter the other conclusions.

[28] These calculations rely on the standard deviation of average scores across states and subgroups for the eighth grade performance, which equals 16.2 scale score points. At the individual student level, the standard deviation in performance is approximately 35 points, varying slightly by test and by year.

Table 4. Differential racial and ethnic impact of accountability on state growth in NAEP reading and mathematics performance (4th to 8th grade), 1992–2002.

	Accountability by Ethnicity	Disaggregation of State Accountability
Consequential accountability	3.40 (2.8)**	3.54 (3.0)**
Consequential accountability × Black	–2.04 (2.0)*	
Consequential accountability × Hispanic	3.10 (2.4)*	
Disaggregated × Hispanic		–2.35 (2.0)*
Disaggregated × Black		3.02 (2.0)*
Report card system	0.72 (0.6)	0.72 (0.6)
% Pop (age 25+) ≥ high school	0.05 (0.7)	0.06 (0.9)
School spending, $/ADM ($1,000)	–1.14 (0.6)	–1.07 (0.6)
Change in exclusion rates	0.50 (3.5)**	0.51 (3.5)**
Black	–6.34 (2.5)*	–6.76 (2.6)**
Hispanic	–10.17 (4.4)**	–9.80 (4.2)**
Minority exposure × Black	–8.59 (2.7)**	–8.16 (2.4)*
Minority exposure × Hispanic	–4.90 (1.4)	–4.98 (1.4)
Observations	348	348
Number of states	42	42
R-squared	0.956	0.956

* Significant at 5%; ** significant at 1%.
Note: All models estimated with state fixed effects. Models include NAEP 4th grade scores for reading and for math (lagged four years) and indicator variables for test and period. Absolute value of robust t statistics (with clustering by state) in parentheses.

At the same time, the subgroup patterns both in performance gains and in the impacts of accountability clearly differ. The differences are most easily seen in Figure 3, which translates the data into the Black-White and Hispanic-White gaps in NAEP performance gains. The Hispanic-White gap in gains falls from 0.63 standard deviations to 0.44 standard deviations when consequential accountability (with disaggregated scores) is introduced in a state. But, the Black-White gap in performance actually increases with accountability (from 0.39 to 0.52 standard deviations).

Accountability systems thus lead to overall improvements in student performance on NAEP mathematics and reading tests, but they do not uniformly meet the objective of also closing achievement gaps. This finding appears to be a simple demonstration of the well-known principle that achieving multiple objectives with a single policy instrument is not generally feasible. We return to this below.

SENSITIVITY ANALYSES

Because of the relatively small samples with just 42 states, it is important to ensure that the results are not being driven by sample or model artifacts. A variety of sensitivity analyses were undertaken.

First, a number of writers have suggested the possibility that certain key states are outliers, with the implication that the results are really driven by a small number of observations. One specific set of arguments has linked performance of some of the fastest improving states to their early adoption of accountability measures. For example, Grissmer et al. (2000) suggest that the linkage of standards, testing, and accountability led Texas and North Carolina to progress at a faster rate on student achievement than California, a state with many common characteristics. At the other extreme, Washington, D.C., has long stood out on the NAEP tests for its poor performance. While heavily concentrated with minority students, it has simultaneously spent a very large amount on its schools and seen its students at the bottom of the NAEP distribution. For example, average eighth-grade math scores in 2003 were far below any of the 50 states, and only one state (Wisconsin) had Black students performing worse than those in the District of Columbia.

In a series of estimates, the commonly identified outliers (CA, DC, NC, and TX) were dropped individually and in combination from the estimation. The main effects of accountability were quantitatively very similar in all of these subsamples. The only change was to reduce the statistical significance of the benefits of accountability for Hispanics when both California and Texas were simultaneously excluded. We interpret this latter result as simply reflecting the loss of important information about Hispanic education by losing California and Texas. Overall, we conclude from these experiments that a few outliers are not driving our accountability results.

Second, the main estimation, while permitting different impacts of accountability by ethnicity, constrains a variety of the other influences on achievement (education, spending, and fourth-grade test scores) to be the same across subgroups of the population. Again, because we concentrate on racial and ethnic differences in achievement growth, it is important to ensure that the constraints on other inputs are not driving the variations across subgroup that we find. To test this, we permit each of the constrained inputs to vary by ethnicity by having a full set of interactions for Blacks and Hispanics. When we do this, all of the prior findings about overall accountability effects and the interactions of accountabil-

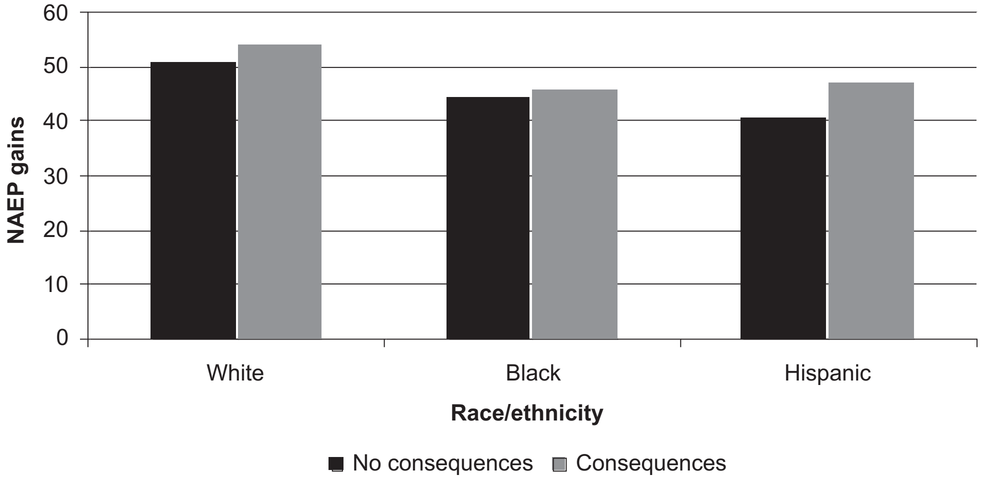

Figure 2. Effect of consequential accountability on achievement by race/ethnicity.

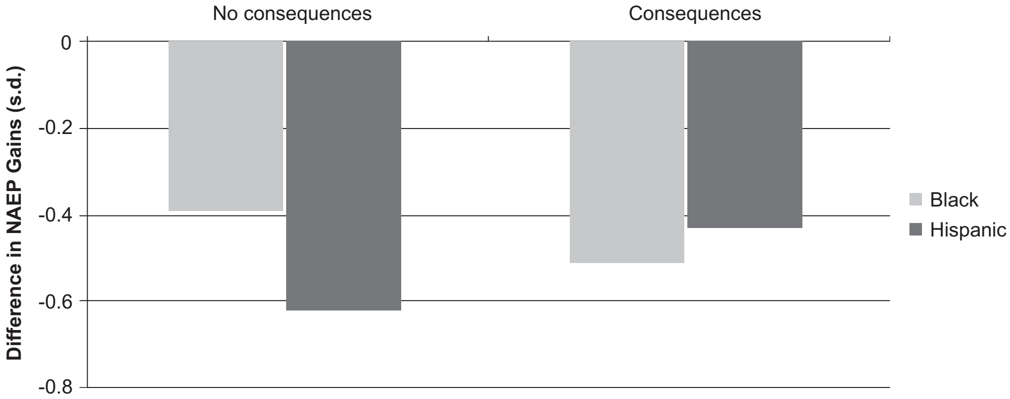

Figure 3. Racial/ethnic gaps by consequential accountability status (NAEP gains relative to whites).

ity with ethnicity still hold, with the exception that Black students no longer have a different reaction to accountability that is statistically significantly different from Whites. The overall impact of accountability in this fully augmented model is nonetheless not significantly different from the estimates in Table 3.

Preliminary Analysis of the Quality of Accountability Systems

The previous analysis employs a relatively blunt measure of whether or not a state introduces a system of consequential accountability for schools. In reality, much discussion has gone into the nature of learning standards, the testing regime, and the precise accountability system that relates to the testing. Indeed, heated debate has occurred over the nature of educational standards and how they should be assessed (Evers, 2001).

The nature of educational policy in the United States is such that states have the ultimate authority for educating their students. As such, each state has developed its own standards and many have designed unique tests that apply to the state's specific learning goals and objectives. Yet, because most of the state programs cover precisely the same subject areas, it is possible to develop qualitative ratings of the different state's efforts.

We develop a preliminary analysis that goes into the quality of the standards and testing. In particular, various ratings of the standards and accountability of different states have entered into the debate about state policies. We employ three separate ratings that give us information on all of the states. The Thomas B. Fordham Foundation has done extensive evaluations of the standards in each state for the major subject areas (Finn Jr. & Petrilli, 2000).[29] In a similar effort, *Education Week,* the leading trade publication for elementary and secondary education, undertook a larger evaluation of state standards, testing, and accountability (*Education Week,* 2001). Finally, Carnoy and Loeb (2002) develop an index of "intensity" of accountability in 2000, based on a count of various components of school and student accountability. We use the ratings in each of these to provide qualitative distinctions among the states adopting consequential accountability.

These ratings are not ideal. They each provide a snapshot of the states for one year, 2000. While there is clear consistency over time in the nature of standards and accountability, there is also some change, and this change will introduce error into our analysis that uses the quality ratings to judge systems at different times.[30] Because of the problems of measurement errors about quality, however, we consider this a preliminary investigation of the range of likely effects.

Table 5 displays alternative estimates that permit the impacts of consequential accountability systems to vary according to quality rankings. (These correspond directly to the estimates in column 1 of Table 4.) The first two columns use alternative versions of the Fordham rankings, the third introduces the *Education Week* ratings, and the fourth investigates the Carnoy-Loeb ratings. The Fordham rankings cover the educational standards in five subject areas of which we use data on mathematics and English to align with our NAEP data. They provide a numerical rating of systems, but the scale of these ratings (which differ by subject) is clearly arbitrary. We transform the ratings into a standardized score (mean 0 and standard deviation 1.) We then use the average for English and mathematics in column 1 and the specific subject linked to the NAEP reading and mathematics gains in column 2. At the mean quality, the estimates of the overall impact of consequential accountability are virtually identical to that found in Table 4: 3.22 or 3.19 added points of growth here versus 3.40 when quality is not considered. The average score provides a statistically significant adjustment to this overall effect when we use the average Fordham grade.

[29] The Fordham Foundation has also evaluated the accountability systems (built on the various standards), but it does not have a comprehensive rating of the states.

[30] The Fordham Foundation had a similar exercise in 1998. In comparing the two ratings of states, there was substantial stability but a noticeable improvement over time (Finn Jr. & Petrilli, 2000).

Higher-quality standards are associated with higher student performance. According to the Fordham rankings, three states with consequential accountability (North Carolina, Alabama, and California) had ratings more than one standard deviation above the mean—implying that students in these high-standard states would be expected to have an additional average gain of two or more points on the eighth-grade NAEP due to introducing consequential accountability with high standards. On the other hand, the two states at the bottom of the Fordham rankings—Rhode Island and Michigan—are estimated to gain nothing from introducing consequential accountability with very weak educational standards.

The *Education Week* evaluations are broader than those of Fordham and used different grading criteria for the states. Interestingly, the two ratings have a correlation of just 0.38.[31] Column 3 of Table 5 provides separate estimates of the impact of these rankings on expected gains. Again, states that rank higher are expected to gain more, although the estimate of the overall impact of accountability at the mean state quality now falls to 2.64 points. At the lower end, Rhode Island again is expected to gain nothing from consequential accountability with its low-rated system. On the other hand, the *Education Week* evaluations place Michigan at the B level—a very high grade.[32]

The final column of Table 5 indicates that the Carnoy-Loeb index of accountability, also standardized, adds little. The overall impacts of consequential accountability remain similar to the others in Table 5, but the modification for intensity in 2000 is not statistically different from zero.

The discrepancies in the ratings between Fordham, *Education Week*, and Carnoy-Loeb lead to some hesitation. These suggest issues of reliability with the subjective ratings. Coupled with concerns about possible changes in ratings over time, we believe that the specifics of these findings should be treated with caution. On the other hand, they suggest clearly that the quality of standards, testing, and accountability systems has important ramifications for student learning.

CONCOMITANT EFFECTS: SPECIAL EDUCATION PLACEMENT

As many people have suggested, there is an immediate incentive in most existing accountability systems to exclude students who might be expected to have low achievement. A method often discussed is to place students into special education and thereby exclude them from testing and from subsequent inclusion in the accountability system. The previous analysis of the impact of accountability on achievement explicitly controlled for alterations in exclusions from NAEP testing, but the exclusion behavior is interesting in its own right.

Several studies have investigated whether schools appear to react to accountability through exclusions. Jacob (forthcoming) considers the introduction of test-based accountability for Chicago public schools. He finds that the large increases in test scores after accountability went into effect were also accompanied by increases in special education placement and by increased grade retentions. Deere and Strayer (2001a, 2001b) and Cullen and Reback (2002) also find apparent increases in special education placement with the introduction of accountability in Texas. Prior work in Kentucky by Koretz and Barron (1998) suggested no strategic use of

[31] The *Education Week* ratings are also standardized to mean zero and standard deviation one. The ratings are available for a common set of 48 states. *Education Week* does not rate D.C., and Fordham does not rate Idaho and Iowa because they lacked codified standards at the time of rating.

[32] If both Fordham average ratings and *Education Week* ratings are simultaneously included, neither is individually significant, although jointly they are significantly different from zero.

Table 5. Effect of quality ratings for standards and accountability on the estimated impact of consequential accountability on state growth in NAEP reading and mathematics performance.

	Fordham Average	Fordham English/Math	Education Week	Carnoy-Loeb
Consequential accountability	3.22 (3.0)**	3.19 (2.8)**	2.64 (2.0)*	2.87 (1.7)$^+$
Consequential accountability × Fordham average	1.92 (2.0)*			
Consequential accountability × Fordham Eng/math		1.07 (1.7)$^+$		
Consequential accountability × Ed Week			1.83 (2.5)*	
Consequential accountability × Carnoy-Loeb				0.93 (1.0)
Consequential accountability × Black	–2.07 (2.1)*	–2.07 (2.1)*	–2.16 (2.0)*	–2.14 (2.0)*
Consequential accountability × Hispanic	3.01 (2.3)*	3.04 (2.3)*	2.56 (1.9)+	2.60 (1.9)$^+$
Observations	348	348	339	339
Number of states	42	42	41	41
R-squared	0.957	0.956	0.957	0.957

$^+$ Significant at 10%; * significant at 5%; ** significant at 1%.
Note: All models estimated with state fixed effects. Models include NAEP 4th grade scores for reading and for math (lagged four years), % greater than equal to high school education, real school spending, change in exclusion rates, minority exposure rates for Blacks and Hispanics, and indicator variables for Black, Hispanic, test and period. Absolute value of robust t statistics (with clustering by state) in parentheses. The *Education Week* and Carnoy-Loeb measures are not available for Washington, DC.

grade retentions. Haney (2000) suggests that both grade retention and increased dropouts were key to improvements in Texas tests, although both Carnoy, Loeb, and Smith (2001) and Toenjes and Dworkin (2002) seriously question this after reanalysis of the data.[33] Any grade retentions are, however, short-run effects that do not provide lasting "accountability" value except if the placement is educationally valuable. Figlio and Getzler (2002) concentrate on special education placement after the introduction of a state accountability system in Florida. The most persuasive evidence is that placement rates increase relatively over time in grades that enter into the accountability system as opposed to those grades that do not.

In each case, the analysis considers changes that occur around the time of introduction of an accountability system. In fact, the key element of most of this research is using the change in accountability to identify the effects on special education

[33] Carnoy, Loeb, and Smith (2001) also find that at least in larger urban areas lower dropout rates are associated with higher student achievement.

placement rates and the like through finding breaks in the patterns of prior placement. Three things are important. First, there are very little relevant data for these analyses, relying on, say, breaks in trends, perhaps compared to trends of other schools outside the specific district. The validity of the interpretation depends crucially on whether or not other things are changing over time that could also affect the patterns of observed changes. Second, since later periods are always compared with earlier periods, there is concern about the general trend in special education placement that has been going on for two decades. Third, each of these analyses provides information just on the short-run immediate effects. Since the incentives change over time, it is important to understand what happens as these systems continue.[34] Because of the recentness of introduction of accountability systems, little is know about the long-run dynamics.

In order to test the importance of this incentive, we study the responsiveness of special-education placement rates to the introduction of an accountability system. We concentrate on the period 1995–2000, the period of large growth in state accountability systems as described in Figure 1. As with the achievement analysis, our basic strategy is to relate special-education placement rates to accountability and other factors that might affect placement.

For this analysis, we turn to an expanded sample with annual information on all states.[35] (Note, however, that it is not possible to disaggregate these data by race and ethnicity, so we concentrate on overall state behavior.) The basic modeling considers special education placement rates across all 50 states plus the District of Columbia. All estimation includes state-specific fixed effects.[36]

The "standard approach" found in the existing literature simply considers comparisons of placement rates before and after the introduction of accountability systems or how placement rates differ with time since the introduction of accountability systems. (The difference between consequential and report card systems was never significant in the estimation, so the analysis relies entirely on the combination of the two.) This standard model is then compared with a simple analysis that allows for national time trends in placement rates.

The standard approach results in Table 6 show that the introduction of an accountability or report card system is associated with a roughly 1.5 percentage point higher special-education placement rates in a state. These estimates are essentially generalizations of difference-in-difference estimators that allow for comparisons across all of the states. The second column indicates that the reaction to accountability occurs over time, with a 1.1 percentage point higher placement rate with consequential or report card systems, and with an increase of 0.4 percentage point each year that the system is in place. Thus, these estimates appear to confirm the estimates from individual states and districts.

[34] Hanushek and Raymond (2003b) consider the incentives that are set up by the design of different accountability systems. Although the method of aggregating student performance and of judging change over time has an impact, the main conclusion is that incentives to exclude are generally largest in the first year of an accountability system and then decline, if not reverse, in subsequent years. This change in incentives results from the fact that exclusions in one year are generally built into the base for the next year, so that exclusions in any year must be maintained in subsequent year or they will lead to potential reductions in scores. Moreover, getting added gains from exclusions over time requires continual increases in the exclusion rates.

[35] Data on special education placement by state comes from annual reports to Congress, U.S. Department of Education (2001) and prior years.

[36] Note that states that always operate under accountability in this period (10) and that never do (10) do not contribute to the estimation of the effects of accountability on special education.

The final two columns that place special education placement into the national picture, however, show a markedly different picture. The final columns introduce a time trend and its square to allow for the strong and ubiquitous increases in special education placement. Columns 3 and 4 show that both the effect of having a consequential or report card system and the effect of how long such a system has been in effect have an insignificant impact on placement rates (in terms of magnitude and of statistical significance) once the overall trends are considered.

We have also duplicated this analysis in terms of the logarithms of special education placement rates. In this formulation, the state fixed effect will provide an estimate of the state-specific trend in special education placement, and the accountability term indicates how the state trend differs with the introduction of accountability. These results (not shown) provide virtually identical findings (which is not surprising given the close similarity of the specifications with the short time series). Additionally, we estimated the basic special education models for the 42 states that entered into our NAEP analysis and found no significant differences.

These estimates do not indicate that gaming is irrelevant to accountability but they do suggest caution in interpreting analyses of the magnitude of gaming under accountability systems. If such gaming were generally important in the case of special education, it should show up in the national data—but it does not. Moreover, the national trends in special-education placement offer a ready explanation for the divergent results.

SOME CONCLUSIONS

Considerable public attention has focused on school accountability. While many states were pursuing their own versions of accountability, the discussion was elevated to new heights during the 2000 presidential campaign when George W. Bush made school accountability a centerpiece of his domestic policy platform. Indeed, the first year of his presidency involved significant pressure on Congress to enact accountability legislation—which it did with the No Child Left Behind Act of 2001.

NCLB mandated that all states introduce accountability systems that included annual testing of all students in grades three through eight by 2006 and disaggregated data on student performance for all schools. This complex law also consid-

Table 6. Effect of accountability on special education placement rate, 1995 through 2000.

	Standard Approach		Allowance for Placement Trend	
Consequential or report card system	1.45 (10.1)**	1.09 (7.9)**	.11 (1.0)	.10 (.9)
Time in place		.38 (7.9)**		–.02 (–.5)
Time trend			.86 (12.4)**	.87 (14.4)**
Time trend squared			–.08 (–6.3)**	–.08 (–6.0)**

** Significant at 1%.

Note: Estimation employs a panel of special education placement rates for all states and the District of Columbia over the period 1995–2000. Estimation includes a fixed effect for each state. The t-statistics appear below each estimate. Time trend = 1 in 1995; = 2 in 1996; and so on.

ered the development of state performance goals along with a variety of sanctions if schools failed to meet those goals.

NCLB has yet to be fully implemented, thus precluding a direct analysis of it.[37] Nonetheless, because NCLB calls for each state to design its own system and because most states have keyed off their existing systems, the analysis here of the impacts of state systems enacted prior to NCLB provides information about what can be expected with full implementation.

The most important result is that accountability is important for students in the United States.[38] Despite design flaws in most existing systems (Hanushek & Raymond, 2003b), we find that they have a positive impact on achievement. This significantly positive effect of accountability holds across a series of alternative specifications of the basic achievement model.

However, the impact holds only for states attaching consequences to performance. States that simply provide information through report cards without attaching consequences to performance do not get significantly larger impacts than those with no accountability. Thus, the NCLB move toward adding consequences to accountability systems is supported. At the same time, however, no existing work provides information about the best set of rewards and sanctions, and more work is needed in that area as accountability systems are further refined.

It is useful to put the detailed subgroup impacts into perspective. Accountability significantly increases student achievement gains, particularly for Hispanics. However, because both Blacks and Hispanics generally show smaller gains relative to Whites on each of the tests, accountability by itself is insufficient to close the gap in learning.

We also find that the effect varies by subgroup, with Hispanics proportionally gaining most and Blacks gaining least. Because Whites gain more than Blacks do after accountability is introduced in absolute terms, the racial achievement gap actually widens with the introduction of accountability.

In addition to accountability, the analysis looks into other determinants of student performance. Most relevant for consideration of where we stand 50 years after *Brown* v. *Board of Education,* Black students are hurt by greater minority concentration in the schools. This compositional effect has no significant influence on White or Hispanic scores, making the effects very similar to those found in Hanushek, Kain, and Rivkin (2002).

These findings, taken together, underscore the fact that there is no one answer that will lead to all of the improvements that we desire. The introduction of consequential accountability systems has a clearly beneficial impact on overall performance. But other forces are simultaneously pushing the distribution of performance—particularly as observed in the Black-White achievement gap—in less desirable ways. First, accountability as seen during the 1990s tended to help White achievement more than Black achievement. Second, the observed movement toward higher minority concentrations in schools has a detrimental effect on Black achievement, again pushing toward a wider distribution of achievement.

[37] Somewhat ironically, when implemented, NCLB essentially precludes analysis of further impacts of overall accountability systems, because it eliminates any comparison group of states without accountability systems. Since, however, individual states will still follow their own locally developed schemes, it will still be possible to contrast the impacts of alternative types of accountability systems and alternative rewards and sanctions.

[38] A variety of countries around the world are also developing testing and accountability systems. Some, such as that in the United Kingdom, are quite well-developed and offer some insights for the United States, including approaches to identifying the value added of different schools.

The finding of differential effects of accountability raises a clear policy dilemma. A prime reason for the U.S. federal government to require each state to develop a test-based accountability system involved raising the achievement of all students, but particularly those at the bottom. It has done that, but not at the same rate across groups. We conclude from this that additional policies are needed to deal with the multiple objectives. Again, as is frequently the case, a single policy cannot effectively work for two different objectives—raising overall student performance and providing more equal outcomes across groups.

The movement toward stronger accountability in schools has also suggested to many that there would be adverse consequences—more exclusions, higher dropout rates, a narrowing of the curriculum, and the like. While some existing research supports these presumptions, we conclude that the negative impacts are likely to be considerably overstated (Hanushek & Raymond, 2003b). Importantly, many of the adverse effects that involve "gaming" the system come from short-run incentives that are unlikely to persist over time. Moreover, our own analysis of special education placement rates indicates clearly that accountability has not had an overall impact through this form of exclusions.

Although we have not dwelled on it, the character of currently available accountability systems is not particularly strong. This concern has two dimensions. First, the educational standards and accountability systems vary dramatically across states, and our preliminary analysis suggests that independent assessments of the quality of different systems are associated with stronger achievement gains. Nonetheless, it is not possible with our data to distinguish clearly among alternative quality ratings.

Second, a majority of the systems concentrates on overall achievement levels (with highly variable passing scores across states). Such systems do not generally provide clear signals about the value-added of schools.[39] Instead, they combine a variety of effects, including those resulting from family background differences and neighborhood effects. As such, they cannot provide truly clear and strong incentives. Moreover, while there is a range of potential consequences incorporated in state systems, it is not possible to investigate whether some specific consequences are more powerful than others.[40] Yet, in the face of the rather blunt incentives from existing systems, the introduction of accountability systems is associated with achievement improvements on the order of 0.2 standard deviations. Such improvements, while not revolutionary, are notable when compared to the failure to find alternative reforms that yield such impacts on a broad basis. As accountability systems evolve, they are likely to have considerably stronger impacts if they move in the direction of more precise incentives for individual schools.

Finally, even if our estimates correctly identify the causal impact of accountability systems and even if they are strengthened and improved, policy cannot stop there. We have pointed out the remaining issues in terms of the distribution of outcomes—guaranteeing that existing ethnic gaps in performance are ameliorated.

[39] The common approach in a majority of states is to track how average school scores change over time. This approach, even though an improvement over just the level of student scores, does not accurately identify the gains that individual students are making—largely because of the high student mobility rates found in most U.S. schools. See Hanushek, Raymond, and Rivkin (2004).

[40] The inability of investigating differences is partially a reflection of the wide range of approaches that the individual states have taken—including both rewards and sanctions—and the lack of any suitable way to characterize the differences. Moreover, most states have yet to unleash the consequences, particularly those arising from sanctions, so the only available response really relates to potential consequences and not to what happens when explicit actions are taken.

Beyond that, however, we need to ensure that any gains achieved through middle school are continued and reinforced in high school and in college. This continuation is not assured, at least from tracing past performance of our schools.[41]

Without doubt, the achievement of our students has direct ramifications for the future well-being of our society (Hanushek, 2004). It should be a very high priority to ensure that all of our students do in fact gain the skills that will be needed as our economy grows and evolves.

Helpful comments were provided by Christopher Jencks, Paul Peterson, and anonymous referees. Lei Zhang provided valuable research assistance. This research was supported by the Packard Humanities Institute.

ERIC A. HANUSHEK is Senior Fellow of the Hoover Institution/Stanford University, Chair of the Executive Board of the Texas Schools Project/University of Texas at Dallas, and Research Associate of the National Bureau of Economic Research.

MARGARET E. RAYMOND is Director of CREDO and Research Fellow of the Hoover Institution/Stanford University.

REFERENCES

Amrein, A.L., & Berliner, D.C. (2002). The impact of high-stakes tests on student academic performance: An analysis of NAEP results in states with high-stakes tests and ACT, SAT, and AP test results in states with high school graduation exams. Tempe, Arizona: Educational Policy Research Unit, College of Education, Arizona State University (December).

Armor, D.J. (1995). Forced justice: School desegregation and the law. New York: Oxford University Press.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? Quarterly Journal of Economics, 114, no.1 (February), 249–275.

Bishop, J., & Mane, F. (2004). Educational reform and disadvantaged students: Are they better off or worse off? CESifo-Harvard University/PEPG Conference on Schooling and Human Capital in the Global Economy: Revisiting the Equity-Efficiency Quandary, Munich (September).

Bishop, J.H., Mane, F., Bishop, M., & Moriarty, J. (2001). The role of end-of-course exams and minimal competency exams in standards-based reforms. In D. Ravitch (Ed.), Brookings papers on education policy 2001, pp. 267–345. Washington, DC: Brookings.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. Educational Evaluation and Policy Analysis, 24(4), 305–331.

Carnoy, M., Loeb, S., & Smith, T.S. (2001). Do higher state test scores in Texas make for better high school outcomes? Research Report Series, RR–047, Consortium for Policy Research in Education (November).

Chambers, J.G. (1998). Geographic variations in public schools' costs. Working Paper No. 98–04, National Center for Education Statistics (February).

Chambers, J., & Fowler Jr., W.J. (1995). Public school teacher cost differences across the United States. Washington, DC: National Center for Education Statistics.

[41] Two pieces of evidence are relevant. First, gains that have been made in NAEP for fourth- and eighth-graders have not translated well into gains for 17-year-olds as they prepare to go on to school or jobs. Second, this fall-off has been apparent in comparing U.S. students to foreign students of various ages on international examinations (Hanushek, 2003b).

Clotfelter, C.T. (2004). After Brown: The rise and retreat of school desegregation. Princeton, NJ: Princeton University Press.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). Equality of educational opportunity. Washington, DC: U.S. Government Printing Office.

Cullen, J.B., & Reback, R. (2002). Tinkering toward accolades: School gaming under a performance based accountability system. Department of Economics, University of Michigan (mimeo).

Deere, D., & Strayer, W. (2001a). Closing the gap: School incentives and minority test scores in Texas. Department of Economics, Texas A&M University (mimeo) (September).

Deere, D., & Strayer, W. (2001b). Putting schools to the test: School accountability, incentives, and behavior. Working Paper 113, Private Enterprise Research Center, Texas A&M University (March 2001).

Education Week. (2001). Quality counts 2001: A better balance. January 11 ed. Washington, DC: Education Week.

Evers, W.M. (2001). Standards and accountability. In T.M. Moe (Ed.), A primer on America's schools. Stanford, CA: Hoover Institution Press.

Figlio, D.N., & Getzler, L.S. (2002). Accountability, ability and disability: Gaming the system? National Bureau of Economic Research, W9307, (November).

Finn Jr., CE., & Petrilli, M.J. (Eds.) (2000). The state of state standards, 2000. Washington, DC: Thomas B. Fordham Foundation.

Fletcher, S.H., & Raymond, M.E. (2002). The future of California's academic performance index. CREDO, Hoover Instiution, Stanford University (April).

Fowler Jr., W.J., & Monk, D.H. (2001). A primer on making cost adjustments in education. Washington, DC: National Center for Education Statistics.

Friar, M.E., Leonard, H.B., & Walder, J.H. (n.d.) The federal budget and the states: Fiscal year 1995. Cambridge, MA: Taubman Center for State and Local Government, John F. Kennedy School of Government, Harvard University.

Goldhaber, D.D. (1999). An alternative measure of inflation in teacher salaries. In W. J. Fowler (Ed.), Selected papers in school finance, 1997–99. Washington, DC: National Center for Education Statistics, 33–54.

Grissmer, D.W., Flanagan, A., Kawata, J., & Williamson, S. (2000). Improving student achievement: What NAEP state test scores tell us. Santa Monica, CA: Rand Corporation.

Haney, W. (2000). The myth of the Texas miracle in education. Education Policy Analysis Archives, 8(41) August.

Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of educational production functions. Journal of Human Resources, 14(3), 351–388.

Hanushek, E.A. (1992). The trade-off between child quantity and quality. Journal of Political Economy, 100(1), 84–117.

Hanushek, E.A. (2001). Black-White achievement differences and governmental interventions. American Economic Review, 91(2), 24–28.

Hanushek, E.A. (2002). Publicly provided education. In A.J. Auerbach & M. Feldstein (Eds.), Handbook of public economics, pp. 2045–2141. Amsterdam: Elsevier.

Hanushek, E.A. (2003a). The failure of input-based schooling policies. Economic Journal 113, no.485 (February), F64–F98.

Hanushek, E.A. (2003b). The importance of school quality In P.E. Peterson (Ed.), The importance of school quality in our schools and our future: Are we still at risk? pp. 141–173. Stanford, CA: Hoover Institution Press.

Hanushek, E.A. (2004). Some simple analytics of school quality. National Bureau of Economic Research, Working Paper 10229, (January).

Hanushek, E.A., Kain, J.F., & Rivkin, S.G. (2002). New evidence about Brown v. Board of Education: The complex effects of school racial composition on achievement. Working Paper 8741, National Bureau of Economic Research (January).

Hanushek, E.A., & Raymond, M.E. (2003a). Improving educational quality: How best to evaluate our schools? In Y. Kodrzycki (Ed.), Education in the 21st century: Meeting the challenges of a changing world, pp. 193–224. Boston, MA: Federal Reserve Bank of Boston.

Hanushek, E.A., & Raymond, M.E. (2003b). Lessons about the design of state accountability systems. In P.E. Peterson & M.R. West, No child left behind? The politics and practice of accountability, pp. 127–151. Washington, DC: Brookings.

Hanushek, E.A., & Raymond, M.E. (2004). The effect of school accountability systems on the level and distribution of student achievement. Journal of the European Economic Association, 2(2–3), 406–415.

Hanushek, E.A., Raymond, M.E., & Rivkin, S.G. (2004). Does it matter how we judge school quality? Paper presented at American Education Finance Association annual meetings, March 11–13, at Salt Lake City, UT.

Hanushek, E.A., Rivkin, S.G., & Taylor, L.L. (1996). Aggregation and the estimated effects of school resources. Review of Economics and Statistics, 78(4), 611–627.

Hanushek, E.A., & Somers, J.A. (2001). Schooling, inequality, and the impact of government. In F. Welch (Ed.), The causes and consequences of increasing inequality, pp. 169–199. Chicago: University of Chicago Press.

Heckman, J.J. (1979). Sample selection bias as a specification error. Econometrica, 47, 153–161.

Iceland, J., & Weinberg, D.H. (2002). Racial and ethnic residential segregation in the United States: 1980–2000, Census 2000 Special Reports. Washington, DC: U.S. Bureau of the Census.

Jacob, B.A. (forthcoming). Accountability, incentives and behavior: Evidence from school reform in Chicago. Journal of Public Economics.

Jacob, B.A. (2003). High stakes in Chicago: Did Chicago's rising test scores reflect genuine academic improvement? Education Next, 3,(1), 66–72.

Jencks, C., & Phillips, M. (Eds.). (1998). The Black-White test score gap. Washington, DC: Brookings.

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

Mayer, S.E. (1997). What money can't buy: Family income and children's life chances. Cambridge, MA: Harvard University Press.

National Assessment of Educational Progress. (2004). The nation's report card, 26 August 2004]. Available at: http://nces.ed.gov/nationsreportcard/; accessed September 2004.

Nelson, H.F. (1991). An interstate cost-of-living index. Educational Evaluation and Policy Analysis, 13(1),103–111.

Orfield, G., & Eaton, S.E. (1996). Dismantling desegregation: The quiet reversal of Brown v. Board of Education. New York, NY: The New Press.

Orfield, G., & Yun, J.T. (1999). Resegregation in American schools. Harvard University, The Civil Rights Project (June).

Peterson, P.E., & West, M.R. (Eds.). (2003). No child left behind? The politics and practice of accountability. Washington, DC: Brookings.

Raymond, M.E., & Hanushek, E.A. (2003). High-stakes research. Education Next, 3(3), 48–55.

Rossell, C.H., Armor, D.J., & Walberg, H.J. (Eds.). (2002). School desegregation in the 21st century. Westport, CT: Praeger.

Schofield, J.W. (1995). Review of research on school desegregation's impact on elementary and secondary school students. In J.A. Banks & C.A. McGee (Eds.), Handbook of research on multicultural education, pp. 597–616. New York: Macmillan Publishing.

Smith, M.S., & O'Day, J. (1990). Systemic school reform. In S.H. Fuhrman & B. Malen (Eds.), The politics of curriculum and testing, pp. 233–267. London: Falmer Press.

Tax Foundation. (1995). Comparing the 50 states' combined state/local tax burdens in 1995, April 2004 1995 [cited 2004]. Available at: http://www.taxfoundation.org/statelocal95.html: accessed September 2004.

Toenjes, L.A., & Dworkin, A.G. (2002). Are increasing test scores in Texas really a myth, or is Haney's myth a myth? Education Policy Analysis Archives, 10(17).

U.S. Census Bureau. (2002). County and city data book: 2000. Washington, DC: U.S. Government Printing Office.

U.S. Commission on Civil Rights. (1967). Racial isolation in the public schools. Washington, DC: Government Printing Office.

U.S. Department of Education. (2001). To assure the free appropriate public education of all children with disabilities: Twenty-third annual report to Congress on the implementation of the Individuals with Disabilities Education Act. Washington: U.S. Department of Education

U.S. Department of Education. (2003). Digest of Education Statistics, 2002. Washington, DC: National Center for Education Statistics.

Welch, F., & Light, A. (1987). New evidence on school desegregation. Washington, DC: U.S. Commission on Civil Rights.

Winter, G. (2002). More schools rely on tests, but big study raises doubts. New York Times, December 28, p. 1.

APPENDIX

Table A1. Number of state observations for analysis by race/ethnicity, test, and sample period.

	White	Black	Hispanic	Total
Mathematics				
1992–1996	35	29	32	96
1996–2000	34	26	32	92
Reading				
1994–1998	32	27	16	75
1998–2002	34	29	22	85
Total	135	111	102	348

Table A2. Variable definitions and sample descriptive statistics: means and standard deviations (in parentheses) by race/ethnic group.

		All	White	Hispanic	Black
Consequential accountability	Proportion of period with school accountability system having consequences for the school; Fletcher and Raymond (2002)	0.39 (0.45)	0.38 (0.44)	0.35 (0.44)	0.44 (0.45)
Report card system	Proportion of period with report card system; Fletcher and Raymond (2002)	0.14 (0.31)	0.14 (0.31)	0.14 (0.31)	0.14 (0.31)
Disaggregated	Proportion of period with school accountability system disaggregated by race/ethnic subgroups			0.37 (0.43)	0.40 (0.45)
%Pop (age 25+) ≥ high school	% of population age 25 and older with a high school degree or greater; interpolation for period of decennial census data by race/ethnicity between 1990 and 2000	71.2 (11.8)	81.7 (5.1)	58.2 (8.3)	70.5 (7.3)
School spending, $/ADM	Average expenditure per pupil in average daily membership for growth period (2000 $)	6109 (1354)	6005 (1273)	6202 (1431)	6149 (1383)
Change in exclusion rates	NAEP exclusion rates: difference in 8th grade and 4th grade lagged four years by test	–0.16 (2.9)	–0.17 (2.8)	–0.25 (2.9)	–0.11 (3.0)
Minority exposure	Average exposure rate to minorities (Black + Hispanic) by school averaged across growth period years	0.38 (0.24)	0.16 (0.17)	0.45 (0.21)	0.57 (0.17)
NAEP8	Average scale score, NAEP 8th grade test	257.4 (16.2)	274.8 (8.5)	249.0 (9.6)	244.0 (6.4)
NAEP4	Average scale score, NAEP 4th grade test	207.6 (16.0)	224.7 (6.2)	200.4 (10.5)	193.5 (7.3)
Special education placement rate	State placement rate for special education (percent)	12.63 (2.0)			
Accountability (report card or consequence)	= 1 if either consequential or report card system in place; = 0 otherwise	0.493 (0.5)			
Time in place	Years since introduction of accountability system	1.046 (1.7)			