Publication version

Improving Educational Quality: How Best to Evaluate Our Schools?

Eric A. Hanushek and Margaret E. Raymond
Hoover Institution, Stanford University

Paper prepared for

Education in the 21st Century: Meeting the Challenges of a Changing World
Federal Reserve Bank of Boston
June 19-21, 2002

Abstract

Test based accountability systems are now a central feature of U.S. education policy. Accountability systems are implemented as a way to improve student outcomes through new, highly visible incentives. In analyzing the effectiveness of such state systems, the correct comparison is not accountability versus no accountability but the differential effects related to the type of system that is employed. The alternative systems that states have developed offer very different incentives.

State accountability systems differ in a variety of important ways. We concentrate on the school activities that are the focus of attention (outcomes or a mixture of inputs and process measures), the scope of the measures included, and the design of reporting and incentives. We then consider the available evidence on impacts of different systems. While research on the outcomes of accountability systems is growing rapidly, it still represents a young and highly selective body of work. The existing research suggests that schools definitely respond to the incentives of accountability systems, but the form and strength of such responses is highly variable. We conclude with new evidence about the impacts of accountability systems on achievement and on special education placement rates.

IMPROVING EDUCATIONAL QUALITY: HOW BEST TO EVALUATE OUR SCHOOLS?

Eric A. Hanushek and Margaret E. Raymond*

It is difficult to be against accountability for public schools. Schools are creatures of state and local government, with all the associated expectations of performance and oversight. The importance of education for individuals and for society is unassailable. But many believe that U.S. schools are not contributing as much as they could and are not competitive in comparisons with those of other countries. Thus, the desire to hold schools responsible for outcomes is natural.

The disagreement comes, however, soon after people acknowledge the importance of school accountability. How should performance be assessed? Is providing information on student outcomes sufficient to get improvement? Should there be explicit sanctions and rewards for students and/or schools? Do unintended consequences overwhelm the intended consequences?

This paper considers some of the basic features of school accountability systems and assesses both the incentives for change that are imbedded in these systems and the existing evidence we have about behavior under different systems. The essential features that we consider are focus, scope of measurement, design, and incentives. "Focus" describes the mix of factors examined within accountability systems. "Scope" considers the extent to which the accountability system captures the full range of school activities for each of the factors under review. "Design" refers to the specific approaches to measuring the schools' contribution and the

* Paul and Jean Hanna Senior Fellow, Hoover Institution, Stanford University, and Senior Research Associate, Green Center for the Study of Science and Society, University of Texas at Dallas; and Research Fellow and Director of CREDO, Hoover Institution, Stanford University, respectively. This research has been supported by grants from the Packard Humanities Institute and the Smith Richardson Foundation.

precision of these measurements. Incentives are created by the interplay of these three aspects of accountability systems and illuminate the ways schools will react to these initiatives. Some prior analyses provide reasonable tests of various incentives in action, and we provide some new evidence about the early impact of accountability systems.

The existing accountability discussion is surprisingly vague both in terms of what is being done and what should be done. Since much of the work to date has focused on single systems or isolated attributes or effects, it is hard to make informed judgments about accountability as a policy. A preliminary step of the analysis is to provide a description of where accountability stands in the United States today. This is essential for any evaluation of where accountability systems should be going.

The Focus of an Accountability System

In almost all the states that have implemented school accountability to date, the overriding concern is the achievement of students. In contrast to policies of earlier periods, the chief focus of accountability is results, not effort. Most of the enabling legislation explicitly states that the purpose of adopting school accountability systems is to reflect student achievement outcomes and school performance.

That having been said, states have made differing choices in program design that have narrowed the range of outcomes and that frequently have involved other school characteristics. The premise of our discussion is that schools will respond most strongly to data elements that are included in the program, and that those aspects of schooling that are not included or measured

will be de-emphasized, distorting school responses.[1] We consider a variety of aspects of this issue below; here we concentrate on the kinds of measures included in accountability systems.

In early 2002, CREDO surveyed each state in an effort to understand the details and effects of accountability systems, an area in which prior data have been very scarce.[2] Most states provide information on the districts in their state, and many have now taken this information down to the level of individual schools. Just providing unprocessed information is, however, considerably different from developing aggregate performance measures and putting rewards and sanctions into place. We distinguish between simple "report cards" and accountability systems by the presence of aggregate measures that can be assessed against a standard, and by the use of rewards and sanctions related to measured performance.

While emphasizing student performance, another distinction between states is that in many cases the accountability legislation calls for the inclusion of other factors that do not measure outcomes. A significant number of states rely on a mix of process and input measures as well as outcome measures. With such a blend, those states hold schools accountable for the way students are taught in addition to considering the outcome of those efforts.

The incentives of hybrid systems are ambiguous: A school could be rewarded for improving its procedures, even if it does not result in additional student achievement. In contrast, an exclusive outcome-orientation creates incentives for schools or districts to direct resources appropriately in order to maximize the outcomes being studied. Outcome measures illustrate most clearly the degree to which schools are achieving the educational goals for their students.

[1] The incentive effects of choice of measures to be used in rewards have been extensively considered in the economics literature about optimal contracts; see, for example, Baker (1992, 2002) and Dixit (2002).

Table 1 lists measures that have been incorporated into school accountability formulae or have been proposed for adoption in legislative bills, divided according to whether they are input, process, or outcome measures. Only 10 of the variables in Table 1 are outcome measures. Six are measures of school activities and are classified as process variables. Six are measures of inputs.

State systems that consist only of student test scores in each school have the virtue of being exclusively outcome-focused. This is not to say that they are perfect, because they can still be rather narrow in their coverage. Expansion of those systems with other outcome measures might add depth to the outcomes picture and still retain clear incentives for schools. However, where states include input or process measures, the strength of the association between the new measure and student achievement determines the degree to which the incentives are dulled. If the relationship between these other factors and student achievement is strong, then the combination would be less compromised than if the strength between them was weak. In short, the potential incentive rests on the degree of alignment between the measured factors and the outcome of interest.

Here we provide a summary of what the education literature indicates about the strength of the relationship of each variable to student achievement. Table 2 classifies each variable in one of three ways based on how strongly it aligns with student achievement and on the weight of existing empirical evidence. If the relationship has not been studied or the evidence is weak or inconclusive, we considered it to have "weak" support for inclusion in a school accountability system. If there is conclusive evidence about a variable but the estimated impact on student achievement was low, we concluded the strength was "moderate." If the conclusive research

[2] See CREDO (2002) for full details and citations of the analysis. CREDO, formerly the Center for Research on Education Outcomes, is an independent research unit at the Hoover Institution and has the mission of promoting and

Table 1
Variables Used or Proposed for Use in Accountability Systems, by Type

Input	Process	Outcome
Teacher Attendance Rate	Student Attendance Rate	State Achievement Tests (various grades)
Condition of School Facilities and Grounds	Percent of Students Taking State Test	College Entrance Exam Scores
Number of Computers	Principal Mobility	Drop-Out Rate
Course Offerings	Student Mobility	Graduation Rate
Number of Non-Credentialed Teachers	Teacher Mobility	Number of Students in Advanced Courses
School Crime Rate	Year-Round School Status	Parent/Community Satisfaction
		Percent of Students Passing End-of-Course Exams
		Percent of Students Passing the High School Exit Exam
		Retention Rate
		Suspension Rate

Source: CREDO (2002).

Table 2
Strength of Relationship to Student Achievement of Accountability Variables in Use or Proposed for Use

Weak	Moderate	Strong
College Entrance Exam Scores	Condition of School Facilities and Grounds	Drop-Out Rate
Course Offerings	Percent of Students Taking State Test	Graduation Rate
Number of Computers	Student Attendance Rate	Number of Students in Advanced Courses
Number of Non-Credential Teachers	Teacher Attendance Rate	Percent of Students Passing End-of-Course Exams
Parent/Community Satisfaction	Year-Round School Status	Percent of Students Passing High School Exit Exam
Principal Mobility		Retention Rate
School Crime Rate		Student Mobility
Teacher Mobility		Suspension Rate

Source: CREDO (2002).

showed a close and robust association, it was designated as having a "strong" relationship. (Note that we consider direct measures of student achievement tests as obviously a strong measure of outcomes and thus do not include them in this part of the analysis).

The resulting classification, when put in terms of the underlying type of measure, shows an interesting pattern, as revealed in Table 3. The input variables were found largely to have a weak relationship with student achievement. Process variables have more mixed relationships to student achievement. Of the three types of variables, the outcome variables show the strongest association with student achievement, with two exceptions. The outcome measures of Parent/Community Satisfaction and College Entrance Exam Scores are weak indicators for the same reason other kinds of measures are weak—they show insufficient correlation with overall school quality. Public opinion research has documented a constant positive regard by parents for their children's schools despite actual differences in performance. College entrance exam scores, while providing some information, are self-selective and reflect only a segment of the student body of a school and thus can provide misleading summaries of the school outcomes because the sampling fractions are generally unspecified.

While we have no formal tests here, we assert that states whose accountability measures are more closely aligned with student outcomes deliver more consistent incentives to their schools. If a school faces consequences—good or bad—for teachers' professional development and for academic achievement, for example, the school will seek to allocate resources and place emphasis on both these dimensions of its operations. In its simplest form, we expect decisions to be made in accordance with the school's ability to change the measured factor, the cost of changing it, and with the reward (or punishment) associated with the change. This response is natural. However, if the two dimensions are not strongly related, a school could end up working

---

assisting in the evaluation of education programs and policies.

Table 3
Pattern of Classification Variables by Strength of Association with Student Achievement[a]

Rating	Input	Process	Outcome
Weak	4	2	2
Moderate	2	3	0
Strong	0	1	8

[a] Value is the number of variables from Table 1.
Source: Authors' calculations based on Tables 1 and 2.

at cross-purposes as they, say, pursue superior opportunities for teacher training and also work to improve student learning. Taking teachers out of the classroom for their professional development activities may actually work against improved learning in their classes.

Since many of the inputs and processes are more concrete than the outcomes—we know how to order more computers or to deliver new programs—they are the low-hanging fruit on the accountability tree. Any elements that are associated closely with the more difficult and desirable objectives of student achievement reinforce the incentives that prompt schools to take corrective action. However, since the majority of the input and process measures currently in use do not meet that standard, they dilute the strength of the output incentives and generally weaken the system.

Scope of Measurement

The scope of an accountability system highlights the breadth of focus that a state elects to adopt. The scope of the accountability system will have an effect on how strong the incentives are and how much latitude or slack schools retain to minimize the impetus to change. Interviews conducted with state officials in 2001-02 suggest that the strength of the incentives is directly related to the comprehensiveness of the program that a state implements (CREDO 2002).[3]

States appear to have been influenced in their choice of scope by several factors. There may be resource constraints that necessitate a narrow focus. Political dissent about implementing accountability in any form may require concessions in the breadth of the program.

[3] Much attention has been given to the potential implications of narrowly focused accountability systems. Most frequently this is raised with respect to the types of achievement-measurement instruments that are employed. For example, do they emphasize just "lower-order" skills or do they concentrate on items most easily included in standardized testing? But the debate also includes issues of whether concentration on basic cognitive skills drives out other elements such as citizenship development, character education, and the like.

Some states may wish to proceed cautiously in order to be able to adjust incrementally as the program matures. And there is some evidence that states may even have genuinely different theories about their appropriate role in gauging school performance. CREDO (2002) mentions all these factors as explanations of the varying structures implemented by state officials and of the differences that are observed across systems.

The implications of many of these larger issues are difficult to assess, in part because they have not been clearly linked to measurable outcomes. We can nonetheless look at some of the factors that enter directly into school teaching programs.

The clearest indication of differences in scope can be seen by considering the number of grades of schooling that the accountability system covers and whether the included grades are sequential or discontinuous. Both aspects have an effect on the incentives that an accountability system produces and their effect on schools.

Table 4 presents the 50 state systems classified by the number of grades included in their testing system. Eighteen states include five or fewer grade levels, 24 states cover between six and eight grades, and eight states have nine or more grades. Even among states with the same accountability model, differences in the strength of the incentives will arise from differences in the scope of their performance focus.

Note that we are not able to judge the quality or breadth of the separate state examinations. We do record whether the tests are criterion-referenced (developed for the specific objectives of each state's schools) or norm-referenced (more generic tests applied across the nation). This division provides some information about the relationship between each state's testing program and its educational goals and standards. Nonetheless, it is a rather coarse cut across the testing programs.

Table 4
Classification of States by the Number of Grade Levels Assessed in 2001

Minimum **Less than 5 Grade Levels**	**Better** **5 – 8 Grade Levels**		**Best** **9 or More Grade Levels**
Connecticut	Alaska	South Carolina	Alabama
Georgia	Arkansas	Texas	Arizona
Hawaii	Colorado	Utah	California
Indiana	Delaware	Vermont	Idaho
Iowa	Florida	Virginia	Mississippi
Maine	Illinois	Washington	South Dakota
Minnesota	Kansas		Tennessee
Montana	Kentucky		West Virginia
Nebraska	Louisiana		
Nevada	Maryland		
New Hampshire	Massachusetts		
New Jersey	Michigan		
New York	Missouri		
North Dakota	New Mexico		
Ohio	North Carolina		
Oregon	Oklahoma		
Wisconsin	Pennsylvania		
Wyoming	Rhode Island		

Source: CREDO (2002).

Among states with more grade levels included in the school scores, differences remain. As reflected in Table 5, which shows the grades and types of assessments currently in use by states, the majority of states capture achievement from an erratic pattern of grade-level testing. Compare, for example, the states of South Carolina, Texas, Utah, and Vermont. All rely on test scores from six grades, but in South Carolina and Texas the grades are consecutive. This is not the case for Utah and Vermont, which both sample from three elementary grades, one middle school grade, and two high school grades. Clearly, states with consecutive grades have an easier time attributing changes in school scores more accurately to their own activities. Beyond that benefit, schools with consecutive grades face steady incentives across those grade levels, which we would expect to result in consistent attention to each grade on the part of schools. With discontinuous grades, the opportunity exists to focus more strongly on the grades under review.

While the pattern of test taking is likely to change with recent federal legislation on testing and accountability, the message at this point is clear. Most states do not have a broad and uniform assessment policy across grades. This disjoint nature of testing both affects how well information can be used to judge school performance and alters some of the incentives faced by schools. It is to these latter points the analysis proceeds.

Design and Incentives

Concentrating on student performance as the key focus of accountability will obviously transform the practice of the past, when a majority of states provided just rudimentary information about their schools, often confined to a few measures of school resources and avoiding any indication of student performance (Hanushek and Raymond 2001). Even where states have created a hybrid system that combines input and outcome regulatory elements,

Table 5

Type of Assessments Being Used in States and the Grade Levels Being Assessed

	Norm-referenced	Criterion-referenced		Norm-referenced	Criterion-referenced
Alabama	3-11	5-7	Montana	4,8,11	
Alaska	4,5,7,9	3,6,8,10	Nebraska	4,8,11	
Arizona	2-11	3,5,8,10,11	Nevada	8,10	
Arkansas	5,7,10	4,6,8,11	New Hampshire		3,6,10
California	2-11	2-11	New Jersey	4,5,8,11	
Colorado		3,4,5,7,8	New Mexico	3-9	
Connecticut		4,6,8,10	New York		4,8,12
Delaware	3,5,8,10	4,6,8,11	North Carolina	3-8,10	
Florida	3-10	3-10	North Dakota	4,6,8,10	
Georgia	4,8	4,6,8,11	Ohio		4,6,9,12
Hawaii	3,6,8,10		Oklahoma	4,5	5,8,9-12
Idaho	3-8	4,8,9-11	Oregon		3,5,8,10
Illinois	3,5,8-12		Pennsylvania		5,6,8,9,11
Indiana	3,6,8,10		Rhode Island	4,8,10	3,7,10,11
Iowa	4,8,11		South Carolina		3-8
Kansas		4-8,10-11	South Dakota	2-11	
Kentucky	3,6,9	4,7,8,12	Tennessee	3-8	9-12
Louisiana	3,5,6,7	4,8	Texas		3-8
Maine	4,8,11		Utah	3,4,5,8, 10,11	
Maryland		3,5,8,9,11	Vermont		2,4,6,8,10,11
Massachusetts		4-8,10-11	Virginia	4,6,9	3,5,8
Michigan		4,5,7,8,11	Washington	3,6	4,7,10
Minnesota	3,5,8,10		West Virginia	3-11	4,7,10
Mississippi	5-8	2-12	Wisconsin		4,8,10
Missouri		3-5,7-11	Wyoming	4,8,11	4,8,11

Source: CREDO (2002).

student outcomes have become a major focus. Yet, the appropriate use of student outcome information is far from obvious. The ways that states compile student achievement measures into school scores and how they treat those results create very different pictures of school performance.

Most of the accountability systems have implicit or explicit goals underlying them. In many cases, the goals are multiple; for example, to improve student achievement and to narrow the historical gap in performance across racial and ethnic groups. Thus the design of a system serves as the vehicle for translating desired goals into incentives that motivate schools toward these goals and capture the results for review. To the extent that a design ignores one or more goals or creates conflicting motivations, the system that relies on that design will likewise distort incentives.

SUMMARY MEASURES

The key to understanding the informational content provided by state systems is to examine the determinants of student performance and how those determinants are displayed within the accountability system in each state. As a foundation, prior work on the determinants of student achievement identifies student outcomes as coming from a variety of influences: families, friends, teachers, and schools. Moreover, a student's knowledge evolves and builds on past learning and on the individual's skills and abilities. How these various influences are recognized and accounted for dramatically influences the ability of state officials to discern the performance of schools and to provide clear incentives.

Accountability systems begin by testing a group of students in each school and then presenting information about school achievement. The actual measure of school achievement

varies. The simplest measure is the average of test scores of the students in a grade or an entire school, although few states end up developing their accountability systems on just school-average achievement. Important variants include distributional information such as the percentage of students scoring above some specific level ("passing" or "proficient"). These variants introduce important elements into accountability systems, but for now, we consider just the average performance measures. Virtually all states, whether they provide just report card information or instead develop accountability structures, report average achievement as one of the components of the information given.

Status Model

The status model simply uses the average performance of students as a measure of the outcomes in each school. (While more important later, we do not distinguish at this point between systems built on calculating grade averages as opposed to school averages).[4] The first point from this is obvious: If the main purpose of the accountability system is assessing the performance of the school, the average test score does it very imperfectly. In addition to school performance, the average achievement will incorporate all of the current and historical inputs to achievement including not only school but also family background and random errors. With the status model, it is not possible to factor out year-to-year changes in student-body composition, or grade-to-grade changes in instructional design or teacher quality. Thus, the simple average score indicates the level of student performance but cannot pinpoint the source of that performance. Despite these imprecise measures, schools are treated according to the result, for better or worse.

[4] For average performance the distinction is unimportant, but a variety of state reward systems are based on such measures as the percentage of students passing a grade-level test. In those, performance requirements or rewards based on separate grades imply different incentives and constraints compared to school-based systems.

This basic confusion between average student achievement and the contribution of schools is well known, and some state accountability and reporting systems provide additional information that might be useful for adjusting these scores to get closer to the impact of schools. For example, some states either provide data on family backgrounds (such as rates of free lunch participation or racial compositions of schools) or describe achievement for reference groups of students judged to have similar family backgrounds. While these measures are usually available, they generally act merely as an external reference, but do not influence the results of the accountability calculations. Thus, these approaches highlight issues of accurate estimation of school performance, because they likely do not adequately identify family differences or cohort differences and they do not capture prior factors that affect current achievement. Nor do they allow for any measurement errors in performance. Most of the attention has focused on ways of trying to allow for differences in the nonschool factors, but existing efforts have simply produced imprecise results, leaving considerable uncertainty about interpretation of scores and little way to separate out the value-added of the school.

One other aspect of status models is important—the relationship between goals and incentives. An underlying element explicit or implicit in most accountability discussions is that schools have systematically left minorities and disadvantaged students behind. In reaction, explicit goals of narrowing and eliminating the existing gaps have been translated into status accountability models built on unadjusted aggregate scores. This confuses goals with the incentives of accountability systems, because each school finds that incentives include aspects of performance that it does not control. Put another way, if one school has students who come to school with poorer preparation than another, that school must meet a higher standard in terms of its value-added to student learning.

A variant of the status model considers performance just for separate grades, instead of aggregate school performance. While the approach is still cross-sectional in nature, and, therefore, vulnerable to shifts in student composition, it provides a more precise focus on school inputs. The approach can help to provide schools with the ability to distinguish between school inputs and student variation. The effect from student migration will still exist, but cohort effects will be seen as they move across grades. With stable programs and teachers, teacher effects will persist over time.[5]

The grade-level variation of the status model of accountability also is used when testing does not cover the range of grades. If, for example, testing is done only at the fourth grade, the accountability system would feature just that grade.

Status-Change Model

The status-change model tracks the average student achievement of a school over time. The idea is easiest seen in terms of an example. The status-change score for a school that has a common examination at a specific grade, say third-grade reading, would appear as the change in the average third-grade reading result between the 2000 and 2001 school years. The status-change model is often calculated for an entire school by aggregating the performance across tested grades.

The status-change model is by far the most common approach to assessing what is happening in schools. Change scores frequently factor heavily in reward systems, but they are

[5] Note that the interpretation of year-to-year grade or status changes depends crucially on which information is used. If looking at just the difference in performance across cohorts of students, the relevant school effect is the change in school quality. If levels of performance are calculated at each year, information about the level of school quality inputs can be obtained.

treated in a wide variety of ways: Examples include absolute levels of change, percentage increments of change, and change relative to an external standard.

The most common interpretation, regardless of form, is that the status-change model provides a measure of the change in performance of the particular grade or school. Thus, for example, states may have goals or rewards related to the "progress" that is measured by the status change. Indeed, recent federal legislation also incorporates change in testing and accountability requirements. Does the accountability system built on status change provide biased estimates of performance improvement that systematically diverge in one direction or another? Are the errors so large that they mute any incentives for schools to do better?

Even if the student body of a school is identical across years, the status-change model is still comparing two different groups of students. Thus, status change has three primary components: the difference in school quality across the two years; the difference in family background and other nonschool factors between the two groups of students; and the average difference in any idiosyncratic errors affecting achievement. Just like the status model that relies on the level of average achievement, the status-change model completely entangles school performance with student-background differences and measurement errors. The best interpretation would be that, if variations in quality improvements across schools were large relative to differences in the other factors, changes in grade or school performance would dominate the changes. But, there is little existing evidence that would support that interpretation.

The situation is, however, even worse than many believe because of the dynamics of student populations. The mobility of the U.S. population has important implications for schools—not only for the way they teach students but also for their accountability systems. The U.S. population moves a surprisingly frequently. From a recent Current Population Survey, we

find that only 55 percent of students live in the same house over a three-year period, and this falls to half for disadvantaged students. Moreover, residential mobility is often related to significant changes in family circumstances such as divorce or job loss and change. In growing states the mobility rates are noticeably higher. The average annual student mobility across schools in Texas, for example, exceeds 20 percent (Hanushek, Kain, and Rivkin 2001) and in California the figure is 15 percent (CREDO 2002).

The implications of mobility for the accountability approaches are clear. As mobility increases, differences in the backgrounds, preparation, and abilities of two groups of students compared over time will influence differences in aggregate performance in the status-change model. At that point, not only do current differences in nonschool factors enter the picture but historical differences also do—and mobility implies that two adjacent cohorts will also diverge in terms of the past schools they attended.

COHORT- AND INDIVIDUAL-GAIN MEASURES

By shifting attention to the progress of students rather than schools over time, it is possible to gain substantial accuracy in the focus of the accountability system. Consider following the same students in a school, year to year, and calculating the improvement or decline for the cohort. The result is a new measure of school performance that has some superior characteristics. With a stable student body (that is, with no in- or out-migration for the school), the historical school and nonschool factors would cancel out (because they influence a cohort's performance both in the current grade and in the prior grade). The cohort-gain score would then reflect what the school contributed to learning plus any differences in idiosyncratic test factors across the two grades. The influence of family differences on current achievement growth rates

would also remain, so that if, for example, disadvantaged students would be expected to have lower rates of improvement in performance than the more advantaged, such differences would remain confounded with school factors. Nonetheless, the cohort model would generally yield a closer measure of the school's contribution than the status model. The family background and ability factors that affect the cohort-gain calculations are ones that affect the rate of growth of learning, not the level. Thus, they would be expected to be relatively small.[6]

The final design that has begun to be used by states further refines the progress model by calculating gain scores for individual students and then creating school summaries by aggregating them by grade and by school. This approach provides the highest level of precision because it controls for family differences and differences in student body composition, and it isolates the year-over-year contribution of schools to student performance. Because it follows individual students, including in-migrants, it minimizes the effects of student variation. Cohort effects are still uncontrolled to the extent that a specific group of students may be brighter or duller than average (perhaps by design through exclusions). Since additionally it focuses on progress, the model can isolate the contribution of individual teachers, although no state makes such information public.[7]

The array of states under the different types of systems is presented in Table 6. The vast majority of states rely on cross-sectional measures and comparisons—even though these approaches generally have the least appealing properties. Only four states (Massachusetts, New Mexico, North Carolina, and Tennessee) currently emphasize student gains. The implications for incentives and results are developed in the next section.

[6] Some practicalities of calculations still remain. The primary question is how to deal with any mobility that might enter into the calculations.

[7] Tennessee produces measures of individual student value-added, but they are not publicly released (Sanders and Horn 1994).

Table 6
Classification of States by the Type of Analysis Model Used in School Rating Systems in 2001

Cross-Sectional Approaches				Student-Change Approaches	
School Status or Status-Change Model		Grade-Level Change		Cohort-Gain	Individual-Gain
Alabama	New Hampshire	Alaska	Louisiana	New Mexico	Tennessee
Arkansas	New York	Colorado	Oklahoma	North Carolina	Massachusetts
California	Ohio	Delaware	Rhode Island		
Connecticut	Oregon	Florida	Vermont		
Georgia	South Carolina	Kentucky	Wisconsin		
Maryland	Texas				
Michigan	Virginia				
Mississippi	West Virginia				
Nevada					

Source: CREDO (2002).

INCENTIVES AND EVIDENCE

It is useful to translate the discussion on the different accountability systems into hypotheses about the incentives introduced by each. We then provide a review of existing evidence about these hypothesized effects. It is important to bear in mind, however, that the recent birth of many accountability systems means that the existing evidence is thin in many crucial places. Indeed, the thinness of the evidence is one of the main points of this analysis.

Accountability systems are designed to increase the exposure of schools by revealing the quality of student performance. Two separate mechanisms operate: the public sharing of performance data and any directly legislated rewards and consequences.

Any school will prefer higher scores to lower ones, even if no explicit consequences follow the awarding of scores. Currently, apparently in the absence of much clear evidence, most parents appear to think that their school is doing a good job (Rose and Gallup 2001). The sharing of accountability evidence has the potential for changing this, perhaps sufficiently enough to overcome the inertial positive regard for local schools. In the absence of direct consequences, one might expect any purely informational incentive to be small relative to organizational pressures to maintain the status quo. Nonetheless, some general evidence on reactions of citizens (in the form of housing prices) to perceived school quality information exists (Black 1999; Weimer and Wolkoff 2001). Moreover, as discussed below, early evidence suggests that public disclosure of scores may in fact produce some strong incentives, both in terms of housing prices (Figlio and Lucas 2000) and other observable outcomes.

The second source of incentives from exposure of performance arises from any consequences that might be directly associated with the school scores. The rewards and

sanctions that many states have built into their accountability systems create the motivation for schools to change behavior. At the same time, one does not expect these incentives to affect all schools equally. For example, schools that have many students scoring close to a threshold might be expected to alter their behavior more than schools with students further away from the established critical thresholds.[8] The interrelationship between the choice of a school-score model, the choice of thresholds, and the location of a given school relative to those thresholds is currently relatively unexplored, but it would be reasonable to speculate that no single design can provide equivalent incentives for all schools. Moreover, it is well known that incentives that emphasize crossing a specific threshold will generally lead to ever greater distribution in behavior.

The following sections consider in more detail the incentives under different accountability models. Within each section, we also provide a review of the existing evidence about the impact of the various incentives.

Cross-Sectional Approaches

As delineated in the preceding discussion, the status model combines one-time scores of student performance into a single school score. Any change in scores from year to year generally is assumed to be a function of school influences. But, since the design does not recognize changes in the underlying student population, the model creates the incentive to include more positive student test scores into the school scores, that is, to adjust the relevant test-taking population.

[8] See, for example, the parallel with past incentives employed in the experiments with performance contracting, where contractors reacted very openly to the notches in the contracts (Gramlich and Koshel 1975).

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 dynamics of these alternative approaches are important. Take the example of a third-grade student from a disadvantaged background who arrived at school less well-prepared than the others in the school and who progressed at a slower rate each year through the third (that is, falls further behind over time). The status model compares performance of individual classes each year to the prior year's class. Thus, if testing begins in the third grade and the system has been going for some time, the school might exclude this slow student through placement in special education or by counseling the student to be absent on the day of testing. If the student is excluded in the third grade, the average of all remaining students would be higher than otherwise, and the school would tend to look better in comparison to the third grade in the prior year. But, the next year's comparison of third grades will be worse because the base comparison has been artificially elevated. Moreover, once the school has excluded a student, there is a continuing incentive to keep the student out of the testing. This continuing incentive puts some restraint into the system, because the school probably cannot increase the exclusion rate year after year. Moreover, since the potential importance of exclusion rates is widely recognized, the school is always at risk that regulatory changes may make it necessary in the future to bring some previously excluded students back into the accountability system.

While the largest effects of exclusion on the school ratings come in the first year of exclusion (when the cumulative effects to the current grade of low preparation plus slow learning are removed), there are some continued accountability benefits to the school from exclusion if the omitted students learn at a slower pace. The status model aggregates across grades, so the slower learning pace will be removed from the calculation of the school average for the student's fourth grade and beyond. The key element of this part of the dynamics is how much the rate of learning might be below average, as opposed to the absolute level of deficit that comes into play in the first year of exclusion.

While there has been widespread attention to such things as test preparation and cheating, these seem to be the clearest cases of one-time effects that are not sustainable after the initial introduction. Specifically, these practices may shift the level of performance in a given year, but, unless their prevalence increases over time, they will not show up in the school gains after the first year. Take, for example, efforts to teach all students how to fill in mechanical scoring sheets for standardized exams. Once students know how to do this—something that might inflate their scores through eliminating errors arising just from coding mistakes—it would not be expected to have any continuing effects on their scores as they progress through the grades. Similarly, any cheating on a given test must be repeated in subsequent years just to stay at the same level, but scores will improve only if the level of cheating is increased over time.

The choice of approach may be assumed to follow rational choice: School officials would select the action that they perceive to have the highest yield, given their planning horizon, budget, and appetite for risk. The preceding discussion highlights the fact that the largest gains from exclusions operate in the first year and that these decline or possibly reverse in subsequent years. Administrators may be very myopic or may have very short time horizons for their

decisions, leading them to overuse exclusions in the first years of an accountability system. Regulatory restrictions frequently are designed in an effort to limit the ability of administrators to increase the use of student exclusions.

The grade-level change variation of the status model of accountability introduces some additional incentives. Some of the dynamics of exclusions are altered. But also there may be incentives to concentrate attention on the tested grade(s), say by placing the best teachers in the relevant testing grades.

Study of the exclusion rates of schools is one way to detect if schools are culling their student ranks prior to testing. Alternatively, one could examine the prevalence of parental waivers, with attention to which students are being held out. Finally, consideration of the effects of state policies on when students who change schools must be included in the new school's score could provide another perspective on exclusions.

Several studies have investigated whether schools appear to react to accountability through exclusions. Jacob (2002) 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 on Kentucky by Koretz and Barron (1998) suggests no strategic use of grade retentions. Haney (2000) suggests that both grade retention and increased dropouts were key to improvements in Texas tests, although this finding is seriously questioned by reanalysis of the data. Both Carnoy, Loeb, and Smith (2001) and Toenjes and Dworkin (2002) find little evidence that testing led to the changes suggested by Haney. Carnoy, Loeb, and Smith also find that at

least in larger urban areas lower dropout rates are associated with higher student achievement. The grade retentions are, however, short-run effects that do not provide lasting 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.

Jacob finds that scores also appear to go up more in subjects that enter into the accountability system than in those that do not. This evidence is consistent with analysis in Texas by Deere and Strayer (2001b). The interpretation is not, however, entirely clear. Schools obviously appear to be responding to the accountability system—which is exactly what the system is supposed to accomplish. On the other hand, one might question whether the weights on different potential outcomes are appropriate. (Zero weight or not paying attention to specific subjects, for example, appears to provide very strong incentives to change the pattern of instruction).

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 placement rates and the like by finding breaks in the patterns of prior placement. Two things are important. First, there is very little relevant data for these analyses—breaks in trends or perhaps comparisons to trends of other schools (such as schools outside of Chicago and its accountability system) convey limited information. 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, 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. Because of the recentness of introduction of accountability systems, little is known about the long-run dynamics.

Hanushek and Rivkin (2002) investigate the impacts of public disclosure of achievement performance. Specifically, before the Texas accountability system included direct consequences or sanctions for performance, the state made information on disaggregated student performance from the Texas Assessment of Academic Skills (TAAS) available to the public. They find that in the largest metropolitan area, competition works to push up average scores.

Greene (2001a, 2001b) analyzes the Florida A+ program that provides exit vouchers to students in failing schools and finds that schools at risk of becoming sites of vouchers make unusually large gains. Carnoy (2001) reviews this evidence and suggests that the reaction to vouchers that Greene identified was more likely a reaction to information. Carnoy finds that similar studies for North Carolina and Texas (Ladd and Glennie 2001 and Brownson 2001, respectively) investigating what happens to failing schools show similar results—dramatic improvements in the year after identification. This occurs even though those states had no voucher threat.

On the other hand, Kane and Staiger (2001) suggest that a portion of the school improvement in North Carolina's failing schools may simply result from measurement errors in the examination scores. They demonstrate that small schools—where the error variance in aggregate tests will be larger—are much more likely to be found at the extremes of the school score distributions. If the measurement errors are independent over time, schools that realized a large error in one period would expect to receive a smaller one the next period, leading to a re-ordering of schools in the second year. Kane and Staiger do not, however, differentiate among

the sources of error of the status model—family differences, teacher and school differences, and measurement errors.

The implications of grade-level versions of accountability have been less studied. Some of the prior work employs differences by grade level primarily as a method of identifying the behavioral effects of the system as opposed to being a focal point of the analysis. Boyd et al. (2002) do consider whether teacher placement responds to the specific grades that "count." They find that exiting from teaching does not appear related to testing regimes. While they have just indirect measures of teacher quality for the New York state sample (experience and quality of college), they do find some attempt in urban schools to place the more experienced teachers in the grades tested when new teachers entered a school.[9]

Student-Gain Approaches

The two variants—cohort gain and individual student gain—produce an average score of student-performance change for a group of students. The distinction between the two in their pure form relates to the group of students included.[10] Student-gain measures allow for the school to isolate school inputs in much the same manner as the grade-level change model above. The superiority of the student-gain model over the grade-level change model lies in its control of student characteristics and in its focus on the level of school performance. Just two states as of fall 2001 (New Mexico and North Carolina) have employed a pure form that examines the same

[9] This evidence is not entirely conclusive about strategic behavior, however. If the grade-level accountability relies just on the levels of achievement in a grade (as most do), schools have an effect that accumulates over time. Thus, getting the effect of a good teacher is possible by placing that teacher in the grade being tested or in a prior grade where students would be better prepared for the material in the tested grade.

[10] In pure form, the largest difference is whether the individual school gets information on the distribution of performance from the individual-gain calculations. An impure form, however, introduces some error in the cohort-gain measure. Specifically, a cohort gain can be calculated by taking the scores in a grade of all students in a school for two years and subtracting the average prior year scores for the previous grade. In this, people who exit between grades are included in the base but not in the current-year score.

cohort of students year over year as they move through a school.[11] The focus on change instead of static performance lends itself to closer association with a school's efforts to improve.

The primary incentives inherent in this approach fall more on improving student scores by improving teaching and programs than for the status model. Exclusions could have an effect on measured performance to the extent that the exclusions eliminate individuals who would have a lower rate of learning. As noted above, however, this impact on the accountability score generally will be considerably less than the impact of exclusions on the status model, because it is only achievement growth and not achievement level that is important.

Since the group of students being examined is constant over time, the model ignores student in-migration. This outcome may interact with district decisions to set school attendance zones and the like—which would eliminate some students from the calculations. To date, no evaluations of the effects of cohort-gain systems on performance are available.

The student-level gain score model follows the progress of individual students and then creates a summary from the net change scores. Of all the models, this approach provides the clearest and strongest incentives for schools to concentrate on the school factors under their control since it minimizes student variation. It enables the fastest and cleanest feedback on any efforts the school undertakes.

With this model, the strength of the incentive will be a function of changes in student-body composition, but the effect will be smaller than for the cohort-change model. Even though

[11] Two aspects of the design of cohort-change systems are important. First, decisions must be made about exclusions of students because of mobility. Based on individual data, it is possible to use initial and subsequent scores for just individuals who start and finish the grade. In general, new entrants during the grade would be excluded from the calculations, but the data would not introduce errors from different groups of students. Second, across each year a decision can be made about whether to update the cohort to the group beginning each grade or whether to maintain the cohort originally identified.

student moves are known to affect scores negatively, as implemented, the school will have students for more than a year before their gain scores are included in the school score.

The model would create the inclination to exclude students who are poor performers. The school will know student-specific performance in the first year of examination and then can follow their progress through the second year, presumably providing information by which to prejudge which students would likely produce negative change scores. By avoiding a second-year test, the gain scores for those students could not be calculated or folded into the school score.

Richards and Sheu (1992) provide an early investigation of the South Carolina incentive system. This system, introduced in 1984, was a sophisticated accountability attempt that considered individual student-gain scores and adjusted rewards for the socioeconomic status of the student body. They find that the reward system yielded gains, although modest, in performance of students (but did not affect teacher attendance, the other attribute of incentive focus). Interestingly, South Carolina subsequently moved away from this incentive system. Ladd (1999) investigates the sophisticated gain-score incentives in Dallas, Texas, during the mid-1990s. She finds that performance in Dallas improved relative to other large Texas districts, although the gains come from white and Hispanic students but not black students. Improvements in terms of student dropout rates and principal turnover rates also appear.

Deere and Strayer (2001a, 2001b) evaluate the impact of Texas incentives on a range of behaviors. They find evidence that schools tend to concentrate on students who are near the passing grade on the TAAS. Moreover, there is some tendency to concentrate on subjects that enter into the accountability system. The evidence also suggests some differential exclusion from testing. They specifically find some sharp increases in overall exemption rates for special

education around the time when these exemptions became most important for accountability. (Note, however, that while the evaluation considers student gains, the Texas incentive system concentrates on overall pass rates.)

In terms of incentives, the objective of rewarding and punishing schools for their contributions to student learning are met in varying degrees by the alternatives. By far the most common alternative—the status model and its grade-level offshoot—provide information that is far distant from the value-added of each school. One aspect of this is the introduction of incentives to change school scores in ways that are unrelated to their learning outcomes. For example, increasing special-education placements or working selectively to decrease test taking can improve scores for a school by changing the rating group. Of course, some alterations work best in the short run—that is, in the year of their introduction—and would be much less effective in later years. The use of these approaches depends on the simple decision-making of administrators and is related to the costs, risks, and time horizons of the administrators.

CUMULATED EVIDENCE ON INCENTIVES

Most accountability systems have been introduced very recently, so the history does not give much scope for analysis. Nonetheless, a variety of investigations have been undertaken recently and provide some, albeit limited, evidence. Table 7 groups these analyses by their focus and by the type of accountability system studied. It seems clear that schools do in fact respond to accountability systems.

Much of the evidence relates to "gaming" the system—actions taken in response to incentives but that are not directly related to improving performance. Thus, as identified in Table 7, several studies indicate that exclusions from the testing tend to increase with the

Table 7
Distribution of Studies of the Impacts of Accountability

Cross-Sectional Accountability Systems	
Outcome effects	
Direct response to consequences	Greene (2001a, 2001b); Jacob (2002); Carnoy and Loeb (2002); Carnoy (2001); Deere and Strayer (2001a, 2001b)
Response to public disclosure	Hanushek and Rivkin (2002); Carnoy (2001)
Measurement errors	
Testing effects	Koretz and Barron (1998); Jacob (2002); Deere and Strayer (2001b)
Random errors	Kane and Staiger (2001)
Exclusions/selectivity	Jacob (2002); Figlio and Getzler (2002); Haney (2000); Cullen and Reback (2002); Toenjes et al. (2000); Carnoy, Loeb, and Smith (2001); Deere and Strayer (2001a, 2001b); Koretz and Barron (1998)
Other responses	
Teacher assignment	Boyd et al. (2002)
Achievement-gain accountability systems	
Outcome effects	
Direct response to consequences	Richards and Sheu (1992); Ladd (1999)

introduction of new accountability systems. None, however, says anything about reactions after the initial response. In most cases, the incentives for these types of reactions will decline over time.

Much less information is available about the range and scope of reactions to improve performance. In most cases studied, the introduction of a performance system has led to achievement improvements. Moreover, the response not surprisingly is more concentrated on the aspects of learning that are measured and assessed as opposed to those that are not. While some people find this to be a negative aspect of the accountability systems, it seems to be just what one would expect. The magnitude of such improvements is nonetheless not easy to characterize. Further, the exact source of the response—whether emanating from the informational aspects of the systems or from the direct sanctions and rewards—is uncertain in states where both mechanisms operate simultaneously.

Important for design considerations, information about the comparative effects of alternative systems is quite limited. Understanding the differences among accountability systems requires comparing states that employ alternative approaches. It is, however, very difficult to do this. For example, Grissmer et al. (2000) interpret estimates of the superior performance of Texas and North Carolina schools on the National Assessment of Education Progress (NAEP) as resulting from their accountability systems, but no attempt is made to test such a hypothesis formally (compare with Hanushek 2001). Carnoy and Loeb (2002) find that accountability systems that have implications for students and schools ("strong accountability") had faster growth in NAEP math achievement. Moreover, this happens not just for low-achievement students but also for high-achievement students. Nonetheless, their categorization cuts accountability systems in different ways than that previously presented. Since a number of states

will soon be adopting new systems as a result of federal legislation, it is important to know which accountability features and designs produce the greatest impact on student performance measures. Specifically, it will become increasingly pertinent to know whether more costly and less understandable systems that focus on value-added measurement are significantly better than status models.

NEW EVIDENCE ON THE IMPACT OF ACCOUNTABILITY

Inferring the impact of accountability systems is difficult both because of the recentness of their introduction in many states and because of the limited information about student performance across different accountability regimes. One source of information on performance, however, offers some possibility for analysis. NAEP has provided performance information for states during the 1990s. These examinations in mathematics track performance across grades. We use these performance measures to assess the impacts of state accountability systems. In this regard, the analysis is directly related to the work of Carnoy and Loeb (2002). It differs largely by looking at longer periods of achievement growth and by employing different measures of accountability. We also investigate whether accountability systems affect special-education placement rates by state.

Impacts on Student Achievement

Understanding the impacts of different state policies on performance is difficult, in part because of the paucity of previous work describing the elements of state policy that are important. Education is the responsibility of state governments, and states have gone in a variety of directions in the regulation, funding, and operation of their schools. As a result, it is difficult

to assess the impacts of individual policies without dealing with the potential impacts of coincidental policy differences.[12]

The basic estimation approach focuses on growth of student achievement across grades. If the impacts of stable state policies enhance or detract from the educational process in a consistent manner across grades, concentrating on achievement growth implicitly allows for stable state policy influences and permits analysis of the introduction of new state accountability policies.

The NAEP testing measured math performance of fourth graders in 1992 and 1996, and of eighth graders four years after each of these assessments. While the students are not matched, the common cohort acts to eliminate a variety of common achievement influences. Our analysis of achievement relies on growth in achievement between fourth and eighth graders over the relevant four-year period (for example, growth in achievement from fourth grade in 1996 to eighth grade in 2000).[13] Understanding the effect of accountability systems is dependent on the introduction of these systems. Table 8 describes the time path of introduction of accountability systems across states by reference to the length of time that accountability systems have been operating in different states. By looking at accountability systems in 1996, it is clear that much of the movement to accountability is very recent. By 1996, just 10 states had active accountability systems, while by 2000, just 13 states had yet to introduce active systems.[14]

The estimation takes two different modeling approaches to understanding the interaction of accountability systems and achievement. First, the two periods of growth between fourth and

[12] Hanushek, Rivkin, and Taylor (1996) discuss the relationship between model specification and the use of aggregate state data. The development here builds on the prior estimation in Hanushek and Somers (2001), and the details of the model specification and estimation can be found there.
[13] We actually rely on differences in logarithms of scores because these implicitly allow state factors to have a multiplicative effect on achievement inputs.
[14] In all analyses, the universe includes 50 states plus the District of Columbia.

Table 8
Distribution of States by Length of Time with Accountability System, 1996 and 2000

Years with an accountability system	1996	2000
0	41	13
1	4	10
2	2	8
3	4	6
4	0	4
5	0	4
6	0	2
7	0	4

Source: CREDO (2002).

eighth grade for the states (1992 through 1996 and 1996 through 2000) are pooled, at times with extraction of state fixed effects. Second, just the latter period is used to look at cross-sectional differences in growth. The former modeling strategy is appropriate if other influences on achievement—both policy and other—are roughly constant over the entire period. The latter concentrates on the period of most activity in accountability but relies on the growth formulation with possible explicit measures of state differences to isolate the effects of accountability systems.

Table 9 presents the basic estimates of the effects of accountability systems on growth in student achievement. The simplest version (columns 1 and 5) looks at whether the state has some form of accountability system in place during the period of observation. Recall that accountability in the United States has taken two general forms—report cards and rewards/sanctions. Report cards serve a public information function whereas rewards and sanctions subject schools to material consequences. The results indicate that the presence of some form of accountability—either report cards or systems with sanctions—produce growth in achievement that is 1 percent higher than it would be without such programs. This is a large effect since the standard deviation of growth in state scores between fourth grade in 1996 and eighth grade in 2000 is just 1.2 percent.[15]

The remaining columns provide additional detail. The second and sixth columns show the implications of having a simple reporting system that either does not have sanctions and rewards or does not summarize the relevant performance of the school. Since reporting systems are less stringent than full accountability systems, one would expect less effect on student achievement growth. Indeed, states with reporting systems achieve about half the growth of

Table 9
Relationship of Presence of Accountability System to Improvements in NAEP Mathematics Performance

	Pooled: 1992-96 and 1996-2000				**1996-2000**		
	(1)	(2)	(3)	(4) With state effects	(5)	(6)	(7)
Accountability or report-card system	0.0084 (3.07)	0.0096 (2.94)	0.0100 (2.56)	0.0089 (1.42)	0.0116 (2.83)	0.0131 (2.96)	0.011 (2.18)
Reporting system		-0.0042 (-1.05)				-0.0057 (-1.25)	
Time system in place			-0.0006 (-0.47)				
Education populated 25-29							0.0006 (0.04)
Real spending per pupil							0.0002 (0.03)

Note: All pooled estimates include an indicator variable for time period. Robust t-statistics are presented below each coefficient. The dependent variable is log ($Achievement_{grade\ 8,\ t}/Achievement_{grade\ 4,\ t-4}$).
Source: Authors' calculations as described in the text.

those with accountability systems (0.42 percent versus 0.96 percent in the pooled sample), although the difference is not statistically significant. Put another way, the results show that the use of sanctions and rewards does not create a significant positive effect over the use of report cards.

With the small number of state observations it is difficult to distinguish between "no effect" and "weak data" such that precise estimation is not possible. Additionally, according to column 3, the time that the system is in effect does not appear to affect performance (that is, achievement growth moves to a higher level once the system is in place but does not continue to improve). The estimate of the overall effect of the use of accountability systems also holds even in the case of state fixed effects (column 4). Finally, while the point estimates are slightly larger when estimated just on the most-recent period of achievement observation, the impact of accountability systems is virtually unchanged from that estimated by pooling the results.[16]

The summary of estimated effects of introducing an accountability system is simple: Accountability systems appear to lead to significantly better growth in achievement. Of course, as discussed above, it would be nice to know more about how variations in the systems employed affect achievement. Unfortunately, the data are rather thin—fewer than 40 states have complete information about achievement growth for the entire period—so it is not possible to say with any certainty whether differences in the accountability systems are important or how important they might be.

---

[15] In all cases the dependent variable is the log of achievement growth. The introduction of an accountability system is a change from 0 to 1, which in the pooled sample corresponds to a proportional increase of 0.008, or roughly 1 percent.

[16] Note that the last column provides estimates of achievement growth where other contemporaneous measures of state differences are included—education level and school spending per pupil of the population aged 25-29 (as a measure of parental education). Neither of these traditional measures of school inputs has an impact on growth in test scores, and the estimates of the effects of accountability are essentially unchanged by their inclusion.

Special-Education Placement

As we discussed, 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 previously discussed literature provides evidence from individual states and school systems suggesting that schools tend to respond in such a manner.

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, when a majority of the accountability systems was introduced. As with achievement analysis, our basic strategy is to relate (logarithms of) special-education placement rates to accountability and other factors that might affect placement. Unlike achievement, however, we have regular measurement of special-education placement, so that we can consider more refined models of the annual patterns in placement. It is also easy in this case to remove state differences in average special education placement (that is, state fixed effects).

Table 10 shows that the introduction of an accountability or report-card system is associated with 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 accountability or report cards, and with an increase of 0.4 percentage point increase each year that the system is in place. Thus, the state estimates appear to confirm the estimates from individual states and districts.

Table 10
Effect of Accountability on Special-Education Placement Rate, 1995 through 2000

	Standard Approach		Allowance for Placement Trend		
Accountability or report card system	1.45 (10.1)	1.09 (7.9)	.11 (1.0)	.10 (.9)	.09 (.7)
Time in place		.38 (7.9)		-.02 (-.5)	
Time trend			.86 (12.4)	.87 (14.4)	.87 (12.5)
Time trend squared			-.08 (-6.3)	-.08 (-6.0)	-.08 (-6.4)
Report card system					.24 (1.2)
Longitudinal system					-.73 (-1.9)

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.
Source: Authors' calculations as described in the text.

The final three columns, however, show a markedly different picture. Specifically, throughout the nation, special-education placement rates have increased over time, and the standard methodology of comparing rates before and after introduction of accountability tends to attribute these overall increases to an effect of accountability systems. Thus, the final columns introduce a time trend and its square to allow for the strong and ubiquitous increases in special-education placement. Columns 4 and 5 show that both the effect of having an accountability 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). The final column introduces the characteristics of the state system. Report card states seem to have a slight positive influence on placement rates. Longitudinal accountability systems (the cohort-change and individual-gain approaches used in several states) lower placement rates, perhaps reflecting regulations on accountability along with the incentives discussed earlier. While neither of these estimates is statistically significant, the impact of longitudinal systems is close to standard levels ($p>0.06$)—even though there are very few observations of such systems.

These estimates suggest caution in interpreting analyses of the gaming of accountability systems. If such gaming were generally important, 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

One of the major conclusions to be drawn from this discussion is that the existing body of evidence about accountability systems is fairly sparse. Moreover, much of it does not help to diagnose the various sources of incentive impacts. Without greater attention across states to

understanding the "signal-to-noise" characteristics of the systems in place, policymakers run the risk of confounding the true effects of their efforts with factors outside of their control.

The analysis provides some simple but powerful messages about state accountability systems. To begin with, on a conceptual level most of the existing systems that have been introduced are not good devices for inferring the quality of individual schools. As a result, they are also not good devices for providing incentives. The incentives do not accurately relate to the activities and performance of the schools, and they are subject to a variety of approaches to "game" the system. These design problems may reflect not having thought out the issues; alternatively they may reflect simple politics that hamstring the introduction of better incentive systems.

The design problems occur in a variety of different forms. Some systems confuse student performance with the inputs and behavior of the schools. Other systems make it difficult if not impossible to separate effects on outcomes that are related to school performance from effects of parents or past educational inputs.

A review of the extant information on how schools react to accountability systems suggests that schools do indeed react to the introduction of accountability systems. At the same time, not all of the reactions appear to be desirable. A variety of investigations of attempts of schools to alter measured achievement without necessarily changing the reality indicates that schools do operate on this margin. Nonetheless, while discovering such unintended consequences is good sport for academics, one would expect the immediate gaming to be much more important than any continual gaming. In other words, this kind of behavior appears largely self-correcting.

Most of the initial investigations also show that the introduction of accountability systems leads states to improve on performance. The confusion with artificial increases through gaming or with responses tailored very specifically to the state testing, however, makes the evidence a little difficult to interpret.

In order to dig more deeply into the effects of accountability systems, we have conducted two new analyses of accountability in the states. We look across the states and investigate whether the introduction of accountability is associated with greater growth in achievement and whether it is associated with more placements into special education. On the first score, we find that achievement growth between the fourth and the eighth grade is 1 percent higher after the introduction of a state accountability system. Further, the difference in impact on achievement between the use of report cards (public disclosure of performance data) and systems that expose schools to direct consequences based on scores are not significant, suggesting that the "power" of accountability lies in reducing barriers to information rather than rewards or punitive measures. The data are not good enough, however, to give us much confidence in whether or not different types of systems have a differential effect.

On the latter score, we find that special-education placement does not appear closely related to the introduction of accountability in a state. Special-education placement rates have increased over time. Once this is allowed for, the introduction of an accountability or report card system has no significant impact on special-education placement, suggesting some caution in interpreting the prior evidence for longitudinal changes within states or districts.

An important element of this analysis is simply setting out some of the features that we believe are most important in thinking about accountability. Specifically, most existing systems—when seen from the perspective of incentives for schools—are seriously flawed. At

the same time, we know that they have an ability to evoke responses from schools. It would be most unfortunate if we lumped all accountability systems together and concluded on the basis of our early observations that they lead to some bad outcomes and thus should be eliminated. This is simply not the message that should be taken from the existing reactions.

If we are interested in student achievement—as we should be—we simply have to focus on student achievement. This is the genius of accountability systems. The perspective should not be whether or not to eliminate accountability but instead how to refine it to provide the kinds of incentives that we want.

Perhaps more important, because accountability is often viewed as a binary choice—you either have it or you don't—it is very likely that some, or even most, of the existing systems will not stand up to expectations. It would be inappropriate, however, to conclude that greater accountability does not work on the basis of results from most existing state systems.

References

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100 (3) June: 598-614.

———. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources* 37 (4) Fall.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114 (2) May: 577-99.

Boyd, Don, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2002. "Do High-Stakes Tests Affect Teachers' Exit and Transfer Decisions? The Case of the 4th Grade Test in New York State." Stanford Graduate School of Education, mimeo.

Brownson, Amanda. 2001. "Appendix B: A Replication of Jay Greene's Voucher Effect Study Using Texas Performance Data." In *School Vouchers: Examining the Evidence*, edited by Martin Carnoy, pp. 41-7, Washington, DC: Economic Policy Institute.

Carnoy, Martin. 2001. *School Vouchers: Examining the Evidence*. Washington, DC: Economic Policy Institute.

Carnoy, Martin and Susanna Loeb. 2002. "Does External Accountability Affect Student Outcomes? A Cross-State Analysis." Stanford Graduate School of Education, mimeo (March).

Carnoy, Martin, Susanna Loeb, and Tiffany L. Smith. 2001. "Do Higher State Test Scores in Texas Make for Better High School Outcomes?" Paper presented at the American Educational Research Association Annual Meeting (April).

CREDO. 2002. "The Future of California's Academic Performance Index." CREDO, Hoover Instiution, Stanford University, mimeo (April).

Cullen, Julie B. and Randall Reback. 2002. "Tinkering Toward Accolades: School Gaming under a Performance Based Accountability System." Department of Economics, University of Michigan, mimeo.

Deere, Donald and Wayne Strayer. 2001a. "Closing the Gap: School Incentives and Minority Test Scores in Texas." Department of Economics, Texas A&M University, mimeo (September).

———. 2001b. "Putting Schools to the Test: School Accountability, Incentives, and Behavior." Private Enterprise Research Center, Texas A&M University, Working Paper No. 113 (March).

Dixit, Avinash. 2002. "Incentives and organizations in the public sector: An interpretative review." *Journal of Human Resources* 37,no.4 (Fall):696-727.

Figlio, David N. and Lawrence S. Getzler. 2002. "Accountability, Ability and Disability: Gaming the System?" University of Florida, mimeo (April).

Figlio, David N. and Maurice E. Lucas. 2000. "What's in a Grade? School Report Cards and House Prices." NBER Working Paper No. 8019 (November).

Gramlich, Edward M. and Patricia P. Koshel. 1975. *Educational Performance Contracting*. Washington, DC: The Brookings Institution.

Greene, Jay P. 2001a. "An Evaluation of the Florida A-Plus Accountability and School Choice Program." New York: Center for Civic Innovation, Manhattan Institute, mimeo (November).

———. 2001b. "The Looming Shadow: Florida Gets Its 'F' Schools to Shape Up." *Education Next* 1 (4) Winter:76-82.

Grissmer, David W., Ann Flanagan, Jennifer Kawata, and Stephanie Williamson. 2000. *Improving Student Achievement: What NAEP State Test Scores Tell Us*. Santa Monica, CA: Rand Corporation.

Haney, Walter. 2000. "The Myth of the Texas Miracle in Education." *Education Policy Analysis Archives* 8 (41).

Hanushek, Eric A. 2001. "Deconstructing RAND." *Education Matters* 1 (1) Spring: 65-70.

Hanushek, Eric A., John F. Kain, and Steve G. Rivkin. 2001. "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools." NBER Working Paper No. 8479 (September).

———. forthcoming. "Inferring Program Effects for Specialized Populations: Does Special Education Raise Achievement for Students with Disabilities?" *Review of Economics and Statistics*.

Hanushek, Eric A. and Margaret E. Raymond. 2001. "The Confusing World of Educational Accountability." *National Tax Journal* 54 (2) June:365-384.

Hanushek, Eric A. and Steven G. Rivkin. 2002. "Does Public School Competition Affect Teacher Quality?" In *The Economics of School Choice*, edited by Caroline M. Hoxby. Chicago, IL: University of Chicago Press.

Hanushek, Eric A., Steven G. Rivkin, and Lori L. Taylor. 1996. "Aggregation and the Estimated Effects of School Resources." *Review of Economics and Statistics* 78 (4) November: 611-627.

Hanushek, Eric A. and Julie A. Somers. 2001. "Schooling, Inequality, and the Impact of Government." In *The Causes and Consequences of Increasing Inequality*, edited by Finis Welch, 169-199. Chicago: University of Chicago Press.

Jacob, Brian A. 2002. "Making the Grade: The Impact of Test-Based Accountability in Schools." Kennedy School of Government, Harvard University, mimeo (April).

Kane, Thomas J. and Douglas O. Staiger. 2001. "Improving School Accountability Measures." NBER Working Paper No. 8156 (March).

Koretz, Daniel M. and Sheila I. Barron. 1998. *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)*. Santa Monica, CA: RAND Corporation.

Ladd, Helen F. 1999. "The Dallas School Accountability and Incentive Program: An Evaluation of the Impacts of Student Outcomes." *Economics of Education Review* 19 (1) February: 1-16.

Ladd, Helen F. and Elizabeth J. Glennie. 2001. "Appendix C: A Replication of Jay Green's Voucher Effect Study Using North Carolina Data." In *School Vouchers: Examining the Evidence*, edited by Martin Carnoy, pp. 49-52. Washington, DC: Economic Policy Institute.

Richards, Craig E. and Tian Ming Sheu. 1992. "The South Carolina School Incentive Reward Program: A Policy Analysis." *Economics of Education Review* 11 (1) March: 71-86.

Rose, Lowell C. and Alec M. Gallup. 2001. "The 33rd Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes toward the Public Schools." *Phi Delta Kappan* (September): 41-58.

Sanders, William L. and Sandra P. Horn. 1994. "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education* 8: 299-311.

Toenjes, Laurence A. and A. Gary Dworkin. 2002. "Are Increasing Test Scores in Texas Really a Myth, or Is Haney's Myth a Myth?" *Education Policy Analysis Archives* 10 (17) March 21.

Toenjes, Laurence, A., A.Gary Dworkin, J. Lorence, and A.N. Hill. 2000. "The Lone Star Gamble: High Stakes Testing, Accountability and Student Achievement in Texas and Houston." Brookings Institution.

Weimer, David L. and Michael J. Wolkoff. 2001. "School Performance and Housing Values: Using Non-Contiguous District and Incorporation Boundaries to Identify School Effects." *National Tax Journal* 54 (2) June: 231-53.