Publication version
August 2002

Lessons about the Design of State Accountability Systems

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

Paper prepared for

Taking Account of Accountability: Assessing Policy and Politics
Harvard University
June 9-11, 2002

Abstract

Test based accountability systems are now a central feature of U.S. education policy. Accountability systems are implemented as a way of improving 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 have developed have very different incentives.

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. This paper characterizes the incentives of different systems and reviews the existing evidence about outcomes.

Lessons about the Design of State Accountability Systems

by Eric A. Hanushek and Margaret E. Raymond

All accountability systems are not alike. They differ in fundamental ways that affect their inherent incentives and potential outcomes. It is easy to conclude that many of the existing systems contain flaws that lead to a variety of undesirable outcomes, particularly in the short run. Yet, it is important that discussion moves past "whether accountability systems are perfect or not." Evidence of flaws should not be taken as general condemnation of accountability systems but instead should lead to focus on how the structure of accountability and reward systems might be improved.

The basic premise of virtually all proposed school accountability systems is that student performance should be the key element. This change, partially forced by federal legislation, will transform the focus of the past when a majority of states provided just rudimentary information about schools in the state, often confined to a few measures of school resources and avoiding any indication of student performance.[1] Even where states have created a hybrid system that combines input and outcome regulatory elements, student outcomes have become a major focus. The appropriate metric for incorporating student outcome information, however, is far from obvious.

The differences across states support a comparative analysis of the structure of the systems and the relationship between structure and performance of the systems over time. Our perspective is that the summary measures of student performance produced by the accountability system are meant to represent the performance of schools and are at least in part intended to introduce incentives for improvement. The question central to this

[1] See the discussion in Hanushek and Raymond (2001)

paper is how different accountability measures reflect the quality and performance of schools and whether we should expect different accountability systems to generate improvements in student outcomes.

The Bottom Line

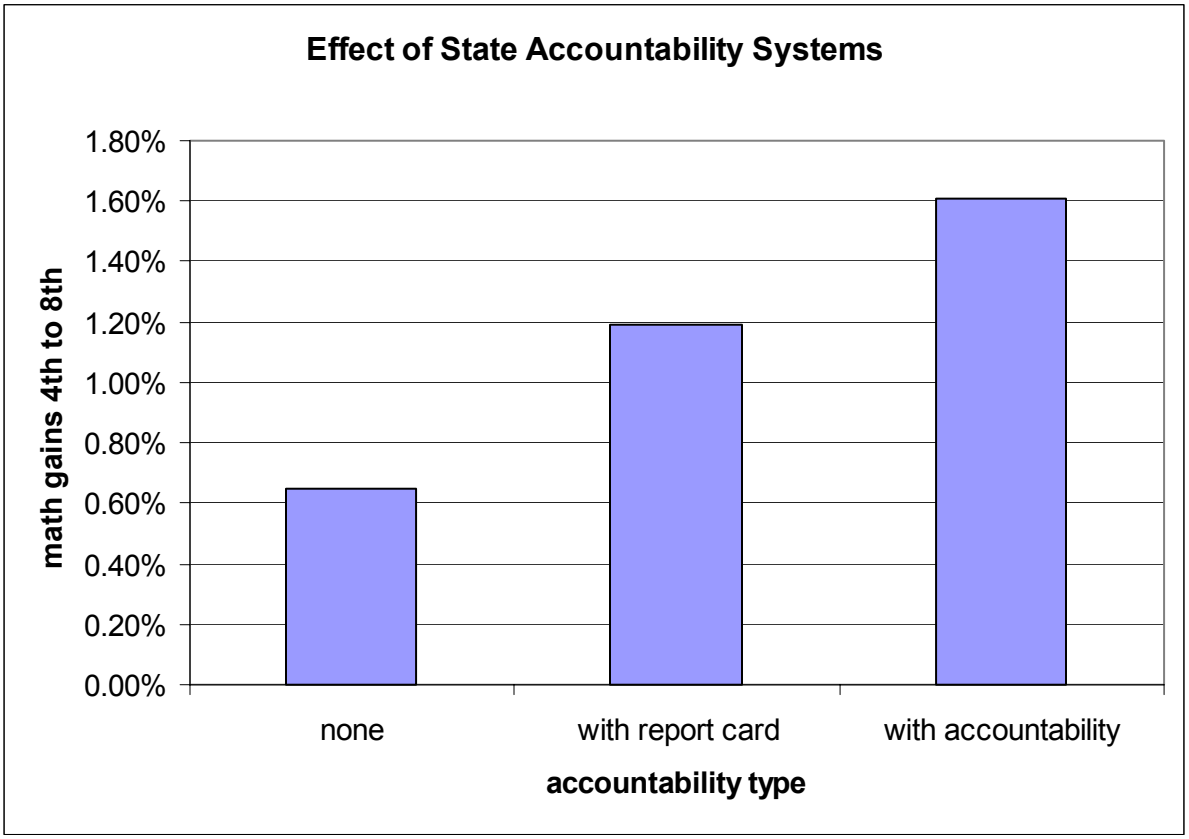
Before analyzing the characteristics of alternative incentive schemes, however, it is useful to motivate the discussion by a quick glimpse of currently observable impacts of state systems. Because the number of states that employed accountability systems changed during the 1990s, it is possible to consider whether schools in accountability states performed differently from those in other states.

Accountability systems can have two kinds of consequences: intended and unintended. In principle, accountability systems alter the incentives faced by schools. In the best of worlds, this would spur states to improve their schools. In the worst, this might induce "bad" before as schools attempted to game the system. For the moment, we concentrate on performance or good aspects.

The impact of existing state systems is illustrated in figure 1, which summarizes our estimates of the gains in mathematics that would be expected between 1996 and 2000 for the typical student who progresses from fourth to eighth grade. These expected gains are calculated from regression analyses of state scores on the National Assessment of Educational Progress (NAEP). The larger empirical work upon which these estimates are drawn separates the potential impact of parental inputs and school spending from the impact of testing and reporting across states.[2] States were classified according to the type of testing and accountability system they had in place at the time of the NAEP test. (A

[2] The details of these estimates can be found in Hanushek and Raymond (2003). The results pool data on NAEP math gains over both the 1992-96 and 1996-2000 period.

Figure 1



Source: Hanushek and Raymond (2003)

state could have its classification change between the two years if it adopted an accountability system). States with "report card" systems display test performance and other factors but neither provide any simple aggregation and judgment of performance nor attach sanctions and rewards. In many ways, these systems serve simply as a public disclosure function. Systems that provide explicit scores for schools and that attach sanctions and rewards are labeled "accountability" systems. The typical student in a state without any formal system would see a 0.7 percent increase in proficiency scores. Reporting systems move the expected gain to 1.2 percent. Finally, states with full accountability systems obtained a 1.6 percent increase in mathematics performance. The performance difference between either reporting systems or accountability systems and no system is statistically significant (although the difference between the two is not). In short, testing and accountability as practiced have led to gains over that expected without formal systems.

This impact makes clear that the issue is not whether to have accountability but how best to have accountability. The focus of this paper is precisely that – what are the alternative approaches to accountability and what do we know about their potential impact?

Alternative Accountability Systems

The key to understanding the informational content (and ultimately incentives) of standard accountability systems is to examine the determinants of student performance and how those determinants are displayed and used. Take the simplest model of student achievement that is consistent with prior work on the determinants of achievement:

$$achievement = school + other.$$

Without getting into controversies about testing at this point, the standard approach is to test students during one or more grades in order to measure achievement. The general idea behind accountability systems is that some aggregation of the tests can be used to assess the contributions of schools – but it is immediately obvious that much will depend on the importance of "other" things.

What is included in *other*? As is well-documented in prior analyses of achievement, many factors outside of the control of schools affect individual student achievement. Students clearly differ in ability, and students get varying input from families and friends. Moreover, measurement of true achievement through common tests is prone to measurement error. Moreover, achievement at any point in time is not determined just by current school and other factors. The historical pattern of each of these also affects the current level of achievement. Thus, we have:

other = *ability* + *family* + *peers* + *history* + *measurement error*.

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 for 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 (e.g., "passing" or "proficient"). These variants introduce important elements into accountability systems, but for now, we consider just the average performance measures. We return to the complications of other kinds of measures later.

Cross-sectional Approaches

The first set of accountability devices begins with the aggregated scores of students in a given year and compares such measures across schools. Virtually all states, whether they provide just report card information or instead develop accountability structures, report average achievement as one of the components of information given. The "status model" simply takes the average performance of students taking the test in a school 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).[3] The first point from this is obvious: If the main purpose of the accountability system is assessing the performance of the *school,* average test score does it very imperfectly. The average achievement will incorporate all of the current and historical inputs to achievement including not only schools but family background and

[3] 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.

random errors included in *other*. 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. That these imperfect scores figure into the determination of sanctions and rewards just adds to the problem.

This basic confusion between average student achievement and the contribution of schools is well known, and most accountability systems introduce additional information to provide context or to show to the impact of schools for particular students. 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 similar in family backgrounds. These approaches still do not allow a very accurate estimation of school performance, because, as suggested by past research, they likely do not adequately identify family differences or cohort differences. Further, even if *family* factors could be adequately accounted for, these adjusted average scores will not capture the prior factors (i.e., *history*) 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, *family*, 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.

Consider an alternative, the "status change" model. In this, the average student achievement of a school is tracked over time. In simplest terms, does the average

performance increase from students in one year to the next? The idea is easiest seen in terms of an example. The status change score for a grade that has a common examination at a specific grade, say third grade reading, is the change in average third grade reading between the 2000 and 2001 school years. The status change model for a school is calculated by aggregating performance across tested grades. For reasons that will soon become obvious, we classify this model as cross-sectional, because it compares snapshots of the school scores across years (as opposed to tracking the performance changes for individual students across years).

The status change model is by far the most common approach to assessing what is happening in schools. The change scores factor heavily in reward systems, but are manipulated in a wide variety of ways: examples include absolute levels of change; percentage increments of change; and change relative to an external standard. Regardless, the most common interpretation is that this 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.

The way to understand this construct is to think of it as providing an estimate of the change across years in value-added of schools ($\Delta school$). It will, however, not be a perfect measure of any school's improvement but will instead contain error. The approach raises two questions. Does an accountability system based 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?

The error in measuring change in school performance goes directly back to the underlying determinants of achievement. The status gain model necessarily compares two different groups of students, only some of whom are common across years. Thus, the status gain has two primary components – the object of interest which is the difference in school quality ($\Delta school$) across the two years and the difference between the two groups of students in family background and other nonschool factors ($\Delta other$). Importantly for some considerations, other differences incorporate any idiosyncratic measurement errors affecting achievement (Δmeasurement *error), and this may have elevated importance*. Just like the status model that relies on the level of average achievement, the status gain 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 are 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.

It might be tempting to argue that local schools in stable communities have similar family inputs and thus $\Delta other$ will be small. But the U.S. population moves a surprisingly large amount. Only 55 percent of students live in the same house for three years in a row, and this falls to half for disadvantaged students.[4] 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 increase noticeably from these national averages. The average annual student mobility across schools in Texas, for example, exceeds 20 percent.

[4] See the overview and analysis of mobility in Hanushek, Kain, and Rivkin (forthcoming)

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

While we have concentrated on school averages, it is common to find these cross-sectional approaches taken to individual grades within a school. The basic motivation for doing this is isolating differential performance by parts of schools. In particular, "grade change" models offer some potential for focusing on school factors when individual cohorts can be tracked over a number of years. Moreover, the use of grade change models become particularly important when passing rates or other distinct elements of the student achievement distribution are highlighted. Nonetheless, these grade approaches still suffer from difficulties in separating *school* and *other* factors.

Longitudinal Approaches

Accountability is quite different when it focuses on the progress of students over time, which we classify as a longitudinal approach. One such approach is the "cohort gain" model. This approach tracks the performance of individual cohorts of students as they progress through school. Consider, for example, comparing the scores of third graders in 2001 with those of fourth graders in 2002. With a stable student body (i.e., 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 grade 3 and grade 4). The cohort gain score would then reflect what the school contributed to

learning in grade 4 plus any differences in idiosyncratic test factors or measurement errors 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 more advantaged, such differences would remain confounded with school factors. The family background and ability factors that affect the cohort gain calculations are, however, ones that affect the rate of growth of learning, not the level. Thus, they would be expected to be relatively small. As a result, the cohort model would generally yield a closer measure of *school* inputs than the status model.

The main concern is how the calculations handle mobility. To the extent that the calculations simply follow the current students in each grade in each year, in and out migration yield the same type of problems discussed previously – the comparisons do not eliminate the differences in nonschool factors across groups.

A number of options for adjusting cohort gains can provide information that is closer to the true impact of schools. One modification simply excludes students entering during the school year from the average achievement calculations. This modification has three advantages for measuring school quality – students who move typically have less learning gain in the year of the move because of the disruption[5]; they have received less than a full dose of the teaching in their current school but part of the teaching in their prior school; and one element of potentially large change in nonschool factors is eliminated. With this modification, the cohort model still compares different groups of students (because those exiting the school between third and fourth grade testing are still included in the earlier achievement calculations but not the second). Moreover, because

[5] Hanushek, Kain, and Rivkin (forthcoming)

mobility is correlated with family backgrounds, the achievement measures are likely to be biased by any differences in student mobility rates across schools. The error would nonetheless be expected to be less than in the no adjustment comparisons.

The influence of mobility suggests an alternative measure for accountability, the "individual gain score" model. This approach improves on cohort change models because it analyzes data at the student level and can include all students with gain scores, not just the students in the original group. If we follow individual students across grades, any historical influences of families and nonschool factors wash out, and the average of individual gains across grades would more closely reflect school quality for the given grade. Nonetheless, it would still incorporate any current influences of family and ability on the growth in achievement and any measurement errors in the separate grade tests.[6]

Refinements and Disaggregations

One obvious fact is that the more aggregated the performance information the less possible it is to pinpoint any causal factors. Thus, for example, accountability models that aggregate all information to the school level (or, worse, the district level) make it difficult to pinpoint the source of any high or low performance. One natural and easy refinement is simply to provide scores for individual grades instead of aggregating these to the entire school level. For example, schools with stable teacher forces could use the grade pattern of cohort gains to unravel the contribution of different groups of teachers. Perhaps the ultimate in this regard is the calculation of teacher value-added as done in

[6] Note that the cohort gain and individual gain models will yield the same results if both school entrants and school leavers are excluded from the cohort and individual gain calculations. The individual gain still nonetheless offers the possibility of further disaggregations by, say, income or entering achievement.

Tennessee.[7] These studies, which are legislatively mandated, provide information to principals and to specific teachers about the student learning gains over time by individual teachers, although the information is not made public.

The validity of the different accountability models for constructing school outcome measures generally relies on a basic stability of underlying nonschool influences and looks at gains in an effort to eliminate the influence of these other factors. An alternative approach is to adjust for outside influences directly.

Consider a situation where there are only two kinds of family influences: good or bad. If we had a measure of these family influences for different students, we could then create a measure of school accountability by simply averaging achievement separately for all students from a good background and all from a bad background. These separate measures would then provide indications of how well a school did with students in the two categories. More generally, it would be possible to expand the calculations to allow for a range of different family backgrounds, including more than two possible levels and including more than a single dimension. States have actively pursued different approaches such as developing indices that rely on weighting different student factors (such as proportion eligible for free or reduced lunch or average education levels of parents) or using statistical approaches (regression analysis) to adjust scores for alternative measures of family background. This is not to imply that some students can't learn, but rather that the pacing may differ, and, for incentive purposes, it is important to separate *school* from *other*. Some adjustments for family background, used in conjunction with individual gain scores, offer perhaps the best chance of isolating the effects of school differences. The individual gain calculations focus the measure on

[7] Sanders and Horn (1994, (1995)

current additions to learning, and the family adjustments eliminate the contemporaneous influence of family factors on the rate of learning by students. As with the simple gains calculations, the effectiveness of these approaches depends on the ability to capture relevant nonschool factors and the ability then to purge the aggregate test scores of things other than school influences.[8] The difficulty in actual application is that normal administrative records typically provide relatively little information about family backgrounds – such as free lunch status and race/ethnicity – and these are crude measures of the relevant family background differences. The paucity of detailed analyses of family effects makes it difficult to assess the impact of alternative specifications and measures of family factors.

Many state systems as described below do not simply report averages of scores, but instead weight the scores against pre-set thresholds to reach judgments about acceptable levels of performance. But this measurement is really no different from the averages in terms of identifying the role of schools. The probability that any given school is above or below any benchmark level for aggregate student performance is directly related to various current and past inputs and to the variance of the random errors – i.e., the *other* factors affecting achievement.

In an insightful paper, Kane and Staiger note that the variance of average measurement error on a test will be inversely related to the number of students tested (by the standard calculations for the variance of a mean).[9] They go on to show empirically

[8] We ignore some of the technical problems in doing the analysis and adjustment. For example, the practice of estimating simple regression analyses based solely on family factors can yield potentially misleading adjustments (Klitgaard and Hall (1975), Grissmer et al. (1994)). Such analysis, which ignores school quality differences, will produce biased estimates of the effects of family background (to the extent that family backgrounds are correlated with school quality differences). These biased estimates will in part incorporate the effects of differences in school quality, the object of the exercise originally.

[9] Kane and Staiger (2002)

how standard calculations of school success in North Carolina lead small schools (with high measurement error variance) to be disproportionately represented among the "successful" schools. Further, if measurement errors over time are uncorrelated, the probability that any school remains as a successful school in subsequent years is very low.

The issues surrounding the variance of test measurement error and its interplay with accountability schemes highlight a set of important trade-offs in designing accountability systems. The first important point is that aggregate achievement scores are error prone measures of school quality because of the error measures of the underlying tests *and* because of the other current and historical factors that are outside of the control of the current school. Thus, viewed from the vantage point of an accountability system for estimating school quality differences, test scores contain both a random component and an error component arising from systematic but unmeasured differences within schools and historical achievement factors. Thus, even if measurement errors could be eliminated, concerns about obtaining unbiased estimates of the effects of schools – the subject of the preceding discussion – would remain.

Clear trade-offs exist. A variety of states are concerned with more than overall performance; they also wish to ensure that high performance reaches distinct subgroups, say by income levels. Quite clearly, as scores are aggregated across smaller groups of students, the variance of measurement error increases and can directly affect rankings of schools depending on how subgroup information is used.[10]

The implications of measurement error depend importantly on the magnitude of such errors and on the magnitude of other factors affecting performance that might bias

[10] Kane and Staiger (2002)

the accountability measure. Kane and Staiger suggest alternative approaches to reducing measurement error. These are most relevant for small schools (say, those with less than 60 students being aggregated into the score). But their recommendations highlight other choices. They propose aggregating test measures over time. In general this will lessen the impact of measurement errors, but it will also bring into play some of the issues surrounding status models unless they can circumvent errors introduced by differences in current and historical factors for cohorts. Specifically, averaging status scores over years does not eliminate the influence of nonschool factors, which bias any estimate of school quality based just on outcomes.

The Distribution of State Accountability Systems

In the summer of 2001, we conducted extensive interviews of the State Education Department in every state about their accountability system.[11] Considerable attention was devoted to the structure of the system, the calculation of school scores, the choice of metrics, and the strength of any consequences that schools faced based on their scores. Recognizing that the practice is evolving and thus is highly fluid, these choices represent a single snapshot of the incentive structures that states chose to provide to their schools.

There are two states that at the time of the survey did not have any measurement or accountability system in place. With the Elementary and Secondary Education Act of 2001, we know these states will adopt some system in short order. Seventeen states had report cards at the school or district level. We found that these states provide information about schools but in a manner that precludes much judgment; for example, a single aspect

[11] Fletcher and Raymond (2002)

of the school is described such as the number of students scoring in the lower quintile, or schools scores are not compared to an independent standard of performance, or the score does not have any consequences associated with it. The remaining 31 states have systems that create a single measure of performance, they have created a scale of judgment about the resulting school scores to determine acceptable and unacceptable results, and they have explicit consequences (sanctions and/or rewards) that schools are exposed to as a result of their score.

The survey of state practices placed states within the four categories described above: Status Model; Status Change Model; Cohort Gain Model; and Individual Gains Score Models. The chief distinction is whether the data are cross-sectional or whether they track student achievement changes over time. Table 1 displays the states with rating systems by the analytic model used to calculate their school scores. The progression from Status to Grade Level Change to Student Change is associated with greater precision in the measures and greater detail about the real impacts of school activity.

The chief information conveyed by these data is the prevalence of using cross-sectional score information. This choice generally precludes sorting out the various components of achievement. Moreover, as we discuss below, this choice tends to increase the incentives for states to manipulate the testing and to attempt to change scores by means other than improving school quality. Specifically, the accounting systems that track student achievement over time improve the incentives for schools, because the results do a better job of explaining the real state of schools without confounding influences mixed in.

Table 1
Classification of States by the Type of Analysis Model
Used in School Rating Systems in 2001

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

Source: Fletcher and Raymond (2002)

Incentives and Evidence on Effects

Accountability systems have an overall influence on schools in two ways: through defining areas of particular attention for schools and through providing rewards or punishments for improving in those areas. We translate the discussion on the different accountability systems into hypothesizes about the incentives introduced and then review the existing evidence. The recent birth of many accountability systems, however, means that the existing evidence is thin in many crucial places.

First, accountability systems focus attention on some details of performance and leave others as irrelevant. A system built solely on test scores, for example, filters out everything except student academic achievement. Similarly, if some subjects are tested and others are not, it is natural to think that attention will go more to the tested areas than the untested areas. Related, part of the debate about testing debate has argued that tests of lower order skills tend to drive out attention of schools to higher order skills.

While these arguments have been discussed quite widely, we know of little empirical work that shows the strength on them. Some general but inconclusive psychometric evidence exists on testing and instruction.[12] More relevant, little work directly links current accountability systems to patterns of time and instruction.

Second, accountability systems increase the exposure of schools in terms of the student performance. Incentives attached to exposure come from two separate mechanisms – indirect pressures and directly legislated rewards and consequences.

Any school will prefer higher scores to lower ones, even if no explicit consequences follow. Currently, apparently in the absence of much clear evidence, most

[12] Dunbar and Witt (1993)

parents appear to think that their school is doing a good job.[13] The provision of accountability evidence has the potential for changing this, perhaps sufficiently 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 impacts) to quality information exists.[14] 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[15] and other observable outcomes.

The second source of incentives arises from any consequences that might be directly associated with the school scores. The rewards and sanctions built into many state accountability systems motivate schools to change behavior. At the same time, one does not expect these incentives to affect all school equally. For example, schools that have scores close to a threshold might be expected to alter their behavior more than schools further away from the established critical thresholds. The interrelationship between the choice of 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.

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.

[13] Rose and Gallup (2001)
[14] Black (1999); Weimer and Wolkoff (2001)
[15] Figlio and Lucas (2000)

Cross Sectional Approaches

Both the status and the status change model confuse the school's influence with other factors. Schools can respond in two ways. First, it can adjust teachers, curriculum, and program in an attempt to improve the teaching and learning that occur. This is, however, a difficult long run proposition. 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.

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, i.e., 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, the school might exclude this slow student through, say, placement in special education or counseling the student to be absent on the day of testing. If excluded, the average of all remaining students would be higher than otherwise, and the school will tend to look better in comparison to the third grade in the prior year.

But, consider the dynamics. The next year comparison of third grades will be worse because the base comparison has been artificially elevated. Moreover, once excluded, there is a continuing incentive to keep the student out of the testing if subsequent grades are also involved in the accountability system. 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.

The largest effects of exclusion on the school ratings come in the first year of exclusion (when the cumulative effect of low preparation plus slow learning are removed). Nonetheless, there are some continued accountability benefits to the school from exclusion if the students learn at a slower pace. The status model typically 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 in the prior example. 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 do not appear 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 only improve 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 "over use" exclusions in the first years of an accountability system. Regulatory restrictions are frequently designed in an effort to limit the ability of administrators to increase the use of student exclusions.

A grade level change version of accountability is used when testing does not cover all grades. If, for example, testing only is done at the fourth grade, the accountability system would feature just that grade. This possibility introduces some additional incentives. Some of the dynamics of exclusions are altered. But also there may be incentives to concentrate attention on the tested grades(s), say by placing the best teachers in the relevant testing grades.

Several studies have investigated whether schools appear to react to accountability through exclusions. Jacob considers the introduction of test-based accountability for Chicago public schools.[16] 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 and Cullen also find apparent increases in special education placement with the introduction of

[16] Jacob (2002)

accountability in Texas.[17] Prior work in Kentucky by Koretz suggested no strategic use of grade retentions.[18] Haney suggests that both grade retention and increased dropouts were key to improvements in Texas tests, although both Carnoy, Loeb, and Smith and Toenjes seriously question this after reanalysis of the data.[19] 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 concentrate on special education placement after the introduction of a state accountability system in Florida.[20] 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.[21] This evidence is consistent with analysis in Texas by Deere and Strayer.[22] 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

[17] Deere and Strayer (2001a, (2001b); Cullen and Reback (2002)
[18] Koretz and Barron (1998)
[19] Haney (2000); Carnoy, Loeb, and Smith (2001); and Toenjes and Dworkin (2002) Carnoy, Loeb, and Smith (2001) also find that at least in larger urban areas lower dropout rates are associated with higher student achievement.
[20] Figlio and Getzler (2002)
[21] Jacob, "Making the Grade: The Impact of Test-Based Accountability in Schools."
[22] Deere and Strayer (2001b)

placement rates and the like through 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, perhaps compared to trends of other schools (such as schools outside of Chicago and its accountability system). 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 know about the long run dynamics.

Our own national work creates questions about the importance of such exclusions. We consider the pattern of special education placement rates across states from 1996-2001.[23] If we simply consider the introduction of accountability systems or the length of time with accountability systems in each state, we find a significant and positive relationship on special education placement. But, if include an overall time trend to allow for the national increases in special education over this time period, the increased use of exclusions disappears.

Hanushek and Rivkin investigate the impacts of public disclosure of achievement performance.[24] 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

[23] See Hanushek and Raymond (2003) for details. All statistical models include separate state fixed effects to allow for different base propensities to classify students in special education programs. Different variants also include overall school spending patterns.
[24]Hanushek and Rivkin (2003)

the public. They find that in the largest metropolitan area, competition works to push up average scores.

Greene analyzes the Florida A+ program that provides exit vouchers to students in failing schools and finds that schools close to being subject to vouchers make unusually large gains.[25] Carnoy reviews this evidence and suggests that the reaction to vouchers that Greene identified was more likely a reaction to information.[26] Carnoy finds that similar studies in North Carolina and Texas investigating what happens to failing schools show similar results – dramatic improvements in the year after identification.[27] This occurs even though those states had no voucher threat.

On the other hand, Kane and Staiger suggest that a portion of the school improvement in North Carolina failing schools may simply result from measurement errors in the examinations.[28] 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-order of schools in the second year. The researchers do not, however, consider all the potential 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 employed differences by grade level primarily as a method of identifying the behavioral effects of the system (comparing a grade included in the

[25] Greene (2001a, (2001b)
[26] Carnoy (2001)
[27] Ladd and Glennie (2001) and Brownson (2001), respectively.
[28] Kane and Staiger (2002)

accountability system to one not included) as opposed to being a focal point of the analysis. Boyd, Lankford, Loeb, and Wyckoff do consider whether teacher placement responds to the specific grades that "count".[29] They find that exiting from teaching does not appear related to testing regimes. While they have just indirect measures of 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.[30]

Longitudinal Approaches

While cohort gain models are more effective at isolating the school's contribution to performance, they have been implemented in just two states as of Fall 2001 (New Mexico and North Carolina). Unlike the status model, the primary incentives in these approaches are toon improve student scores by improving teaching and programs. 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 will generally 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. In purest form, the group of students being examined is constant over time, student in-migration is ignored, potentially interacting with district decisions to set school attendance zones and the like. Nonetheless, to date, no evaluations of the effects of cohort gain systems on performance

[29] Boyd et al. (2002)

[30] 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 all 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.

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.[31] Since additionally it focuses on progress, the model can isolate the contribution of individual teachers, although no state makes such information public.[32]

The model provides an inclination and an ability 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 pre-judge which students would likely produce negative change scores. By avoiding a second year test, the gain scores for those students cannot be calculated or folded in to the school score for two years (i.e. not as the change year nor as the base for the following year).

Richards and Sheu provide an early investigation of the South Carolina incentive system.[33] This system, introduced in 1984, was a sophisticated accountability attempt that considered individual student gain scores and adjusted rewards for the SES 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 investigates the sophisticated gain score incentives in Dallas

[31] 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).
[32] Tennessee produces measures of individual student value-added, but it is not publicly released (Sanders and Horn (1994)).
[33] Richards and Sheu (1992)

during the mid1990s.[34] She finds that performance in Dallas improves relative to other large Texas districts, although the gains come for white and Hispanic students but not black students. Improvements in terms of student dropout rates and on principal turnover also appear.

Deere and Strayer evaluate the impact of Texas incentives on a range of behaviors.[35] They find evidence that schools tend to concentrate on students who are near the passing grade on the Texas Assessment of Academic Skills (TAAS) and 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).

Summary of Evidence

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. Table 2 summarizes the hypotheses about the kinds of effects that might be expected under different accountability regimes. Importantly, we have strong evidence about very few of these components, particularly for the longitudinal accountability schemes. The boldface hypotheses in Table 2 indicate areas where we have no systematic evidence. Most accountability systems have been introduced very recently, so the history does not give much scope for analysis. The prevalence of bold in the table, particularly about any long run effects, is unfortunate.

[34] Ladd (1999)
[35] Deere and Strayer (2001a, (2001b).

Table 2. Hypothesized Impacts of Accountability

(BOLD implies no existing evidence on hypothesis)

	Cross sectional accountability systems	Achievement gain accountability systems
Outcome effects		
Direct response to consequences	SR: muted positive school quality improvements that might be overpowered by other reactions **LR: increasing pressures to improve quality than SR**	Stronger impact on outcomes than cross-sectional, especially in SR but also in LR
Response to public disclosure	Same pattern as effects to direct consequences but less strong	**Same pattern as effects to direct consequences but less strong**
Measurement errors		
Testing effects	Movement toward areas in accountability measure	**Movement toward areas in accountability measure**
Random errors	**May lessen incentives for quality improvement**	**May lessen incentives for quality improvement; comparison to cross-sectional systems unclear**
Exclusions/selectivity		
	SR: large incentives to adjust tested population **LR; considerably dampened incentives to alter population**	**SR & LR: relatively modest incentives to alter population, similar to long run in cross sectional systems**
Other responses		
Teacher decisions/assignment	Higher exit rates of teachers (and principals) with accountability systems	**Higher exit rates of teachers (and principals) with accountability systems**

Note: SR = short run or immediate responses; LR=long run responses.

The clearest story is simply that schools do in fact respond to accountability systems. When introduced, schools appear from the outcomes that are observed to react to the varying incentives.

The most common accountability alternative chosen by states – the status model and its grade level offshoot – provides 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. The largest volume of evidence actually relates to "gaming" the system – actions taken in response to incentives but actions that are not directly related to improving performance. Thus, several studies indicate that exclusions from the testing by individual states and districts tend to increase with the introduction of new accountability systems. None, however, say anything about reactions after the initial response. This is unfortunate since most such actions work best in the short run, i.e., in the year of their introduction, and would be much less effective in later years. In most cases, the incentives for these types of reactions will decline over time. Moreover, the aggregate picture from looking across states in recent years suggests that general increases in special education placement are much more important than specific state reactions to accountability. The impact of special education exclusions does not show up at the state level.

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 does in fact lead to achievement improvements. Moreover, the responses not surprisingly appear 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 nature of the response – whether emanating from the informational aspects of the systems or from the direct sanctions and rewards – is uncertain.

Our generalization to overall state performance on NAEP suggest that accountability improves learning, not just responses to specific tests. The validity and reliability of NAEP, often called the "nation's report card," are well accepted. It is a test for which students cannot easily be prepped and, since the performance of individual school districts, schools, or students is not reported, there is little incentive to cheat or even to prepare for the test. Since the test was adopted before the advent of state accountability systems, it also provides a neutral standard for assessing the effects of state policies. Thus, improvement there reflects more general learning, not just responses to the specific state testing instruments.

Importantly for design considerations there is little information about the comparative effects of the alternative systems. 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. 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.[36] Carnoy and Loeb find that accountability systems that have implications for students and schools ("strong accountability") had faster growth in NAEP math achievement.[37] Moreover, this happens

[36] Grissmer et al. (2000) For an analysis of this conclusion, see Hanushek (2001)
[37] Carnoy and Loeb (2002)

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. Our work reported previously finds differences between accountability and report card systems, although the difference is not statistically significant. It is not possible to distinguish in this between weak data from the limited number of state observations and true design impacts. Finally, using a different coding for states employing "high stakes" tests, we again confirm significant differences in NAEP performance across states that are related to accountability.[38]

Since a number of states will soon be adopting new systems as a result of federal legislation, it is important to know, say, whether more costly and less understandable systems that focus on value-added measurement are significantly better than status models. The bold items in Table 2 are simply central to the most significant design issues for state accountability systems.

Implications for Policy and Research

A prime implication of this review is that more extensive and focused analysis is needed before we can make many strong statements about the effectiveness of accountability for raising student performance. While the accountability movement appears to hold significant promise for improvement of schools, its potential has yet to be fully realized. Part of this is a simple reflection of the newness of most state accountability systems.

But part of the uncertainty results from the particular forms of accountability systems that have been adopted. The vast majority of existing systems use performance

[38] Raymond and Hanushek (2003)

measures that confuse changes in school performance with other factors that the school does not control – families, student abilities, neighborhood effects, and simple measurement errors.

Aspects of the confusion have been explored, but our current knowledge is skewed. Even though the theoretical discussion above indicates that student gain models provide superior precision to cross-sectional models, they remain largely unexplored. Moreover, much of the work to date on cross-sectional models has been useful in identifying unintended consequences or edge cases, but these aspects are likely to be addressed through refinements over time. Further, we can expect most of these incentives to die out naturally over time. It is the central features of the systems that will eventually be most relevant, and much opportunity remains to fully explore their impact. It will be necessary to fill in with additional studies before we can fully judge these systems as a general policy or to know whether any success achieved by some approaches is generalizable to others.

The degree of precision in these systems directly affects the strength and clarity of the incentives they create. In addition to knowing if accountability systems create better outcomes, it is also important to learn more about the manner in which schools react. At present, most of these proposed mechanisms for how schools respond are unexplored.

References

Black, Sandra E. 1999. "Do better schools matter? Parental valuation of elementary education." *Quarterly Journal of Economics* 114,no.2 (May):577-599.

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. Washington, DC: Economic Policy Institute:41-47.

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." *Educational Evaluation and Policy Analysis* 24,no.4 (Winter):305-331.

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).

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." Working Paper 113, Private Enterprise Research Center, Texas A&M University (March 2001).

Dunbar, Stephen B., and Elizabeth A. Witt. 1993. "Design innovations in measuring mathematics achievement." In *Measuring what counts: A conceptual guide for mathematics assessment*, edited by National Research Council. Washington, DC: National Academy Press:175-200.

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

Figlio, David N., and Maurice E. Lucas. 2000. "What's in a grade? School report cards and house prices." Working Paper W8019, National Bureau of Economic Research (November).

Fletcher, Stephen H, and Margaret E. Raymond. 2002. "The future of California's academic performance index." CREDO, Hoover Instiution, Stanford University (April).

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

———. 2001b. "The looming shadow: Florida gets its 'F' schools to shape up." *Education Next* 1,no.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.

Grissmer, David W., Sheila Nataraj Kirby, Mark Berends, and Stephanie Williamson. 1994. *Student achievement and the changing American family*. Santa Monica, CA: Rand Corporation.

Haney, Walter. 2000. "The myth of the Texas miracle in education." *Education Policy Analysis Archives* 8,no.41 (August).

Hanushek, Eric A. 2001. "Deconstructing RAND." *Education Matters* 1,no.1 (Spring):65-70.

Hanushek, Eric A., John F. Kain, and Steve G. Rivkin. forthcoming. "Disruption versus Tiebout improvement: The costs and benefits of switching schools." *Journal of Public Economics*.

Hanushek, Eric A., and Margaret E. Raymond. 2001. "The confusing world of educational accountability." *National Tax Journal* 54,no.2 (June):365-384.

———. 2003. "Improving Educational Quality: How Best to Evaluate Our Schools?" In *Education in the 21st Century: Meeting the Challenges of a Changing World*, edited by Yolanda Kodrzycki. Boston, MA: Federal Reserve Bank of Boston.

Hanushek, Eric A., and Steven G. Rivkin. 2003. "Does public school competition affect teacher quality?" In *The Economics of School Choice*, edited by Caroline M. Hoxby. Chicago, IL: University of Chicago Press:23-47.

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. 2002. "Volatility in school test scores: Implications for test-based accountability systems." In *Brookings Papers on Education Policy 2002*, edited by Diane Ravitch. Washington, DC: Brookings:235-269.

Klitgaard, Robert E., and George R. Hall. 1975. "Are there unusually effective schools?" *Journal of Human Resources* 10,no.1 (Winter):90-106.

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,no.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. Washington, DC: Economic Policy Institute:49-52.

Raymond, Margaret E., and Eric A. Hanushek. 2003. "High-Stakes Research." *Education Next* 3,no.3 (Summer):48-55.

Richards, Craig E., and Tian Ming Sheu. 1992. "The South Carolina school incentive reward program: A policy analysis." *Economics of Education Review* 11,no.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.

———. 1995. "The Tennessee Value-Added Assessment System (TVAA): Mixed model methodology in educational assessment." In *Teacher evaluation: Guide to effective practice*, edited by Anthony J. Shinkfield and Daniel L. Stufflebeam. Boston: Kluwer Academic Publishers:337-376.

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,no.17 (March).

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,no.2 (June):231-253.