Lessons and Limits of State Accountability Systems

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

PEPG/02-13

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 and Limits of State Accountability Systems
by Eric A. Hanushek and Margaret E. Raymond

The desire to have accountability in the schools has spread across the nation, leading to a new attention to the performance of students. In response to the few states that resisted their own development, the federal government has moved to ensure a minimal level of testing and reporting that only a decade ago would have been unthinkable. Yet for all of the attention to accountability even the underlying concepts, to say nothing of effectiveness, are surprisingly vague and ill thought out.

All accountability systems are not alike. They differ in fundamental ways that affect the incentives and the potential outcomes. While there has been some discussion of the tests that are used, of details of the testing and reporting requirements, and of potential rewards and sanctions that are attached, there has been much less attention to the overall structure of the systems.

It is important that discussion moves past “whether accountability systems are good or not.” Many of the existing systems involve a variety of flaws that lead to some apparently undesirable outcomes, particularly in the short run. Such evidence should not, however, 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.

This paper begins with a discussion of fundamental alternatives in accountability systems along with a description of the current array of systems used by states. Following that, there is a discussion of the incentives created by the different types of
systems and the evidence about behavioral reactions to these incentives. It concludes with some observations about policy and research implications.

**Testing and accountability**

The basic premise of virtually all proposed school accountability systems is that student performance should be a 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 (Hanushek and Raymond (2001)). Even where states have created a hybrid system that combines input and outcome regulatory elements, student outcomes have become a major focus. The appropriate use of student outcome information, however, is far from obvious.

The basic accountability structure begins with testing some subset of students in a state using a common examination. We concentrate here entirely on testing and reporting regimes that go to the individual school level. The accountability system then presents information about aggregate student performance and sometimes incorporates additional information about the students or the school.

The summary information about student performance is used by a majority of states to form an accountability system whose ultimate purpose is the improvement of school performance and student outcomes. The path to improvement from state

---

1 While new reporting regimes are moving to the school level and even disaggregations within the school (say by race or income), many traditional accountability systems have provided information just at the district level.

2 As described below, we distinguish between accountability systems that provide information about acceptable levels of performance, expected outcomes, and the like from pure reporting systems that might present school or district information on achievement without any information on how to judge this performance.
accountability systems involves incentives to schools, either explicit or implicit. Our perspective here is that the summary measures produced by the accountability system are meant to be a reflection of the performance of the school and that the schools are led to improve these measures, i.e., that the incentives relate directly to these accountability measures. The question here is how different accountability measures relate to the quality and performance of schools and whether we should expect different accountability systems to generate improvements in student outcomes. We begin with a simple conceptual structure so that we can be precise about the measures and their interpretations. Subsequent discussions carry this framework into a more explicit discussion of the incentives generated by different systems along with a review of the existing evidence about effects.

The key to understanding the informational content of the reporting is to examine the determinants of student performance and how those determinants are displayed within the accountability system. Take the simplest model of student achievement that is consistent with prior work on the determinants of achievement. At any grade (G), the achievement of an individual student (i) will be related to school inputs (δ^s^G^t^) and to nonschool inputs including family background, friends, individual ability and the like (X). In this simplified presentation, we assume that the impact of a given school (s) is the same for all students in a grade (G) during a specific school year (t). Over time, improvement in school performance – coming from hiring better teachers, improving the curriculum for a grade or whatever – would be captured by increases in δ^s^G^t^ between two

\[ s \]

\[ G \]

\[ t \]

\[ \delta \]

\[ ^s^G^t \]

\[ X \]

\[ ^s^G^t \]

\[ \delta \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]

\[ ^s^G^t \]
school years. Moreover, our measurement of individual achievement may contain error from several sources: these include errors from the particular test instrument used; to idiosyncrasies of the testing environment to random factors impinging on the measurement at any point of time ($\varepsilon$). Finally, what a student knows at the end of any grade depends not only on what went on during that grade but also on prior learning and the achievement in earlier grades. Thus, denoting past inputs by a tilde, symbolically we can write achievement at the end of any grade $G$ as:

$$A_{iG} = \bar{X}_{iG} + \delta_{iG} + \varepsilon_{iG} + \tilde{X}_{iG} + \tilde{\delta}_{iG} + \tilde{\varepsilon}_{iG}$$

(1)

where $\bar{X}_{iG} = \sum_{g=0}^{G-1} X_{ig}$, $\tilde{\delta}_{iG} = \sum_{g=0}^{G-1} \delta_{ig}$, and $\tilde{\varepsilon}_{iG} = \sum_{g=0}^{G-1} \varepsilon_{ig}$.

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 average test scores of the students in a grade or an entire school ($\bar{A}$), 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.

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. Average test performance is simply the average performance
across all of the students taking the test in a school in any given year, which can be seen through adding up the average of the various current and past inputs displayed in equation 1. The status model simply takes 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). The first point from this is obvious: If the main purpose of the accountability system is assessing the performance of the school (i.e., $\delta^s$), $\overline{A}^s$ 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 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. Certainly, to the extent that schools chiefly control their own contribution ($\delta^s$), the current achievement level only very imprecisely measures the contribution of schools. That these imperfect scores are used for 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 provide additional information to adjust these scores or 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.

---

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.
judged similar in family backgrounds. These approaches still do not allow a very accurate estimation of school performance, because they likely do not adequately identify family differences or cohort differences and they do not capture prior factors (the tilde’s in equation 1) 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, X, 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. We consider common alternative approaches to supplementing this status model that have been introduced into accountability systems.

Perhaps the most frequently considered added measure is described in 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 change in performance across time for a grade in the school is simply:

\[
\Delta \overline{A}_{Gi} = \overline{A}_{Gi} - \overline{A}_{Gi-1}
\]

where \(t\) and \(t-1\) refer to successive students enrolled in a school. The idea is easiest seen in terms of an example. The status change score for a school and grade that has a common examination at a specific grade, say third grade reading, is something like the change in average third grade reading between the 2000 and 2001 school years. The status change model can also be 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. The change scores factor heavily in reward systems, but are treated in a wide variety of ways: examples include absolute levels of change; percentage increments of change; and change relative to an external standard. Nonetheless, 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 in value-added of schools \( \Delta \delta^{s}_{Gt} = \delta^{s}_{Gt} - \delta^{s}_{Gt-1} \). It will, however, not be a perfect measure of any school’s improvement but will instead contain error. The key to evaluating this in an accountability framework is considering the properties of the error in estimating school quality. Does the accountability system 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?

We can understand the characteristics of the estimator of school quality in the status model by referring back to equation 1, which describes what leads to the achievement of students.

\[
\Delta \overline{A}^{s}_{Gt} = \overline{A}^{s}_{Gt} - \overline{A}^{s}_{Gt-1} = \Delta \delta^{s}_{Gt} + v^{s}_{Gt}
\]

In this, the estimate of the change in the school’s contribution \( \Delta \delta^{s}_{Gt} \) provided by the accountability system for grade G of school s in year t is the true value plus an error in estimating the true contribution \( v^{s}_{Gt} \). Understanding the composition of this error is the essential element for understanding the accountability system.
Even if the student body is identical across years, the status gain model is still comparing two different groups of students. Thus, the status gain has three primary components – the difference in school quality ($\delta$) 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 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.

The situation is, however, even worse than described. A key element of consideration is the dynamics of student populations. The mobility of the U.S. population has important implications for the schools – not only for the way they teach students but also for the accountability systems.

The U.S. population moves a surprisingly large amount. Table 1 presents nationwide average mobility rates (changes of residence) for elementary students over a three-year period. Only 55 percent of students live in the same house for all three years of observation, 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 increase noticeably. The average annual student mobility across schools in Texas, for example, exceeds 20 percent (Hanushek, Kain, and Rivkin (2001)).
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 differences 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.

By following the progress of students over time, it is possible to slightly modify the approach and gain substantial accuracy in the results. Student change models capture student achievement and progress in a manner that preserves a greater amount of detail about a school and its students. These models isolate the effects of changes in student body composition, teacher differences and instructional design shifts across grades; the results describe more clearly the performance of schools. Because the unit of measure is improvement, it can be applied to students regardless of the actual level of initial performance. One such approach is the cohort gain model. It tracks the performance of individual cohorts of students as they progress through school:

\[ \Delta \bar{A}_{Gt}^s = \bar{A}_{Gt}^s - \bar{A}_{G-1,t-1}^s \]

The scores of students in a school are matched year to year and the extent of improvement or loss is calculated. With a stable student body (i.e., with no in or out migration for the school), the historical school and nonschool factors (S, X) would cancel out (because they influence a cohort’s performance both in grade G and in the prior grade, G-1). The cohort gain score would then reflect what the school contributed to learning in grade G (i.e., \( \delta \)) plus any differences in idiosyncratic test factors (\( \varepsilon \)) 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. Nonetheless, the cohort model would generally yield a closer measure of $\delta$ 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.

The main concern is how the calculations handle mobility. To the extent that the calculations simply follow the 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.

One modification of the cohort model generally improves its measurement. Students entering during the school year are frequently excluded from the 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 (Hanushek, Kain, and Rivkin (2001)); 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. Even with this modification, the cohort model still compares different groups of students (because those exiting the school between G-1 and G are still included in the first achievement calculations). Moreover, because mobility is correlated with family backgrounds, the achievement measures are likely to be biased by any differences in student mobility rates across schools.
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.⁵

*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 make it difficult to pinpoint the source of any high or low performance.⁶ 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 Tennessee (Sanders and Horn (1994, (1995). 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.

---

⁵ 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.

⁶ Even less useful are systems that aggregate information to the district level. Current movement is almost exclusively toward providing information at the school building level.
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 actually 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.

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 current additions to learning, and the family adjustments eliminate the contemporaneous influence of family factors.

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.\textsuperscript{7} 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 use simple averages of scores, but instead transform scores to reflect judgments about acceptable levels of performance. For example, it is possible to calculate the proportion of students that achieve above some level deemed to be proficient or passing. In any grade, the probability that a student exceeds the cutoff for the test is a function not only of past learning (i.e., $\bar{X}$, $\vec{\delta}$, and $\vec{\varepsilon}$) but also of current circumstances plus current measurement errors. The scores of individual students are then aggregated to the school.

The combination of the different factors are perhaps easiest to see in terms of a state accountability system that reports the percentage of students over some specific level of performance, say A*. Then, from equation 1, the proportion of students above the cutoff is going to be probabilistically related not only to current school quality ($\delta^*$) but also to $\bar{X}$, $\vec{\epsilon}$, $\vec{X}$, $\vec{\delta}$, and $\vec{\varepsilon}$ (current and past average levels of family background and measurement error plus average past school quality for students currently in the school. The probability that any given school is above or below any benchmark level for

\textsuperscript{7} 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.
aggregate student performance is directly related to various current and past inputs and to
the variance of the random errors.

In an insightful paper, Kane and Staiger (2001) 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). They go on to show empirically
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 $\varepsilon$ and $\bar{\varepsilon}$ 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 ($\varepsilon$) and an error component arising from systematic but unmeasured
differences ($\bar{X}$) and the historical achievement factors ($\bar{X}$, $\delta$, and $\bar{\varepsilon}$). 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.

Additionally, 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 Kane and Staiger (2001)).

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 the accountability measure. Kane and Staiger (2001) 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 some of the clear trade-offs that exist. They propose aggregating test measures over time. In general this will 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.

More importantly, different cohorts involve different elements of school quality. To the extent the accountability system provides incentives to improve on school performance, one expects school quality ($\delta^*$) to change over time such that averaging across years misses key elements of school performance changes and distorts incentives. Thus, the best balance between annual measurement and correcting for measurement is not easily described.
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 (CREDO (2002)). 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 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 essentially cross-sectional or whether they track student achievement changes over time. Table 2 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.

**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 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 rather thin in many crucial places. Indeed, the thinness of the evidence is one of the main points of this analysis.
First, accountability systems impose filters onto schools that focus attention on some details of performance and leave others as irrelevant. A system built solely on test scores filters out everything except student academic achievement whereas a system that is built on multiple factors would filter out less activity but tend to weaken any achievement motivation. 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. Specifically, does the existence of specific kinds of accountability lead to changes in the scope and focus of school operations? Is this a causal influence? Some general but inconclusive psychometric evidence exists on testing and instruction Dunbar and Witt (1993). More relevant, little work directly links current accountability systems to patterns of time and instruction.

Second, accountability systems are designed to increase exposure of schools in terms of the quality of 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 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 (cite). 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 (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 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.

*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 is generally 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.

Schools can respond in two ways. First, it can adjust teachers, curriculum, and program 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, i.e., falls further behind over time. The status model compares performance of individual classes each year to the prior years 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, say, through placement in special education or counseling the student to be absent on the day of testing. If excluded in the third grade, 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, 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. 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 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 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 variant of the status model considers individual grades, instead of aggregate school performance, over time. 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 begins to provide the ability to distinguish between school inputs and student variation. The student effect from 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.8

The grade level change version of accountability also is used when testing does not cover the range of grades. If, for example, testing only is done at the fourth grade, the accountability system would feature just that grade. This possibility introduces some

8 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.
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.

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 in Kentucky (Koretz and Barron (1998)) suggested 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 (2001) 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 (2002) 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 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.

Hanushek and Rivkin (forthcoming) 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 close to being 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 in 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 failing schools may simply result from measurement
errors in the examinations. 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. They 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 employed 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 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*

This approach produces an average score of student performance change for a
group of students. In doing so, it allows for the school to isolate school inputs in much
the same manner as the grade level change model above. Its superiority over the grade
level change model lies in its control of student characteristics and in its focus on the
level of school performance. While implemented it just two states as of Fall 2001 (New
Mexico and North Carolina), it is has 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 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.
students year over year as they move through a school. 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 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.

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. 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

---

10 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.
focuses on progress, the model can isolate the contribution of individual teachers, although no state makes such information public. Of all the models, this approach provides the clearest and strongest incentives for schools to concentrate on the school factors under their control. It enables the fastest and cleanest feedback on any efforts the school undertakes.

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

The model increases 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 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.

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 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

---

11 Tennessee produces measures of individual student value-added, but it is not publicly released (Sanders and Horn (1994)).
incentive system. Ladd (1999) investigates the sophisticated gain score incentives in Dallas during the mid-1990s. 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 (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 Texas Assessment of Academic Skills (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).

**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. 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 to selectively decrease test taking can improve scores for a school by changing the rating group. Of course, some alterations work best in the short run, i.e., 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.

Most accountability systems have been introduced very recently, so the history does not give much scope for analysis. There have, nonetheless, been a variety of investigations that have been recently undertaken. These analyses, summarized in Table 3, provide some albeit limited evidence about the effects of accountability systems.

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.

Much of the evidence relates to “gaming” the system – actions taken in response to incentives but actions that are not directly related to improving performance. Thus, as identified in Table 3, several studies indicate that exclusions from the testing tend to increase with the introduction of new accountability systems. None, however, say 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 will in fact lead 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 nature of
the response – whether emanating from the informational aspects of the systems or from
the direct sanctions and rewards – is uncertain.

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. (2000) interprets
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 (cf.
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, say, whether more costly and less understandable systems that focus
on value-added measurement are significantly better than status models.
Complementary Issues

The preceding discussion has concentrated on school accountability issues, but states are also interested in individual accountability. The debate about “high stakes” testing has largely, although not entirely, focused on the use of tests with significant consequences for individuals – such as when passing an examination is a requirement for high school graduation.

The discussion of high stakes individual testing has gone in a variety of directions. An important element of these discussions centers on incentives. In a series of papers, John Bishop (1995, (1996; Bishop et al. (2001) has argued that the role of the student in performance should not be ignored and that external examinations provide incentives for students to work harder.

A majority of attention has nonetheless centered on issues of fairness and reliability (see, for example, Heubert and Hauser (1999)). Specifically, similar to the prior discussion, if there is a high variance to the measurement error component \( \varepsilon_i \), the passing probability will have a large random component, leading many to be concerned about the consequences of preventing graduation for individuals who have a bad test day. Of course, no state currently has a single test administration to determine graduation so that multiple examinations are effectively given to eliminate occurrences of random errors that have dire individual consequences.

An important aspect of individual examinations is what is being measured. Individual scores have the components identified in equation 1 and therefore involve not only individual effort (part of \( X \)) but also school quality. As such, scores are partially under the control of the individual and partly not. This aspect has implications for the
incentives that are produced by holding individuals accountable for scores on examinations. In other words, if graduation requirements are put into place to ensure high levels of student effort, achievement levels can be viewed as a measure of effort that contains considerable error introduced both by randomness in the test scores and by school quality – which is largely outside the control of students.\textsuperscript{12}

While Bishop (1995, 1996; Bishop et al. 2001) finds evidence that examinations and graduation requirements have positive effects, an analysis by Jacob (2001) suggests otherwise. He finds that 12\textsuperscript{th} grade math and reading achievement are not affected by a state’s having graduation requirements, but that dropout rates for low achieving students rise. The analysis of Carnoy and Loeb (2002) is nonetheless also relevant. That analysis finds higher math score gains on the NAEP in states that have tough graduation requirements for students.

Of course, examination requirements can fulfill two separate purposes – providing incentives and indicating or identifying the level of individual proficiency. These two purposes are not identical and in fact may push in different directions (Betts and Costrell 2001), so it is important to be clear about the use of examinations and what they entail.

**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

\textsuperscript{12} Some discussions of student graduation requirements consider these as incentives for the schools to perform better. Such incentives may result through public disclosure and its impact on school behavior, but the incentives are also very indirect.
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 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. Table 3 reflects the current body of knowledge on the effectiveness of accountability systems in creating outcomes. It is clear that gaps exist, particularly as concerns the individual gain models. 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, most of these incentives 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.

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. Table 4 summarizes the previous discussion about the different incentives created by
each accountability model. As we learn more about the responsiveness of schools to the accountability systems they face, these hypothesis (and other competing ideas) will need to be tested. At present, most of these proposed mechanisms for how schools respond are unexplored.
Table 1. Aggregate U.S. Moving Outcomes from Students age 9-14, 1994-96 by poverty status

<table>
<thead>
<tr>
<th></th>
<th>Nonmovers</th>
<th>Movers within district</th>
<th>Movers across districts</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>55.4%</td>
<td>31.3%</td>
<td>13.3%</td>
</tr>
<tr>
<td>Disadvantaged</td>
<td>50.5%</td>
<td>34.7%</td>
<td>14.9%</td>
</tr>
<tr>
<td>not disadvantaged</td>
<td>59.3%</td>
<td>28.7%</td>
<td>12.1%</td>
</tr>
</tbody>
</table>

Source: Hanushek, Kain, and Rivkin (2001) from the NLSY79 data.
<table>
<thead>
<tr>
<th>Cross-sectional Approaches</th>
<th>Student Change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>School Status Model (or</strong></td>
<td><strong>Grade Level Change</strong></td>
</tr>
<tr>
<td><strong>Status Change)</strong></td>
<td></td>
</tr>
<tr>
<td>Arkansas</td>
<td>Appalachian</td>
</tr>
<tr>
<td>Alabama</td>
<td>Arkansas</td>
</tr>
<tr>
<td><strong>California</strong></td>
<td>Boise</td>
</tr>
<tr>
<td>Connecticut</td>
<td>Colorado</td>
</tr>
<tr>
<td>Georgia</td>
<td>Delaware</td>
</tr>
<tr>
<td>Mississippi</td>
<td>Florida</td>
</tr>
<tr>
<td>Maryland</td>
<td>Georgia</td>
</tr>
<tr>
<td>Michigan</td>
<td>Hawaii</td>
</tr>
<tr>
<td>Nevada</td>
<td>Indiana</td>
</tr>
<tr>
<td></td>
<td>Illinois</td>
</tr>
<tr>
<td></td>
<td>Iowa</td>
</tr>
<tr>
<td></td>
<td>Kansas</td>
</tr>
<tr>
<td></td>
<td>Kentucky</td>
</tr>
<tr>
<td></td>
<td>Louisiana</td>
</tr>
<tr>
<td></td>
<td>Maryland</td>
</tr>
<tr>
<td></td>
<td>Massachusetts</td>
</tr>
<tr>
<td></td>
<td>Michigan</td>
</tr>
<tr>
<td></td>
<td>Minnesota</td>
</tr>
<tr>
<td></td>
<td>Missouri</td>
</tr>
<tr>
<td></td>
<td>Nebraska</td>
</tr>
<tr>
<td></td>
<td>Nevada</td>
</tr>
<tr>
<td></td>
<td>New Hampshire</td>
</tr>
<tr>
<td></td>
<td>New York</td>
</tr>
<tr>
<td></td>
<td>Ohio</td>
</tr>
<tr>
<td></td>
<td>Oregon</td>
</tr>
<tr>
<td></td>
<td>South Carolina</td>
</tr>
<tr>
<td></td>
<td>Texas</td>
</tr>
<tr>
<td></td>
<td>Virginia</td>
</tr>
<tr>
<td></td>
<td>West Virginia</td>
</tr>
</tbody>
</table>

Source: CREDO (2002)
Table 3. Distribution of Studies of the Impacts of Accountability

<table>
<thead>
<tr>
<th>Outcomes</th>
<th>Cross sectional accountability systems</th>
<th>Achievement gain accountability systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome effects</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct response to</td>
<td>Greene (2001a, 2001b); Jacob (2002); Carnoy and Loeb (2002); Carnoy (2001); Deere and Strayer</td>
<td>Richards and Sheu (1992); Ladd (1999)</td>
</tr>
<tr>
<td>consequences</td>
<td>(2001a, 2001b)</td>
<td></td>
</tr>
<tr>
<td>Response to public</td>
<td>Hanushek and Rivkin (forthcoming); Carnoy (2001)</td>
<td></td>
</tr>
<tr>
<td>disclosure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Measurement errors</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing effects</td>
<td>Koretz and Barron (1998); Jacob (2002); Deere and Strayer (2001b)</td>
<td></td>
</tr>
<tr>
<td>Random errors</td>
<td>Kane and Staiger (2001)</td>
<td></td>
</tr>
<tr>
<td>Exclusions/selectivity</td>
<td>Jacob (2002); Figlio and Getzler (2002); Haney (2000); Cullen and Reback (2002); Toenjes et al. (2000);</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Carnoy, Loeb, and Smith (2001); Deere and Strayer (2001a, 2001b); Koretz and Barron (1998)</td>
<td></td>
</tr>
<tr>
<td>Other responses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher assignment</td>
<td>Boyd et al. (2002)</td>
<td></td>
</tr>
</tbody>
</table>
**Table 4. Hypothesized Impacts of Accountability**

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


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)


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).


Toenjes, L, AG Dworkin, J Lorence, and AN Hill. 2000. "The lone star gamble: High stakes testing, accountability and student achievement in Texas and Houston."
