

School-Based Educational Accountability Systems: The Promise and the Pitfalls

INTRODUCTION

Demands for more accountability and results-based incentive systems in K-12 education come from many directions and currently dominate much of the education policy discussion at both the state and the federal levels. Accountability in education is a broad concept and could be addressed in many different ways—through the design of political systems to assure democratic accountability, the introduction of market-based reforms to increase accountability to parents and children, the development of peer-based accountability systems to increase the professional accountability of teachers, or the use of top-down administered accountability systems designed to drive the system toward higher student achievement. This article focuses on the last of these approaches and pays particular attention to programs that use the individual school as the primary unit of accountability.

The programs of interest here operate within the traditional public school system, are usually part of a broader standards-based reform effort, and rely heavily on student testing (Elmore, Abelman, and Fuhrman, 1996). Forty-five states now have report cards on schools, and 27 of them rate schools or identify low performing schools (*Quality Counts, 2001*). Several of them have full school-based accountability programs, in that they rate schools based on their students' performance, provide rewards for improved performance, and provide some combination of sanctions and assistance to low performing schools. Included among these are North and South Carolina, Texas, and Kentucky. School based accountability systems were also used during the mid-1990s in some school districts, such as Dallas, Texas and Charlotte-Mecklenburg, North Carolina. The state of North Carolina's relatively comprehensive and carefully designed program serves as the exemplar for much of the discussion in this paper.

One set of fundamental concerns that arises in connection with these administered programs is their heavy reliance on student test scores, usually in just a few key subject areas, such as math and reading. Critics contend that the emphasis on test scores will narrow the curriculum, induce teachers to teach narrowly to the test, and promote a shallow approach

Helen F. Ladd
*Sanford Institute of
Public Policy,
Duke University,
Durham, NC 27708*

National Tax Journal
Vol. LIV, No. 2

to learning (Kohn, 2001). Support for some of these concerns comes from organizational theorists, who point out that when incentives are focused on only one portion of a school's mission, attention to other elements will be reduced (Milgrom and Roberts, 1992). Although these concerns are serious and deserve additional research and debate, I do not address them directly in this paper. In particular, I do not address how administered accountability programs affect the behavior of teachers within the classroom, and focus instead on incentives related to the allocation of resources and capacity both within and across schools.¹

Another concern is that accountability systems may exert their impacts on failing schools more through the fear of being sanctioned than through positive interventions designed to assist and support low performing schools. With respect to this issue, my bias is clear: I believe it is undesirable to design an education system in which the primary motivating force is one of fear of failure and sanctions. Even if that motivation were to increase student achievement in the short run, it would be hard to sustain over time given that good teachers and school principals have options other than to teach in such schools. Instead, educational accountability systems should be used primarily to identify problems in particular schools so that those problems can be addressed in a positive and constructive manner.

The basic argument I make in this paper is the following. First, subject to some important qualifications related to funding and capacity, schools are an appropriate unit for accountability purposes and have clear advantages compared to other possible units of accountability, such as school districts, individual teachers, and

students. Second, as a means of rating the effectiveness of schools, value-added measures designed to measure the contribution of schools to student learning are superior to alternative measures such as average student achievement, and this conclusion holds despite the practical difficulties of implementing true value-added measures. Third, we know little about the success of school-based accountability programs in raising student achievement but, fourth, we do know that such programs can have powerful effects—both intended and unintended—on how principals behave.

An important subtext running throughout this discussion is that school-based accountability systems make most sense when schools have adequate funding and when support systems are in place to assist the ailing schools. I conclude that state policy makers need to move cautiously as they implement such systems to make sure that accountability for results does not become the end rather than the means to the end of higher achievement for all students. The corollary is that Federal policy makers should be careful not to push states to implement accountability systems more rapidly than is justified by either the federal government's or the state governments' willingness to provide the funding and support systems needed to make them work constructively in a positive, rather than a punitive, manner.

SCHOOLS AS THE PRIMARY UNIT OF ACCOUNTABILITY

Top down accountability programs could focus attention on districts, schools, individual teachers, or students, either exclusively or in some combination. Report cards for districts, merit pay for indi-

¹ See work by Allan Odden and his colleagues on the potential for school based accountability systems to affect the performance of teachers (Odden et. al., 1997; and Odden and Kelley, 1997.)

vidual teachers, and no-social promotion policies directed at students are all examples of accountability and incentive programs. I argue in this section that the use of schools as the focal point is a logical extension of the earlier focus on school districts, that it is preferable to focusing on individual teachers, and that some form of school-based accountability needs to precede any serious effect to increase accountability through testing with high stakes for students.

A comprehensive school-based accountability system, such as the one in North Carolina, starts with a curriculum and clear content standards describing what the state or district wants children to know and be able to do. As part of that effort, policy makers have to develop a consensus on which subjects are most important and are to be the focus of the accountability system. North Carolina, for example, focuses attention on reading, math, and writing in grades 3–8 and on a range of subjects in high school, while other states, such as Kentucky, have chosen to include a larger set of subjects even at the elementary school level (Elmore, Abelman, and Fuhrman, 1996).

The state must develop assessment tools that generate reliable and valid measures of how well students have mastered the specified curriculum. Student performance on those measures then serves as the basis for measuring how effectively schools increase the learning of their students. How best to measure the performance of schools is a thorny issue and one to which I will return below. The state then provides for a system of rewards and positive incentives for schools to increase student performance and various sanctions or intervention strategies for low-performing schools. The rewards in North Carolina are \$1,500 bonuses for teachers and staff in “exemplary” schools. With respect to “low performing schools” the North Carolina legislation initially called

for automatic dismissal of school principals and intervention by state assistance teams. The state no longer automatically dismisses principals but continues to send in state assistance teams and is now discussing whether to provide additional support for such schools, including bonuses for teachers.

Schools versus Districts

As in many other states, North Carolina’s early outcomes-based accountability efforts focused initially on school districts, rather than on schools. The most common approach in various states was to publicize student achievement results at the district level in the form of district-level report cards designed to increase public pressure to improve the performance of the system. In this manner, states were trying to use public information as a policy lever to induce change. Some states, such as Mississippi, went one step further, offering rewards, in the form of greater flexibility, for districts with high achieving students and sanctions and intervention strategies for districts with low performing students (Elmore, Abelman, and Fuhrman, 1996).

A key issue in the design of such report cards was how they should account for the differing socio-economic backgrounds of the students in each district. While most states chose to make no explicit adjustments in the student performance measures, they typically included information on family background along with the performance measures, so that the public could make comparisons among comparable districts if they so chose. North Carolina, in contrast, explicitly adjusted for such differences in its district-level report cards during the period 1991–93 and drew attention to how well students in each district were doing relative to what would be expected given the socio-economic and racial characteristics

of the students.² The idea was to reduce the complacency of districts that were not contributing as much to the learning of their high socio-economic status (SES), high performing students compared to other comparable districts and to raise the morale of those districts that were contributing a lot to the learning of their low SES students.

Increasingly, states have extended their public reports and their accountability systems down to the level of individual schools. This shift to the schools is consonant with the interest throughout the country in site-based management, which places more managerial authority in the schools, and also with the introduction of charter schools that operate as autonomous units separate from districts. Moreover, the growth in computer capacity has made it more feasible than in previous periods for the state to use the school as the unit of accountability and to provide user-friendly electronic reports directly to individual schools. A major advantage of focusing on individual schools is that it minimizes the chances that poorly performing schools will escape public scrutiny. When the state focuses attention on individual schools, a district cannot ignore the plight of some of its schools in favor of maximizing the performance of the district as a whole.

Against these advantages of extending accountability down to the school level is the complication that arises because students are much more likely to move in and out of schools both during the year and from year to year than they are to move among districts. This movement complicates the task of accountability. Fair treatment of school personnel requires that schools be held accountable only for the

performance of the students they serve. Some states, including North Carolina, are now doing just that. At the same time, one wonders who is being held accountable for the students who move from school to school.

A more significant problem arises because of the relationship between local school districts and their schools. Because the district generates the revenue for schools, decides how to allocate the money among schools, and may make many of the decisions about student and teacher assignment among schools, an individual school does not have control over many aspects of the educational environment in which it is operating. As a result, even if a state-based accountability system took into account the mix of students in rating the school's performance, a school would still be at a disadvantage if it had little control over who teaches in the school and if it had inadequate funding given the mix of students it served.

Hence, for the focus on schools to make the most sense, school boards would need to change the way they function. They would need to get out of the business of micro managing individual schools and, instead, take on the role of assuring that every school within the district has the capacity (both financial and otherwise) to meet the state standards (Elmore, 1996). This new role, however, would not absolve the boards of responsibility for providing additional support for low performing schools.

Schools Rather than Individual Teachers

Since the 1920s, school districts and states have periodically experimented

² The state used multiple regression analysis to predict the average test scores of students in each of four subject areas based on district characteristics such as the percentage of enrolled students who were frequently absent, the percentage receiving free and reduced price lunches, and measures of the educational background of the parents. In the first year, the state also included the racial mix of the students, but dropped that variable in subsequent years. Results were reported for each district as deviations from the predicted results for that district. By this measure, some districts that had above-average test results appeared to be performing poorly and other districts with below-average test scores appeared to be performing well (Clotfelter and Ladd, 1994).

with programs that provide additional pay for the most effective—or most meritorious—teachers. However, most of these programs for individual teachers have not been very successful and few have survived. Those that have survived tend to be ones in which additional pay ends up being given more for additional work or responsibility than for meritorious performance (Murnane and Cohen, 1986).

The problems with merit pay programs are well known. For a variety of reasons, teachers do not like them. In part this reflects the difficulty that principals have had in developing criteria for measuring the effectiveness of teachers that correspond to professional standards of good practice. Contrary to the hopes of some advocates, the availability of student test scores does not solve the problem of measuring teacher merit. Even in a state with annual testing of students, which technically would make it possible to measure the gains in scores for a teacher's students during a year, the sample sizes would typically be too small to put much weight on the gains as a high-stakes measure of a teacher's effectiveness. In addition, to the extent that student outcomes are the joint product of many people, it is difficult to isolate the contributions to learning of individual teachers. The biggest concern about merit pay, however, is that it encourages teachers to compete with one another rather than to work together toward a common goal.

By shifting attention away from individual teachers to the school as a whole, school-based accountability programs set up a more appropriate and powerful set of incentives. First, school-based incentive rewards provide an incentive for all school personnel to work cooperatively toward a well specified goal. Second, assuming that schools are given more management authority than in the past, schools have more mechanisms (such as rearranging the use of resources, spending more on professional development, or altering

class sizes) for increasing student performance than is the case for the individual teacher. Third, school-based rewards have more potential to change the culture within a school. To the extent that John Bishop (1994) is correct that the level of work demanded of students is the outcome of negotiations between teachers and their students, school-based incentive programs might improve student performance by strengthening the power of the teachers relative to the students. Finally, incentives for schools, in contrast to individual teachers, would bring education more in line with the private sector, which is increasingly relying on group or team incentives (Nalbantian and Schotter, 1997).

The free-rider problems present a possible disadvantage of the focus on schools as the unit of accountability. Because school-based incentive programs typically provide financial rewards for all teachers (and possibly support staff as well) in a school deemed to be effective, individual teachers receive rewards regardless of their contributions to the total enterprise. Productive teachers may resent the fact that teachers who are less productive also benefit. That problem can be solved if school administrators are willing to give honest evaluations of teachers and to dismiss unproductive teachers, but not all administrators are willing to pursue that costly and time consuming process. A further and potentially more serious problem with the focus on schools is that good teachers in low performing schools may prefer to leave such schools in favor of schools where their chances of earning a financial bonus are higher (see below for further discussion of this point).

Schools versus Students

Many teachers and other school officials object to being held accountable for the performance of their students, given that students themselves play an active role in

their own learning. As David Cohen emphasizes in his critical essay on accountability programs, teachers differ from carpenters who are working with inert pieces of wood in that the students come to school with different degrees of readiness and willingness to learn (Cohen, 1996, p. 30). The economist John Bishop points out that, with the exception of students applying to elite colleges, how students perform in high school has little effect on their opportunities after high school. Many colleges attach little or no weight to high school grades, and employers seldom look at high school transcripts. Hence, argues Bishop, students need to be held more accountable through the use of external exams (Bishop et. al., 2001).

Eighteen states have now raised the stakes for students by requiring that they pass an exit exam to receive a high school diploma, and some states and districts—most notably Chicago—require that students pass end-of-grade tests in some grades in order to move on to the next grade. North Carolina is about to implement a no-social promotion policy in grades 3 and 5, and is also planning to require students to pass an exit exam to graduate. In grades 3 and 5, the students will have to score at grade level or above on the same curriculum-referenced tests that are the basis of the state's school-based accountability system. Such high-stakes tests for students are controversial and raise a number of issues beyond the scope of this paper.³

This is not the place to make the case for or against high-stakes testing of students. Suffice it to say that if policy mak-

ers choose to go in that direction, then well-designed school-based accountability systems can be viewed as a logical prior step. If schools do not have the appropriate incentives to teach children to the state's standards, it is unfair to hold the students accountable for their failure to learn. Also crucial are the resources necessary to assure that the schools have the capacity to respond to the incentives. In recognition of that point, the state Board in Maryland, a state with a school-based accountability system, has recently delayed implementing its new requirement that students pass an exit exam to graduate on the grounds that the legislature had not appropriated sufficient funds to finance the necessary intervention (Grasmick, 2001).

MEASURING THE PERFORMANCE OF SCHOOLS⁴

One of the thorniest issues in any school based accountability system is measuring the effectiveness of schools. Of particular interest, here, is whether the method generates the appropriate incentives for schools to improve, both in the short run and the longer run.

States and districts have approached this task in different ways. Some, such as Texas and Indiana, essentially use average levels of test scores, Kentucky focuses on incremental progress toward a school-specific goal, Charlotte-Mecklenburg measured performance relative to annual growth targets, and others, such as South Carolina, the Dallas Independent School District, and North Carolina, use achieve-

³ Three brief points are worth noting. First, states need to be careful in using tests designed for one purpose, e.g., holding schools accountable, for the different purpose of holding students accountable. Tests validated for one purpose may not be valid for another. Second, good practice (National Research Council, 1999) requires that states not use the outcomes of a single test to determine the education fate of a child. Exactly what that means in practice is unclear. If a state gives a child a second or third chance on the same test, does that count as the use of a single test or not? Third, the best new evidence of the effects of stringent student-based accountability comes from Chicago's experience with a tough no-social promotion policy. That evidence indicates that high-stakes testing of students in the middle school is more effective than at the third grade level in providing positive incentives for improvement. (Chicago Consortium on School Research, 2001).

⁴ This section draws heavily on Ladd and Walsh, forthcoming.

ment gains or value-added measures that are based on test score data for students that is matched from one year to the next.

Several considerations are relevant for determining the relative desirability of various approaches to school effectiveness. One is the usefulness of the measure to the state or the district for diagnosing the strengths and weaknesses of individual schools or for providing incentives for schools to improve the performance of their students. A second consideration is the usefulness of the measure to children and their parents as they make decisions among schools. A third is whether it treats the targets of accountability fairly. Typically, the target is teachers and principals (and often the staff as well) in each school. An approach would be unfair if it attempted to hold the teachers and principal of a school accountable for factors beyond their control. Fairness is important not only for its own sake but also because of how it affects the decisions of teachers and principals to respond to the new incentives and their willingness to continue working in the education system.

States have generally used one of three stylized approaches for measuring a school's effectiveness based on student test scores: 1) the school's average test scores (either absolute or relative to other schools), 2) its rate of improvement relative to its school-specific target rate of improvement, or 3) its value added. By my criteria, a value-added measure is the preferred approach but even that approach is difficult to implement fairly and in a way that provides appropriate incentives.

Average Test Scores or Pass Rates

With this approach, a school would be deemed effective if the average test scores (or pass rates) of its students in the relevant subjects were high in either a relative or absolute sense. For the relative approach, schools would first be ranked by their average scores (or the percentages of stu-

dents passing a test), and schools higher in the ranking would be deemed more effective than those lower in the ranking. A state would typically specify the percent of schools to be deemed effective in any one year. Under the relative approach, the effectiveness of a school would depend not only on its own performance during the year but also on the performance of all other schools. More consistent with the rhetoric of the standards based reform movement are absolute cutoffs that are linked to the state's standards. Under this approach, state officials would set cutoff levels of average scores (or for the percentage of students passing a test), above which schools would be deemed effective, or cutoffs below which schools would be deemed ineffective or low performing.

Because these measures do not control for factors that are outside the control of schools, such as the family backgrounds of the students, these measures would provide at best weak positive incentives for schools to improve student performance. Schools serving advantaged youngsters may become complacent and schools serving disadvantaged youngsters may believe they have no chance of ever being deemed effective. To the extent they do drive change in the low performing schools, that change is likely to be driven primarily by the fear of public humiliation and of negative sanctions. Moreover, such measures are not useful to state policy makers for diagnosing problems. Low school performance by this measure is more likely to reflect students' backgrounds rather than any specific problems in the school (Coleman, 1966; Clotfelter and Ladd, 1996). The measures may well harm the schools with low performing students by inducing good teachers and principals to leave those schools in favor of schools where their efforts are more likely to be recognized and rewarded.

Similarly, such measures provide potentially misleading information to children and parents about how much the school

is likely to add to student learning. Instead, the measures primarily provide information about the background of the typical student in the school. While that information is not irrelevant to parents (and could be very important to some), it differs from information about the contribution of the school to student learning.

Texas adds an additional element to its accountability rating standards. For a school to reach a specific rating (such as exemplary or recognized), not only must all students meet the passing requirement (90 percent for exemplary and 75 percent for recognized), but each student subgroup must meet it as well. The student groups are African-American, Hispanic, white, and economically disadvantaged. The policy goal is clear: Texas wants to assure that no group of students is left behind.

High Rates of Improvement Relative to Target Rates of Improvement

An alternative approach would label a school effective if it increased the learning of its students during a specified period, such as a year or two years, relative to a target specified by the state. How the state sets the targets becomes crucial for judging this approach.

Consider, for example, Kentucky, the best known of the programs that employ this approach (Elmore, Abelman, and Fuhrman, 1996). Policy makers in Kentucky start with the same target levels of student performance for all schools and then they work backward to set target growth rates. Roughly, the ultimate goal is expressed in terms of the percentage of students who meet a state-specified proficiency standard. Because the state would like all schools to reach the goal within 20 years, it sets two-year target growth rates as 1/10th of the difference between student performance in

the school in the first year and the 20-year goal. This approach means that schools serving low performing students must increase student performance more in each period than those serving high performing students. This approach is designed to assure that all schools eventually meet the state standards. To keep schools from meeting their targets by focusing primarily on the higher performing students, Kentucky requires that schools increase the scores of their lowest performing student (those at the novice level) at the same overall rate required for the school as a whole.⁵

At one level this target approach is highly appealing. The focus on incremental gains solves some of the problems that arise with the focus on levels of the previous approach. Moreover, each school has clear targets to aim for during a one- or two-year period, and, in Kentucky, these targets are consistent with the long-term goal for each school. The problem arises because the target rates of improvement are driven by what the policy makers want to have happen, not necessarily by what is feasible. Only if the targets are consistent with what well-functioning schools could be expected to do given the set of students they serve and the resources they have would the approach treat principals and teachers fairly. Moreover, it is likely that the targets will be easier to meet in some schools than in others. Schools that start near the 20-year goal will be deemed effective—and the principals and teachers rewarded—simply for continuing what they were doing in the past whether or not the school's value added is high. In contrast, schools starting with low performing students may find it difficult, if not impossible, to meet their above-average growth targets. The danger here is that principals and teachers in those schools could well be pe-

⁵ A similar approach was used in the Charlotte-Mecklenburg school district in the early 1990s. Known as the benchmark goals program, performance goals were set for each school and rewards given to the schools that met their goals each year. To reduce the gap between the high- and low-performing students, more ambitious goals were set for schools with low-performing students than for other schools.

nalized for doing a reasonable job given the resources they have simply because they did not meet their unrealistic targets.

Another element of inequity built into the Kentucky approach is that the effectiveness measure is based on comparisons of different groups of students from one year to the next, such as fourth graders in one year and fourth graders in the next. To the extent that the composition of the students changes from year to year, the state is attempting to hold principals and teachers accountable for factors outside their control.

Value-Added Approaches

A school's value added can be roughly defined as the amount of student learning during the year that is attributable to the school. Notice the ambiguity here. What does "the school" mean? Is it just the teachers and the staff? Or is it all the resources available to the school (including the budgeted resources and those provided by parents, foundations, and local corporations)? Should the school's contribution include the effect of the composition of the student body, which, through peer effects, may affect the learning of others in the classroom? These questions are important since the answers will determine both how fair the resulting measure is to the teachers and principals and could affect its efficacy as an incentive.

Such models have three major strengths compared to the other approaches. They are preferable to measures of average levels of student performance in that they attempt to measure the contributions of the school to the gains in learning from one year to the next. They are preferable to the incremental approach used in Kentucky in part because the measures of school effectiveness are based on student performance gains that are related to what past performance has demonstrated to be feasible or relative to what other schools are able to do. Also, in contrast to approaches that are based on all of a school's students (in one

or more grades), the student-based value-added approach makes it possible to measure each school's performance based on students in all grades and based only on the students who were in the school for a large part of the year. This consideration is particularly important in urban school systems serving low-income students whose families tend to move a lot each year.

Measuring the value added of a school is difficult, especially if the goal is to develop a measure that will be used as the basis for rewarding teachers and principals within the school. The challenge arises because of the difficulty of separating the contributions to student learning of the school personnel from those of other factors. To isolate the contributions of the school personnel, one would want to control statistically for other variables such as the family background characteristics of each student, community characteristics that might affect student motivation, and, to the extent they are outside the control of the schools, the composition of the student body in the student's class and the resources available to the school.

For a variety of conceptual, practical and political reasons, it is difficult to specify an appropriate set of control variables and hence to generate a good measure of how much the school's teachers and staff contribute to student learning. One problem is that state agencies typically do not have all the data required for each student and some of the data they may have, such as parents' education as reported by the student, are likely to be measured with error. Second, it is unclear which characteristics of the students should be included. The most controversial characteristic is the student's race. With the exception of the Dallas accountability system, programs do not control for a student's race on the grounds that some people might interpret the adjustment for race as sending the inappropriate signal that the expectations for minority children are less than those for white

children. A third problem is the difficulty of controlling appropriately for peer effects within the school given the statistical problems associated with the fact that a child's peers are determined in part by the choices made by parents and hence are endogenous. Finally, controlling for resource levels is far more difficult than one might expect given the state often does not maintain such data for individual schools and would find it difficult to allocate expenditures by the central office to the schools. In addition, one would not want to control for additional resources that the school generated through its own entrepreneurial activity. Dallas, the one district that tried to implement the full value-added model, fell short by including only a few family background variables, by not including resource variables, and not controlling for peer effects. Hence, the experience in Dallas supports the conclusion that the full model is indeed hard to implement.⁶

Other states, including South and North Carolina, have opted for simpler approaches that require data only on test scores (Ladd and Walsh, 2001). Both states use test score data that are matched by student and measure the gains in scores from one year to the next. In North Carolina, schools are rewarded for having gains in excess of the gains that would be expected given the prior year test scores of the students they serve. The absence of controls for factors outside the control of school personnel, including the socio-economic mix of the students, means that the measured effects should be interpreted as all the effects on student learning associated with the school, including both those that are within the control of school personnel and those that are not. A key policy question is whether these resulting measures are useful and, if so, for what purposes.

Such measures provide useful information for typical parents and students in their

capacity as choosers of schools. However, the measures should at best be used with caution as the basis for rewards and sanctions for principals and teachers. If schools do not have adequate resources that appropriately account for the mix of students they serve, and hence for the fact that children from some family backgrounds have an easier time learning than those from other backgrounds, the use of these measures as the basis for rewards could reduce teacher morale in low performing schools and could hurt those schools further in the long run by providing an additional incentive for good teachers to leave those schools in favor of other schools where they would have a greater chance of being rewarded.

Nonetheless, such measures can still drive policy in productive directions. The measures provide information to parents, citizens, and policy makers in their role of assuring that all schools are facilitating an acceptable amount of learning during a year. Poor performance by these measures indicates that something needs to be done to make the schools better. Importantly, however, some schools may need more resources to offset the effects of factors outside the control of school personnel in addition to harder and smarter work by existing teachers. Thus, to the extent that an accountability system based on school effectiveness measures of this type generates pressure on policy makers to try to determine the causes of school ineffectiveness and to intervene in ways to make the low performing schools more effective, such measures would be useful.

North Carolina's experience with such an approach indicates both its feasibility and also some of the statistical problems that it poses. Recent research identifies two types of statistical problems. One problem of measurement error is that, as shown by Ladd and Walsh (2001), it tends to bias the

⁶ Dallas' goal was to create a fair accountability system, one in which every school has an equal chance of being a winner. The district succeeded in that the winning schools each year are distributed among all schools with no apparent bias based on the race or socio-economic status of the students in each school.

school effectiveness measures in favor of the schools serving the more white and more affluent children. A second problem arises because of the variation in the size of schools. For small schools the amount of random error, or noise, is large relative to the effect that policy makers are trying to measure. The result is that such schools are both more likely to be deemed effective and also to exhibit the greatest swings in measured, although not necessarily in true, effectiveness from one year to the next (Kane and Staiger, 2000).

These statistical problems notwithstanding, such measures of school effectiveness based on student gains from one year to the next are far superior to measures based on average levels of student performance. Hence, these gain measures should be used more for the purpose of driving change than as a vehicle for placing blame. To the extent that low measured value added in a school may reflect factors that require additional resources or support systems that the teachers currently do not have access to, the shortfall in student performance should not be attributed to the teachers alone.

DO SCHOOL-BASED ACCOUNTABILITY AND INCENTIVE PROGRAMS INCREASE STUDENT ACHIEVEMENT?

Given the current attention to educational accountability, surprisingly little is known about how accountability programs affect student achievement. Grissmer and Flanagan (1998) have observed that the two states with the largest gains in National Assessment of Educational Progress (NAEP) scores during the mid-nineties, North Carolina and Texas, both have well developed accountability systems. However, one should not draw

causal inferences, especially given that North Carolina did not implement its comprehensive school-based accountability system until 1997. The Texas program has received the most attention from scholars, but as pointed out by Murnane and Levy (2001) the effects of that program on student achievement, retentions, and school dropouts are in dispute. While the Texas Education Agency touts large gains in test scores, a closing of the performance gap between minorities and whites, and declines in school dropout rates, independent scholars dispute the validity of these claims.

Even if observers could agree on the magnitude and distribution of the gains in Texas test scores, they still would not know the extent to which the gains were attributable to specific components of the Texas accountability system such as the high stakes exit exams for students or the assignment of effectiveness rating to schools, or to other changes, such as the infusion of substantial new revenue in 1995. Measuring impacts for state-wide programs is further complicated by the lack of a clear counterfactual to which to compare the new system.

The value-added accountability system introduced by the Dallas Independent School District provides the best laboratory I am aware of for examining the impact of a school-based accountability system on student outcomes. First, the effects of the Dallas accountability system can be reasonably cleanly estimated given the absence of other major confounding policy changes in Dallas. Second, the counterfactual problem can be solved by comparing gains in student performance in Dallas to those in other comparable Texas school districts, all of which are operating within the same statewide policy context.⁷ In Ladd (1999), I examined gains in student performance on the Texas Assessment of Academic

⁷ That method assumes that none of the other urban Texas school districts had their own local reform initiatives. To the extent some of them had such initiatives and that they were successful, students in Dallas would have had to do even better for the Dallas program to generate statistically significant positive impacts. Hence, the basic model represents a relatively stringent standard for evaluation. An extended model changed the basic comparison to schools in other districts not engaged in local reform efforts.

Skills (TAAS), a test that is linked to the state's curriculum and serves as the basis for statewide accountability, and also at drop-out rates during the period 1991-95. Given that student performance on the TAAS was one of the outcome variables included in the Dallas accountability system, it was reasonable to expect a successful Dallas program to generate better student performance on the TAAS.

Some gains in achievement emerged for seventh grade students, but even these results are subject to interpretation because the big gains in test scores in Dallas schools relative to those in other districts occurred before the program was fully in operation. The most favorable interpretation of those early gains was that they reflected the new focus on outcomes and the extensive publicity associated with the program during the summer of 1999.

The findings differed for subgroups of the population, with the results positive for white and Hispanic students and small and insignificant for African-American students. The magnitude of the impacts depends on the appropriate interpretation of the coefficients of the first-year variable. If those results represent true program impacts then the impacts by 1995 were on the order of 10 to 20 percentage points relative to the state average for whites and Hispanics. If, however, the first-year results reflect regression to the mean or simply the gains from narrow teaching to the test, then the program gains by 1995 would be smaller, on the order of 4 to 7 percent for Hispanics and 5 to 10 percent for whites (Ladd, 1999).

The observation that most of the gains occurred during the first year, with few gains after that year, is important in that it suggests that the accountability system did not serve as a catalyst for the significant changes necessary to generate ongoing gains in student achievement over time. To supporters of standards-based systemic reform, this outcome should not

be particularly surprising. Instead of incorporating the accountability system into a broader reform effort that assured that the schools had the capacity to teach the students they served, Dallas put all its policy efforts into the accountability system itself.

IMPACT ON THE BEHAVIOR OF SCHOOL PRINCIPALS

Even if we do not know much about the impacts of school-based accountability systems on student achievement, we do have evidence from North Carolina that such systems can serve as a potentially powerful tool for changing the behavior of one set of key adults in the system, school principals. This evidence comes from a recent study of the responses of elementary school principals to North Carolina's highly touted ABCs program (Ladd and Zelli, 2001).

The ABCs program is designed to hold schools Accountable for the *B*asic skills of reading, math and writing, while giving the schools more local Control. The schools were given greater managerial flexibility in return for being held accountable for gains in the achievement of their students. Each school is expected to provide at least a year's worth of learning for a year's worth of schooling, where the expected gains are based on statewide average gains with some minor adjustments. School teachers and administrators working in schools that exceed their expected achievement gains by more than 10 percent receive \$1,500 bonuses. Low performing schools, that is, those that meet neither the school's expected growth in achievement nor the state's performance standard of having more than half the students at grade level are subject to increased scrutiny and interventions from state assistance teams.

Our analysis was framed within the context of a principal-agent model in which we view state policy makers as the

“principal” with the goal of improving educational outcomes in the state and the school principals as the “agents” of the state. We were interested in the extent to which North Carolina’s ABCs school accountability system has succeeded in altering the behavior of school principals in the direction desired by the state. To that end, we surveyed principals both about their attitudes toward the ABCs program and about how they have altered the way they allocate their own time and the use of resources within the school. We administered the survey first in the summer of 1997 after the program’s first year and again in the summer of 1999.

The principals generally supported the ABCs program, with over 60 percent having a positive view overall (at the level of four or five on a five-point scale) and the others being evenly divided between a neutral and a negative view of the program. This positive attitude toward the program, and also the principals’ general agreement with many of the basic components of the program, is largely attributable to strong educational leadership at the state level and to the state’s efforts to communicate with local school officials. Further, it reflects the fact that the ABCs program itself was part of an accountability movement that had been evolving throughout the decade.

The survey results indicate that school principals were well aware that student performance on the end-of-grade tests (EOG) would play a larger role than in previous years in the direct evaluation of their own performance and also would affect the rating of their schools. In addition, two out of three principals perceived that the ABCs program had increased their ability to make their teachers more effective, presumably in part by highlighting the potential for teachers to earn financial rewards or public recognition. Without such empowerment, principals could well have believed that they did not have the tools necessary to effect change at the classroom level.

The ABCs program clearly changed the behavior of school principals. By comparing the actions the respondents described as ongoing at the time of the 1997 survey to the actions described as ongoing and new in the 1999 survey, we were able to conclude that principals had become much more active in a wide range of policy areas. In response to the ABCs program, we found that school principals increased their use of EOG as a diagnostic tool to help teachers improve instruction, developed extracurricula programs focused on math and/or reading, spent more time with teachers on classroom instruction, and encouraged greater focus on math and reading in the teaching of other subject areas. All of these actions are fully consistent with the state’s goal of raising student achievement. In addition, principals redirected resources from other subject areas to math and/or reading and encouraged teachers to spend more time on teaching test-taking skills, new actions that the state did not necessarily intend and that in some ways could be detrimental to students.

Importantly, the incentives of the program were sufficiently strong to induce even those school principals who opposed the ABCs program to change their behavior in ways similar to the supporters of the program. While principals who agreed with the basic thrust of the ABCs program were initially more aggressive than the others in changing their behavior, by 1999, the actions of principals who did not support the ABCs program were indistinguishable from those who did.

Within schools, principals overwhelmingly said they focused new attention on the low performing students, with that focus increasing from about 63 percent of the principals to over 80 percent in 1999. Some principals indicated that they focused as well on the high performing students. The students receiving the least new attention under the ABCs program were the students at grade level in the middle of the distribution.

From responses to open ended questions, we learned that in some cases principals had access to additional funding that they could use to assist the low performing students. In many cases, however, it appears that principals had either to shift resources away from the other groups of students or had to ask teachers to spend additional "voluntary" hours after school or on Saturdays working with these students. Among the other actions taken by various principals were increased use of tutoring, more individual assistance, more grouping of students by ability, greater use of computer technology (mentioned by only a few principals), and restructuring of the reading program.

The greater focus by principals on low performing students within schools is a desirable outcome that is fully consistent with the objectives of state policy makers. Across schools, however, the situation is more complex. One concern emerges from the observation that principals in schools with large proportions of low performing students were less supportive of the ABCs program than were other principals and that they were less optimistic than other principals about their increased power to remove low performing teachers. Such views could reduce the future willingness of ambitious and effective principals to serve in such schools. Second, the surveys suggest that schools with low performing students may find it even more difficult than in the past to attract the higher quality teachers. While such teachers have always had incentives to move from such schools to schools with more motivated and easier-to-teach students, the ABCs program enhances those incentives. By moving to "better" schools, teachers can increase the chances of receiving a bonus and can minimize the chances of being associated with a publicly identified failing school. As a result, the ABCs program could ultimately reduce the quality of staff in the schools serving low performing stu-

dents. Unless countered by other specific policy actions, this systemic effect could potentially outweigh any short run positive effects of increased effort within schools directed toward low performing students.

CONCLUSION

This paper has addressed none of the controversial issues surrounding testing *per se*. Nor has it focused on how administered accountability systems affect the behavior of teachers in the classroom. Though both of these issues are important for a full evaluation of top-down administered accountability systems, my focus here has been on how school-based accountability systems affect how schools operate, with particular attention to the incentives, both positive and negative, that they create for changes within and among schools. My primary conclusions are that in many ways schools are the most logical starting point for a top-down administered accountability system and that value-added methods for evaluating schools can be designed that have the potential to drive the system in productive directions.

However, such systems must be used carefully and introduced cautiously. Evidence from North Carolina documents the power of such systems to change the behavior of school principals. While many of the behavioral responses of the principals can be viewed as positive and consistent with state goals, not all are. Importantly, the North Carolina experience hints at some basic incentive problems that arise in such systems, including, for example, the possibility that such systems could induce the higher quality, more mobile teachers and principals to avoid the schools serving the students who are most difficult to educate. To avoid this outcome, state policy makers will need to assure that the funding for such schools is adequate, and perhaps to provide special

salary bonuses or supplements to induce high-quality teachers and principals to work in such schools.

The lessons for the federal government are quite clear. The efforts of George W. Bush and his new Secretary of Education to force all states to develop more test-based accountability may well be counterproductive if states do not have the capacity and the systems in place to provide the schools serving low performing students the support they need to be successful.

Acknowledgments

Many of the ideas in this paper are based on a longer paper commissioned by the Committee on Education Finance of the National Academy of Sciences. The author thanks the national academy for financial support for that project and Stacy Zotter for research assistance.

REFERENCES

- Bishop, John H.
"Signalling, Incentives, and School Organization." Cornell University Working Paper No. 94-25. Ithaca, NY: Center for Advanced Human Resource Studies, New York State School of Industrial and Labor Relations, 1994.
- Bishop, John H., Ferran Mane, Michael Bishop, and Joan Moriarty.
"The Role of End-of-Course Exams and Minimal Competency Exams in Standards-Based Reforms." In *Brookings Papers in Education Policy 2001*, edited by Diane Ravitch. Washington, D.C.: Brookings, 2001.
- Clotfelter, Charles T., and Helen F. Ladd.
"Information as a Policy Lever: The Case of North Carolina's School Report Card." Paper presented at 1994 Annual Research Conference of APPAM, 1994.
- Clotfelter, Charles T., and Helen F. Ladd.
"Recognizing and Rewarding Successful Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 23-64. Washington, D.C.: Brookings, 1996.
- Cohen, David K.
"Rewarding Teachers for Student Performance." In *Rewards and Reform, Creative Educational Incentives That Work*, edited by Susan Furhman and Jennifer A. O'Day, 60-112. San Francisco: Jossey Bass, 1996.
- Coleman, James, E.Q. Campbell, C.J. Hobson, J. McPartland, A.M. Mead, F.D. Weinfeld, and R.L. York.
Equality of Educational Opportunity. Washington, D.C.: U.S. Department of Health, Education, and Welfare, Office of Education, 1966.
- Consortium on Chicago School Research. Annual CPS Test Review Data Brief, 2000. Research Data Brief, 2001.
- Elmore, Richard.
"Accountability in Local School Districts: Learning to do the Right Things." Paper prepared for the First Edwin O'Leary Symposium on Financial Management at the University of Illinois at Urbana-Champaign, October, 1996.
- Elmore, Richard F., Charles Abelman, and Susan Fuhrman.
"The New Accountability in State Education Reform: From Process to Performance." In *Holding Schools Accountable: Performance-Based Reform in Education*, edited by Helen F. Ladd, 65-98. Washington, D.C.: Brookings Institution, 1996.
- Grasmick, Nancy S.
"Mistake We Can't Repeat." *Education Week* 20 No. 16 (January 10, 2001): 88.
- Grissmer David, and Amy Flanagan.
Exploring Rapid Achievement Gains in North Carolina and Texas. Washington, D.C.: National Education Goals Panel, 1998.
- Kane, Thomas J., and Douglas O. Staiger.
"Improving School Accountability Measures." NBER Working Paper No. 8156. Cambridge, MA: National Bureau of Economic Research, 2001.
- Kentucky Department of Education.
KIRIS Q&A, August, 1996.

- Kohn, Alfie.
 "Fighting the Tests: A Practical Guide to Rescuing Our Schools." *Phi Delta Kappan* 82 No. 5 (January, 2001): 349–57.
- Koretz, Daniel M., Sheila Barron, Daren J. Mitchell, and Brian M. Stecher.
 "Perceived Effects of the Kentucky Instructional Results Information System (KIRIS)." Santa Monica, CA: Rand Corporation, 1996.
- Ladd, Helen F.
 "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes." *Economics of Education Review* (1999).
- Ladd, Helen F., and Randall P. Walsh.
 "Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right." *Economics of Education Review* (forthcoming).
- Ladd, Helen F., and Arnaldo Zelli.
 "School-Based Accountability in North Carolina: The Responses of School Principals." Duke University. Mimeo, 2001.
- Milgrom, Paul, and John Roberts.
Economics, Organization, and Management. Englewood Cliffs, NJ: Prentice Hall, 1992.
- Murnane, Richard J., and David K. Cohen.
 "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and A Few Survive." *Harvard Educational Review* 56 No. 1 (February, 1986): 1–17.
- Murnane, Richard J., and Frank Levy.
 "Will Standards-Based Reforms Improve the Education of Students of Color?" *National Tax Journal* 54 No. 2 (June, 2001): 401–16.
- Nalbantian, Haig R., and Andrew Schotter.
 "Productivity Under Group Incentives: An Experimental Study." *American Economic Review* 87 No. 3 (June, 1997): 314–41.
- National Research Council. Committee on Appropriate Test Use.
High Stakes: Testing for Tracking, Promotion, and Graduation, edited by J.P. Heubert and R.M. Hauser. Washington, D.C.: National Academy Press, 1999.
- North Carolina Department of Public Instruction.
School-Based Management & Accountability Procedures Manual.
- Odden, Allan, Herbert Heneman, David J. Wakelyn, and Jean Protsik.
 "School-Based Performance Award Designs: A Case Study." Paper presented at AERA, Chicago, March, 1997.
- Odden, Allan, and Carolyn Kelley.
Paying Teachers for What They Know and Do: New and Smarter Compensation Strategies to Improve Schools. Thousand Oaks, CA: Corwin Press, 1997.
- Quality Counts.
Education Week 20 No. 17 (January 11, 2001).