Sorting Out Accountability Systems

Eric A. Hanushek and Margaret E. Raymond

Accountability has been a catchword in education for decades, for who could be against it? It has not been a reality, however, because accountability threatens many and because, even when desired, it is difficult to implement. There are signs, however, that times are changing. Today, accountability is not only taken more seriously but also sometimes promises to have real teeth. Yet the future is far from certain. Although many states and districts are moving forward with accountability schemes, they are likely to run into real problems that compromise and distort these programs' impact. Though it seems natural to measure outcomes and hold schools responsible for them, the mechanics of how to do that appropriately are complicated. Creating effective accountability schemes will require a deeper understanding of how these programs alter incentives in schools and in turn the dynamics of accountability.[1] Understanding these issues is important because many people tend to generalize erroneously from problems imbedded in specific accountability systems to assertions of inherent weaknesses in all accountability systems.

[1]These considerations are also not unique to schools. The recent growth of research into corporate accountability systems underscores how the simplicity of the idea contrasts with the reality of the application.

Considerable controversy accompanies accountability in schools. Parents, teachers, policymakers, and the American public frequently enter into debate about various elements and uses of accountability systems. These debates are motivated by different underlying views about how best to improve the education of our youth as well as by self-interested reactions. This discussion does not dwell on the controversies but instead focuses on the key elements that enter into the incentives that are created by them.

The origin of today's need for accountability can be traced to the historical development of the U.S. educational system, which is thus briefly reviewed here. The structure and function of current accountability systems are then described. Following that, issues brought up by implementation and program impacts are discussed.

The importance of the accountability movement should not, however, be missed. By focusing attention on student performance, the policy debate has dramatically shifted. The challenge now is capitalizing on this movement to bring about improvements in outcomes.

THE STATE OF U.S. EDUCATION

Understanding the dynamics of the U.S. educational system sheds light on the current thrust toward accountability and the issues facing today's policymakers. In simplest terms, student performance has stagnated while costs have steadily increased. These simple facts have led to the realization that just providing more resources within the current structure is unlikely to be effective. Nor does adding further regulation offer much promise.

This stagnation is illustrated by the results of the National Assessment of Educational Progress (NAEP), which annually tests students across the country in different subject areas. The tests, which have been conducted over the past three decades, start with a random sample of students from different grade levels. A summary of the performance of 17-year-olds over time is provided in Figure 1. This figure tracks average scores

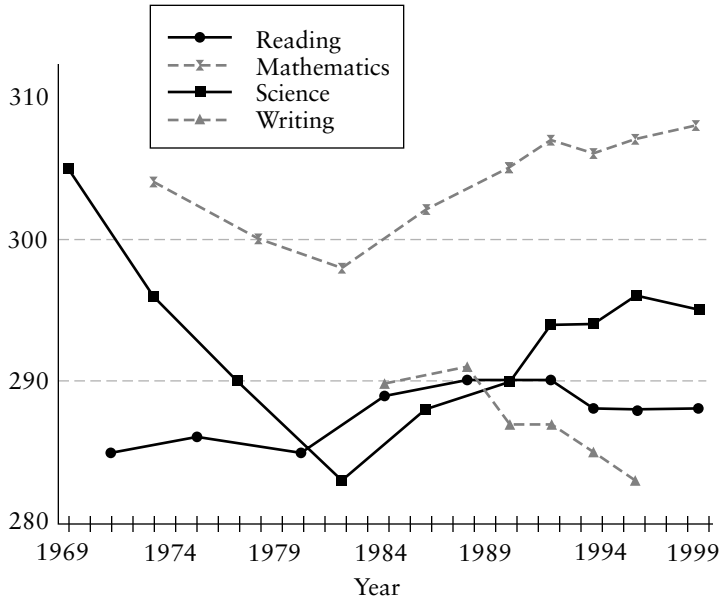

FIGURE 1. National Assessment of Educational Progress—17-year-olds

in reading, math, science, and writing.[2] The story is one of flat achievement. Reading and math scores are slightly higher at the end of three decades, whereas science and writing appear to have noticeably declined.

Level performance would not be a matter of serious concern except for two important additional trends. First, it parallels mediocre performance on the international level, where the United States has performed at or below average since the 1960s.[3] Second, the lackluster U.S. performance has not been for want of trying. As Table 1 shows, school resources have been increased over the same period of time. Real spending per

[2]The writing tests were first introduced in 1986 and then dropped after 1996 because of concerns about both the expense and the reliability of the tests over time.

[3]At least in recent years, these results do not reflect international differences in selectivity of schooling or test taking but instead appear to reflect more fundamental forces. A summary of the performance of countries across the tests along with references to the basic data can be found in Eric A. Hanushek and Dennis D. Kimko, "Schooling, Labor Force Quality, and the Growth of Nations," *American Economic Review* 90, no. 5 (2000): 1184–1208.

TABLE 1
Public School Resources in the United States, 1960–1995

Resource	*1960*	*1970*	*1980*	*1990*	*1995*
Pupil-teacher ratio	25.8	22.3	18.7	17.2	17.3
% teachers with master's degree or more	23.5	27.5	49.6	53.1	56.2
Median years teacher experience	11	8	12	15	15
Current expenditure/ADA (1996–97 $s)	$2,122	$3,645	$4,589	$6,239	$6,434

student more than tripled between 1960 and 1995.[4] This increase in resources was accomplished in the way typically called for by reformers and policymakers: by significantly reducing pupil-teacher ratios, by increasing the training of teachers, and by developing a more experienced teaching force.

The dominant approach to policymaking over much of this period has been regulation of education inputs and processes. Efforts have been concentrated on providing resources for specific programs in the schools. This approach has been especially appealing to legislatures and courts—the places where overall fiscal decisions tend to be made—because it is easy to set resource policy. But, as shown in the aggregate data, increased resources have not improved performance. Moreover, these overall impressions have

[4]Some have argued that the simple data on resources overstate what is available for schools for improvement. Specifically, because of productivity increases in other industries, wages of educated workers in schools (teachers) are driven up, and the price deflators for school spending might be too low (Richard Rothstein and Karen Hawley Miles, *Where's the Money Gone? Changes in the Level and Composition of Education Spending*, Washington, DC: Economic Policy Institute, 1995). Additionally, increased demands such as those generated by laws for special education may draw resources away from the regular education students who are tested by NAEP. Each of these arguments has some legitimacy but cannot eliminate the significant rise in real resources devoted to schools (see Eric A. Hanushek and Steven G. Rivkin, "Understanding the Twentieth-Century Growth in U.S. School Spending," *Journal of Human Resources* 32, no. 1 (1997): 35–68).

been reinforced by similar findings of analyses of performance across classrooms and schools.[5] And there is little evidence that the special emphasis of the courts on the distribution of outcomes (expressed, however, in terms of required changes in funding distributions) has narrowed variation in student results.[6]

This lack of improved performance has brought attention to alternative means of effecting change in schools. This attention has been manifested in a variety of forms (discussed below), but a common theme has been the regulation of outcomes rather than the more traditional regulation of process and inputs. Previous efforts were based on providing or prescribing specific inputs (such as reduced class size in specific circumstances) and hoping that these led to improved student performance. Often, however, these decisions were based on little information that would indicate high probabilities of success. The new regulatory frameworks tend to emphasize objective outcomes while letting schools decide how they would meet demands for achievement. The underlying idea is that public monitoring and reporting of student outcomes would drive innovation and competition in schools and would bring about improvement.

A prime example of the change to performance focus is the development of Goals 2000. Because of concerns about school performance, the nation's governors met in an unprecedented summit with President George H.W. Bush in 1989. As a result of this meeting, a commitment was made to a set of national educational goals. These goals included such resolutions as "the United States should be first in the

[5]Eric A. Hanushek, "Assessing the Effects of School Resources on Student Performance: An Update," *Educational Evaluation and Policy Analysis* 19, no. 2 (1997): 141–164.

[6]Thomas A. Downes, "Evaluating the Impact of School Finance Reform on the Provision of Public Education: The California Case," *National Tax Journal* 45, no. 4 (1992): 405–419; Eric A. Hanushek and Julie A. Somers, "Schooling, Inequality, and the Impact of Government," in *The Causes and Consequences of Increasing Inequality*, Finis Welch, editor, Chicago: University of Chicago Press, 2001.

world in science and math performance by 2000." Though this bit of wishful thinking later was belied by international test scores, it nonetheless underscored the movement toward measurable goals based on student outcomes.[7]

The Goals 2000 ideas blended into what is today perhaps the most acclaimed path to educational improvement: the so-called "standards-based reform." This approach to education reform relies on setting educational goals and measuring progress toward them. Public disclosure of both is considered a feasible way to ensure goal achievement. Nonetheless, there are many ways to implement this approach, suggesting that achieving the results desired is not automatic.

For the present discussion, it is sufficient to note that attention to results created by these reform efforts has moved most states to begin development of accountability systems. The design, use, and impact of such systems is the subject of this analysis.

The underlying perspective throughout this analysis is that accountability systems should be viewed as an inherent source of incentives designed to push schools toward desired outcomes. The ultimate impact of accountability efforts depends upon the precision and force of the incentives they create.

Sea Change in Policy Perspective

Accountability systems have been developed almost universally across the states to deal with the aggregate performance shortcomings that are now widely recognized. That history has shown that we do not know how to link programs, resources, and other inputs to student outcomes so that

[7]Subsequent modifications of the original goals have added confusion, however, by moving more toward inputs as opposed to outcomes. Instead of considering just school completion, performance, and so forth, the goals now include expanding parental participation in education and ensuring safe and drug-free schools.

regulation of inputs cannot be assumed to satisfy outcome objectives. The sea change of moving from a basic regulatory environment to one that emphasizes performance and outcomes can be interpreted as recognition that something else has to be done.

The importance of this changed perspective should not be underestimated. *If one is interested in outcomes, one should focus on outcomes.* As simple as this principle might be, it has not been recognized previously.

States now routinely develop snapshots of how students are doing in each year. To varying extents, they also use these snapshots to provide views about the performance of schools and teachers.

These systems are premised on an assumption that a focus on student outcomes will lead to behavioral changes by students, teachers, and schools to align with the performance goals of the system. Part of this is presumed to be more or less automatic (i.e., a public reporting of outcomes will bring everybody onto course to improve those outcomes). But part also comes from the development of explicit incentives that will lead to innovation, efficiency, and fixes to any observed performance problems.

The governance of schools is, nonetheless, currently in transition. States have not entirely bought into an exclusive focus on outcomes. They are reluctant to let go completely of a long tradition of input regulation. A benign interpretation is that tracking inputs and processes can provide important comparative data to understand better the distribution of outcomes. However, a more prescriptive treatment of inputs, though conflicting with the overall intent and working of accountability for outcomes, may also reflect a combination of uncertainty about how to design an outcome-based system along with political pressures to do other things.

Our question is simply, Given what states are doing, is it likely that we will get to the improved performance that is desired and expected?

Current Practice[8]

The basic skeleton of accountability systems involves goals, content standards, measurement, consequences, and reporting. Although states differ in significant ways, a general description of the structure of these systems is useful in comparing actual plans and how their elements interact. Below is a description of each element, followed by a look at currently unanswered questions.

Goals. An accountability system begins with a set of goals about what is to be accomplished by the accountability system. Though this is often phrased in very general and lofty terms (e.g., "ensure that all students have sufficient skills to participate in society"), the goals have a distinct role because precise standards and measurements are usually based on these goals. Nonetheless, most states' goals are created in the underlying statutes that create their accounting systems, leaving them frequently ambiguous and difficult to measure. Though perhaps necessary to ensure legislative approval, such vaguely worded goals leave real ambiguity about what is to be done by whom.

A notable aspect of the goals statements of accountability systems that can have real impact is where they place the focus of attention—that is, whether they focus on students, schools, or teachers. In the current development, it is frequently suggested that each group feels targeted, although the degree of attention to each differs significantly across states. The differences in focus are often related to the strength of incentives ultimately generated for each participant in the system.

Content Standards. Content standards typically present the details of what is expected. They create boundaries or domains for attention. The typical student outcome standards

[8]The profiles of current accountability system practices are based on data from "Quality Counts 2001: A Better Balance," *Education Week,* January 11, 2001.

delineate to what extent students should demonstrate mastery of a body of material that has been designated by an authoritative body to represent a minimum acceptable set of knowledge. Forty-nine states have established academic standards for student achievement: thirty-six states have standards in English or language arts, forty-four in mathematics, forty-three in science, and twenty-seven in social studies. Many researchers believe that this movement to explicit measurement of performance is key to current school reforms.[9]

Standards involve selection of a subset of all possible elements in a domain to both represent the whole and to be used to extrapolate more generalized performance. Although apparently straightforward, the creation of precise standards has been fraught with difficulty. Tension exists between the need for a representative set of elements and the need for the elements to be testable (discussed below). Tensions also exist between standards and learning goals; take the opposition, for example, between those who advocate rigorous standards and those who say such standards do not adequately assess higher-order curriculum or reasoning.

Standards are introduced in order to change behavior. The current standards-based reform explicitly argues that the development of standards will (almost necessarily) lead to better performance. As a snapshot of the interim effect of adopting standards, a national survey asked teachers if they had altered their classroom behavior.[10] The majority of respondents indicated that standards have necessitated more challenging curriculum and a focus on material delineated by the standards. This internal view has, however, yet to be matched with evidence that student performance has been

[9]Cf. Richard F. Elmore, Charles H. Abelmann, and Susan H. Fuhrman, "The New Accountability in State Education Reform: From Process to Performance," in *Holding Schools Accountable: Performance-Based Reform in Education*, Helen F. Ladd, editor, Washington, DC: Brookings Institute Press, 1996: 65–98.

[10]Belden, Russonello, and Stewart, *Making the Grade: Teachers' Attitudes toward Academic Standards and State Testing*, Washington, DC: Belden, Russonello, and Stewart, 2000.

affected. The possible discrepancies between teacher-reported changes and student outcomes also highlights the fact that changes in input don't always produce changes in output.

Standards have proved controversial because they sometimes go beyond simple educational goals and become embroiled in disputes about the best methods of instruction. Although at first glance developing standards appears a straightforward process, in reality it is difficult and political because of ambiguous goals and disagreement over what makes effective teaching. With diffuse goals, differences of opinion on what and how to teach become the source of intense battles. For example, controversies over math instruction have involved a perceived dichotomy between the importance of knowing basic math operations and the need to have broad conceptual knowledge. Although each is clearly important, various curricula and approaches to mathematics instruction have tended to place more weight on one over the other, leading to conflicts over standards.

Measurement. The biggest controversy in accountability, however, probably surrounds how standards compliance should be measured. Proving that the standards have been met requires some sort of measurement. Assessing compliance requires several decisions: who to measure, what approach to use, how to create valid indices, and, frequently, where to set the critical value or cut-point for meeting the standard.

The centerpiece of current state accountability systems is the testing of student performance. This performance is then aggregated to, say, the school or district level, and some summary of the test scores is made public.

Though obvious, it is important to note first that direct assessment of performance focuses on students. Consistent with the goals and standards related to learning, all fifty states test students. Other influential parties, such as governors, legislatures, parents, and state boards of education, are currently excluded from direct performance measurement, even though they affect the ways schools behave and the

ways students perform.[11] As discussed below, translation of student test results into measurement of performance for other participants is central to the incentives provided by various accountability systems.

A key issue in choosing valid test items is whether the material that is captured by a given standard can be directly translated into a test format. The dilemma lies less in the testability of content than in the feasibility of applying some independent scale of measurement to it. Some crude evidence about the movement in this direction is found from the use of criterion-referenced assessment, assessments that are designed to align closely with the learning standards and curriculum.[12] Forty-five states use criterion-referenced assessment in English, forty-three in mathematics, twenty-three in history/social studies (largely in middle and high schools), and twenty-nine in science.[13] The capacity of the various state criterion-referenced tests to capture existing standards has, however, not been generally assessed.

Another ongoing policy debate involves the mechanics of performance measurement. The mapping of standards to either observable or measurable dimensions necessarily requires abstraction and thus carries a degree of (unknown) error. Current tools used for students and/or teacher testing

[11]Teachers are frequently evaluated in a variety of ways, although this evaluation is seldomly systematically related to student performance and to state accountability systems. Teachers are also frequently tested, but this testing is designed to screen who gets into teaching. Thirty-nine states use tests on content knowledge for beginning teachers. No state has elected to test teachers periodically during their careers. Whether the current preservice testing improves student performance depends on the quality of the test as a predictor of performance, something that remains uncertain. See John M. Goff, *A More Comprehensive Accountability Model*, Washington, DC: Council for Basic Education, 2000.

[12]Criterion-referenced tests are frequently scored in terms of what percentage of the curriculum is mastered by the student. The common alternative is norm-referenced tests, which provide information on how well students do in comparison to a reference group of students and which are not as directly linked to any specific curriculum.

[13]"Quality Counts 2001: A Better Balance," *Education Week*.

include multiple-choice standardized tests, observational studies, expert assessments of portfolios of work, essay or other examples of written work, or short-answer tests. The options involve different tradeoffs in reliability, validity, ease of administration, and cost, but at this point in the evolution of accountability systems, not enough is known of the errors for various types of measurements or their distributional characteristics.[14]

The testing techniques used by states are presented in Table 2. Forty-nine states use standardized tests with multiple-choice format. Fewer states, thirty-eight, add short-answer questions to the testing format. Essays are used primarily for assessing English compositional skills in all but four states. Only two states, Vermont and Kentucky, employ the intensive method of assessing portfolios of student work.

States are not, however, always content with relying exclusively on outcome measures. Many states add in other factors, such as attendance rates (nine states), drop-out rates (fourteen states), or patterns of course enrollment (three states), when assessing the performance of schools. Use of these latter measures appears less directly related to outcome standards than test measures (although they may enter into the diagnosis of what is behind test performance).

Deriving Composite Measures. While most of the public attention has gone to the development of standards and how to measure compliance with them, the use of resulting data, particularly when there are multiple objectives, is equally important. The goal of an accountability system is improving student performance, but performance is the outcome of a variety of factors: student ability and effort,

[14]A final decision in measurement typically is to determine the score that will be treated as the break between passing and failing. The choice is in one sense completely arbitrary; that is, choosing "70 out of 100" as the cut-point is more selective than "60 out of 100" but cannot be related in any systematic way to the underlying measures and metrics. And because it relies on aggregate test information, the choice of a cut-point cannot address any weaknesses or limitations in the underlying measures.

TABLE 2
Choice of Testing Items Used to Assess Students
(Value is number of states using method)

	Multiple Choice	*Short Answer*	*Essay Answer*	*Portfolio of Work*
Elementary School	49	36	44	2
Middle School	49	35	44	2
High School	48	28	43	2

Source: Author's tabulations from *Education Week* (2001).

parental inputs, teacher inputs, and school programs and resources. Even with accurate and reliable data on student performance, the outcome statistics produced must reflect the actions of the relevant people if they are to enter appropriately into performance incentives.

The issue of disentangling underlying elements of performance is most frequently raised in assessing the performance of teachers and schools. If we take accountability down to each of these levels, it is common sense that nobody should be held responsible for bad performance by others. For example, if a teacher starts with low-performing students but does a terrific job of improving their performance, she should not be penalized if the resulting performance level is still lower than, say, the national average. Similarly, a teacher starting with a high-performing group should get credit for her job in improving them but not for their initial preparation. The implication is that any measurement of teacher quality should focus on the teacher's addition, or value added, to student learning—and this requires adjusting the measurement of student performance according to the initial preparation of students. Similar arguments can be made that student accountability should focus on the gains of students after allowing for differences in the quality of teachers.

The best way to separate the different factors that influence student performance is currently unclear. A variety of approaches have been proposed and experimented with in the

states. The most obvious starting measure—applied in virtually every existing accountability system—is the average of all student test scores for a district or a school. This aggregate summary, however, mixes all sources of performance. Other alternatives proposed and used in different places include:

- Annual change in school average score over time
- Average of the mean individual gains in scores
- Average scores of a school relative to state average scores for students of similar background
- Regression adjusted scores to remove individual background differences

The list could be extended, but these illustrate that performance measurement can take many forms. Importantly, as discussed below, these derived measures differ in the degree to which they reveal the contribution of the underlying factors and thus in their value in developing good incentive systems.

These measures also highlight a fundamental tension between the incentives that are created by the way a given accountability system is structured and the overall performance goals they are supposed to promote. For example, for many incentive uses it may be desirable to pinpoint the value added by each school, but even a high-value-added school may start with students sufficiently ill-prepared so that the school does not bring them up to the desired levels of student performance. Looked at from the viewpoint of enforcing high standards of student performance, this school might be judged as falling short—though from the incentive side, this school would deserve praise. This apparently simple issue illustrates the difficulties of using student performance data simultaneously for multiple goals. A common approach is for states to create incentives involving a combination of the level of score and the school change in score over time (such as seen in school reward systems in North Carolina and California). Nonetheless, such approaches, though recognizing a range of measures, may still not align the measurement and incentives appropriately.

Finally, the possibilities for deriving value-added measures relate directly to the choice of measurement approach. Given the current level of inexperience, it is important not to exclude the ability to examine performance from multiple vantage points. For example, as discussed below, the methods that are best suited for tracking teacher and school performance appear to be ones that track the performance change of individual students over time. This approach can control better for ability and background differences across students, which bias simple aggregates that do not consider variations in the cohorts being assessed. But tracking individuals over time cannot be done in systems that use sporadic testing (e.g., those testing only fourth, eighth, and twelfth graders). Moreover, testing regimes that involve portfolios of work, though subject to reliability concerns at any point in time, generally defy consideration of growth in performance over time.[15]

Reporting. Report cards for schools are prepared and published in forty-five states, but the calculations differ widely, making comparisons impossible. In addition, thirty-four states are also producing a district-level report. Two additional states will join the school report card practice in the future, leaving Idaho and Montana the only states that provide no public information on the performance of their education efforts.

To help the public interpret the statistics, seventeen states (with another six planning to) have created aggregate ratings systems of available outcome information and/or input data. Another ten states (with two more in the next few years) use ratings only to identify poor-performing schools. In both practices, however, additional information may be incorporated into the rating, at the state's discretion. Table 3 shows the types of information that states use to rate their schools.

[15]Cf. Daniel Koretz, Brian Stecher, Stephen Klein, Daniel McCaffrey, and Edward Deibert, "Can Portfolios Assess Student Performance and Influence Instruction: The 1991–92 Vermont Experience," CSE Technical Report 371, Rand Institute on Education and Training, 1993.

Many states that incorporate multiple measurements into their ratings do not explain the breakdown, so we are unable to judge which ratings accurately reflect school performance. The lack of computational transparency and consistency could lead to future problems when consequences are attached to performance.

Uses and Consequences.[16] Goals, standards, and measurements create an accountability system. But the mechanics of such a system are largely unrelated to the way states put the results to use.

In most states, accountability systems have multiple objectives—including creating a measuring rod for outcomes, improving school instruction, creating incentives, and creating rewards/punishments for performance.

The standards and accountability movement strives to induce alignment among standards, teaching, and student performance. In contrast to a regulatory approach, the underlying philosophy of accountability is letting the responsible parties maintain control of a process whose outcomes are scrutinized. Consequences—both positive and negative—are the fulcrum that gives leverage to the other players in the education system. If schools or students do not expect any decisive actions as a result of their performance, there is little to motivate attention to the outcomes they produce.

The clearest use of performance standards is to judge student accomplishments. Test scores are used as a graduation requirement in eighteen states (with another six to follow suit in the next three years). Three states use test scores as a promotional criterion from grade to grade. Students with high performance are eligible for scholarships in six states.

The picture with respect to other uses is less clear. Before the movement to greater usage of formal accountability

[16]See the compilation in "Quality Counts 2001: A Better Balance," *Education Week*.

TABLE 3
Information Used by States to Create School Ratings

Source of Information	*Number of States Using Source*
Student test scores only	14
Multiple sources:	
Test scores/drop-out rates	4
Test scores/drop-out rates/other	1
Test scores/attendance	1
Test scores/drop-out rates/attendance	2
Test scores/drop-out rates/attendance/other	7

Source: Author's tabulations from *Education Week* (2001).

systems, many states were accustomed to making judgments about individual school and district performance. All told, 5,613 schools were identified as low performing in the 1999–2000 school year—with many almost certainly making this list primarily because of the average level of student test performance.[17] Many of these, however, were not the result of newly adopted accountability systems. With the advent of new accountability systems and ratings (as shown in Table 3), these judgments are likely to become more systematic.

Creating incentives for schools and teachers is more complicated than creating them for students. Perhaps due to the newness of the policy, accountability systems across the country rely primarily on rewards for good performance or significant improvement. The snapshot of possibilities and actions in the 2001 school year is instructive. For example, twenty states reward schools, and sixteen give teacher bonuses for good performance. Far fewer states, however, impose sanctions. Only fourteen states are authorized to close, reconstitute, or take over a failing school. Of those states, only four have actually followed through with consequences, in a total of seventy schools. Sixteen states are

[17]Ibid.

permitted to replace teachers or principals, but only two cases have been pursued. Just nine states allow students in consistently poor-performing schools to enroll in other schools; widespread court challenges have delayed this option in other states. Clearly, if Florida courts ultimately uphold the A+ program (which permits students in schools that twice receive failing ratings to enroll elsewhere), more schools will likely exercise this option. Eleven states are authorized to revoke accreditation. However, because accreditation can be reinstated with plans to improve—instead of on proven performance—this option is not considered as strong a consequence as others. Only Texas reports using students' test scores to evaluate their teachers.

Incentives and the Application of Accountability Data

The effectiveness of accountability systems rests on three legs: standards, measurement, and consequences. Yet at the most fundamental level, the relationship between these three things is frequently ignored. Consider who is being judged. The most common direct incentives built into state accountability systems revolve around student requirements. As described, about half of the states have consequential test requirements for students, and others are sure to follow. However, few have been binding yet because of phase-in requirements and test experimentation. This aspect of learning has been well documented, although the impact of differing performance requirements on student achievement is less well understood.[18]

[18]The importance of student incentives has been most thoroughly developed by John Bishop (e.g., John Bishop, "Signaling, Incentives, and School Organization in France, the Netherlands, Britain, and United States," in *Improving America's Schools: The Role of Incentives,* Eric A. Hanushek and Dale W. Jorgenson, editors, Washington, DC: National Academy Press, 1996). He argues that external testing leads to significant changes in the motivation of students in their subsequent effort and results. Nonetheless, the best form of such incentives in the context of state accountability systems requires further attention.

But though accountability systems create direct incentives for students, they produce only indirect ones for schools and teachers.[19] Accountability systems work well only if they provide a direct link between outcomes and the behavior of each person in question. Thus, consequences for teachers must be directly related to their effect on student performance. If related to overall levels of student performance, the system would obviously be unfair for teachers who worked with students entering their classrooms with large deficits. They would be punished for something outside of their control. Instead of promoting better performance by teachers, such a system might be expected to have more significant effects on the choices of schools by teachers. Improper measurement can break the link between actions and consequences.

These issues have led to the various approaches delineated, both academic and governmental, to produce reliable estimates of the value added of different schools. The previous discussion of test measures provided a partial listing of the choices currently being made. Nonetheless, even though several states report alternatives and actually issue school rewards based on them, the properties of alternative approaches are not completely understood.

Measurement Accuracy

Some obvious concerns have been raised in previous discussions about accountability, but they have yet to be resolved. For example, the high level of student mobility across schools means that cohort changes over time can have significant effects on measured performance when individual student gains are not considered. For perspective,

[19]Some argue that if students are finding it difficult or impossible to pass the required tests, pressures will be placed on schools that will lead to their improvement. These pressures might be self-generated by school personnel who wish to do a good job, come from school boards and parents, or be the result of Tiebout pressures from school district choices. Little analysis is available to show the strength of such indirect incentives.

in Texas schools, one-third of the students will change schools between grades 4 and 7 (after eliminating all structural moves associated with moving to middle schools from elementary schools). These moves are also more frequent for low-income and minority students.[20] Similarly, mobility of teachers and principals makes it difficult to infer who is responsible for any performance changes of schools over time. For example, average teacher movements in the mid-1990s in Texas show that less than 80 percent of the teachers in any given year remain at the school they were in the prior year.[21] Thus, any simple comparisons of school average scores over time yield ambiguous performance information.

Measurement errors in individual tests can also lead to score changes for small schools over time without being related to any fundamental differences in performance.[22] This presents a dilemma, since error can be reduced by averaging over time, but such averaging makes it difficult to pinpoint any performance changes. And different adjustment methods, such as those previously identified, lead to differing rankings without any clear superiority in terms of true differences in school performance.[23]

These issues also introduce a dilemma. In order to emphasize performance of schools in elevating scores of poor and minority children, states have both required reporting of disaggregated scores and moved to link reward to such distributional information. But scores disaggregated by sub-

[20]Eric A. Hanushek, John F. Kain, and Steve G. Rivkin, "Disruption Versus Tiebout Improvement: The Costs and Benefits of Switching Schools," Cambridge, MA: National Bureau of Economic Research, 2001.

[21]Eric A. Hanushek, John F. Kain, and Steve G. Rivkin, "Why Public Schools Lose Teachers," Cambridge, MA: National Bureau of Economic Research, 2001.

[22]Thomas J. Kane and Douglas O. Staiger, "Improving School Accountability Measures," Cambridge, MA: National Bureau of Economic Research, 2001.

[23]Charles T. Clotfelter and Helen F. Ladd, "Recognizing and Rewarding Success in Public Schools," in *Holding Schools Accountable: Performance-Based Reform in Education*, Helen F. Ladd, editor, Washington, DC: Brookings Institute Press, 1996, pp. 23–63.

populations necessarily involve smaller numbers of students and thus are more subject to random measurement errors. Balancing these requires not only care in the design of incentives but also detailed technical considerations that go beyond just the goals of the system.

A further issue, which extends the previous concerns about separating sources of performance, is the use of derived measures to assess individual teacher performance. The growing databases in states on annual school performance permit measurement of student achievement gains that are directly related to individual teachers.[24] Again, many of the issues raised about school accounting are relevant but more severe here because of the smaller numbers of students involved in the performance measurement.

Using Single-Cut Scores

The strength of incentives is affected by the standards and measurement. It seems natural to many to judge performance as meeting standards or not, that is, to define an acceptable level of knowledge. Obviously, the determination of the passing score is somewhat arbitrary and has a variety of political ramifications. Without going into details about those, the important point here is that differing cutoffs for passing can produce some undesirable incentives. Systems based on raising students over an absolute passing score cause schools and teachers to focus more on students close to the cut-score—because those are the students who can usually be moved across the boundary most easily. At the

[24]Rivkin, Hanushek, and Kain show how teacher quality can be separated from student factors by using panel data on different cohorts of students (Steven G. Rivkin, Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," Cambridge, MA: National Bureau of Economic Research [revised], 2001). The state of Tennessee has actually implemented an alternative approach to identifying individual teacher impacts and uses this in its internal school management. See William L. Sanders and Sandra P. Horn, "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment," *Journal of Personnel Evaluation in Education* 8, 1994: 299–311.

same time, the cutoff weakens the incentive to work with students far below or far above the cutoff.[25] This problem becomes especially acute when considering heterogeneous populations. For such populations, it is very difficult to set absolute cutoffs that don't unfairly penalize disadvantaged and minority students, who more frequently begin with poor performance. At the same time, failing to reward higher-performing students would also render the accountability system ineffective.

Heterogeneous Populations

Inherent in our current organization of education is the assumption that all students can progress at roughly the same pace. The choice of passing score marks the "finish line" that all students are expected to cross in a given year. But the real world presents a different picture. Dealing with passing scores in states or districts with very heterogeneous populations introduces fundamental difficulties of both a political and a conceptual nature. From the political side, there are tensions between having stringent and demanding standards and the need to deal fairly with different populations. It is politically unacceptable to leave disadvantaged or minority students behind, but as currently constructed, many of the accountability systems do just that in their efforts to be challenging to higher-performing students. As long as we continue to expect homogeneous rates of progress in absolute time

[25]The implications of such incentives based on passing scores can be seen from prior work on "performance contracting." In an effort to understand the potential of contracting with private employers to provide remedial education, the Office of Economic Opportunity attempted an experiment. The contract, which provided no payment for any student performing below grade level and a ceiling on the largest payment, led several private providers to ignore both the poorest and best performing students (see Edward M. Gramlich and Patricia P. Koshel, *Educational Performance Contracting,* Washington, DC: The Brookings Institute Press, 1975). More recent observations of this kind are made for Texas by Deere and Strayer (Donald Deere and Wayne Strayer, "Putting Schools to the Test: School Accountability, Incentives, and Behavior," Department of Economics, Texas A&M University, 2001).

frames, we will hinder our ability to address differences in starting points or rates of progress. Just as schools are left to find the best way to achieve the performance measures set for them, we may also eventually consider greater flexibility for students in reaching the performance standards set for them.

The conceptual issues involve the uses and interpretation of the measurements in the accountability system. The system can be used simply to identify levels of student performance and signal to others—employers, colleges, and the like—who is below the cutoff at a given point in time. It can also be used to separately provide incentives for higher performance. As Betts and Costrell demonstrate, these alternate uses lead to some unexpected outcomes when they interact with varying cutoffs in heterogeneous populations.[26] In particular, the different uses can conflict, requiring a clearer delineation of goals and objectives.

One implication of consideration of passing scores is that the binary nature of the scores leads to a set of complications that are avoided by simply providing more detailed information about the distribution of underlying scores, as opposed to relying on just "pass" and "fail." Although such a system does not have the same political appeal, it does permit both information about overall performance and good incentives to students to coexist.

This issue, of course, has different implications when considering accountability based on value added for teachers and schools. The development of passing scores and the building of incentives on them applies most directly to consideration of overall level of scores. If, on the other hand, the system assesses how far a teacher moves a student toward the standards, the cutoffs have less important implications. Concentrating on the reporting of score levels could facilitate the assessment and reward/punishment of teachers, for

[26]Julian R. Betts and Robert M. Costrell, "Incentives and Equity under Standards-Based Reform," in *Brookings Papers on Education Policy: 2001,* Diane Ravitch, editor, Washington, DC: Brookings Institute Press, 2001.

it could track how far a teacher moves a student toward standards. Cutoffs, therefore, would have fewer political dilemmas. One option may be to set fixed standards for diplomas or graduation but permit flexible time frames for meeting them, with the accountability and incentive systems looking at how schools and teachers contribute to progress over time. Nonetheless, in order to provide incentives in different parts of the performance distribution, some sort of "enhanced diploma" would still be useful.

Some Issues of Implementation

The newness of strong accountability systems leaves individual states to make guesses about what is best. It also opens up larger political problems.

Feasibility. The political nature of the standards and accountability process leads to huge tensions. No state wishes to be known for setting standards that are too low or that can be construed as not challenging. On the other hand, standards that are too high become infeasible—and could involve serious harm, depending on the consequences for not meeting them.

Consider the actions of the state of New York. In 1999, the Board of Regents decided that it should do away with lower levels of diplomas and require all students to obtain its premier diploma, the Regents diploma. The Regents diploma requires passing a series of rigorous subject-area examinations that are linked to a difficult underlying curriculum. At the time of development of this standard, some 40 percent of graduating high school students in the state obtained Regents diplomas. Twenty-one percent of graduating students in New York City obtained a Regents diploma. Simply mandating that all students move to the new standard is likely to leave many who previously would have received some sort of diploma without a diploma—arguably a very harmful situation. The hope of the new standard is that it would lead

students to work harder and would lead schools to do a better job. On the other hand, it also looks generally infeasible under the new standard for the school systems in many parts of New York State, most particularly for New York City, to obtain graduation rates close to those previously achieved.

One response, followed by New York State, is to stretch out the time period before the standards are applied. Thus, though they originally were to be operative today, the phase-in period has been extended into the future. Whether this will permit full phase-in depends on how well school systems can respond (i.e., on whether the goals move closer to being feasible). Currently, this possibility is unclear.

The Need for Both Rewards and Sanctions. The incentives that derive from the design and use of accountability systems work only to the extent that they motivate students, teachers, and schools to examine their performance and make changes to improve, if necessary. Without consequences, incentives disappear. But people will react to consequences differently. Some are motivated positively by bonuses, whereas others find them offensive. Some people, but not all, are only moved by the wish to avoid negative consequences. Relying solely on rewards may not be sufficient to overcome the inertia of habit; but likewise, the existence of only sanctions can demoralize and undermine sustained effort. This suggests that accountability systems should acknowledge different patterns of motivation and incorporate these differences into the design. Having both negative and positive consequences creates avoidance incentives and attraction incentives and can address more fully the range of motivations. As the previous state data suggest, nonetheless, most school incentives are currently heavily weighted toward rewards.

Testing the Premises. At the outset, it is important to recognize that there is little experience in the design and

operation of educational accountability systems and their elements. In many ways, this does not differ from many other educational policies that are introduced more on superficial plausibility than on any evidence. As an example, there is uncertainty about how schools and teachers react to the incentives introduced. If the implicit weights in the incentive system favor a certain set of subjects at the expense of others, does it lead to undesirable distortions in the balance of teaching? Does the incentive structure lead to cooperation among the teachers? Does it change the amount of teacher turnover?

One implication of this is that states must be prepared to review and revise as experience reveals better information about the underlying linkages. But the ways that states go about these important steps introduce potential problems that need mentioning.

There is a distinct trade-off between adjusting incentives and maintaining a strong set of incentives. School personnel today are accustomed to frequently changing programs and perspectives of schools, leading to some cynicism about the staying power of any innovation. Accountability systems also face unique problems of adjustments going beyond those of normal programs. Many of the potential adjustments that are feasible are long-term responses—reflecting better selection and motivation of teachers, improved student effort, better matching of students and programs, and the like. In order for incentives to elicit these long-term impacts, the participants must believe that the incentives will remain in place over the long term. But balanced against this is the difficulty in designing incentive structures given our current knowledge.

We do not know much about how best to accumulate knowledge or even about which directions schools might take to improve. Some hope comes from other states. With so many states launching these systems at about the same time, states can look to each other to learn from their similarities and differences without each needing to continuously vary their own design.

Ensuring That Accountability Systems Effect Change

Once an accountability system is in place, there are still many things that need to be done to ensure it actually produces change. Some of these things, which bring up further unresolved issues, are discussed below.

Dealing with Poor Performance. Consequences for poor performance are far from straightforward. What many fail to realize in the context of accountability systems is that appropriate consequences vary according to the causes of bad performance. For example, if deficient student background is the determining factor, simply increasing school resources may be the answer; if, on the other hand, poor performance is due to poor teaching, a different solution is required. Many people tend to assume that all poor performance by students is either one or the other: poor preparation that must be compensated for or bad teaching and school management. In reality, observed poor performance almost certainly has elements of each, if not in individual schools, at least across different schools. Accountability schemes must not continue to ignore this fundamental issue, for it has important consequences for program design.

The correct answer requires sufficient evidence to distinguish the causes of poor performance. Though this is largely an implication of prior discussions about aligning results with the people responsible for each of the components, it has obvious importance to overall design issues. And the current systems have not been demonstrated to be effective at this.

Incentives and Efficiency. To return to the motivation for accountability systems in the first place, namely that student performance has remained flat while resources have grown dramatically, one of the unanswered questions of the new accountability systems is the degree to which they yield efficiency gains through the incentives they create. Will incentives to improve student achievement outputs naturally lead to better use of resources? The answer is not obvious.

Although a simple version would be that schools redirect their resources to the places of highest payoffs, this cannot be assumed. For example, if the largest incentive and impact of incentives comes through student effort, there might be little impact on efficiency of resource usage. Or if the direct incentives for teachers and school personnel are less than the value they put on current resource usage, there might be little impact on efficiency. This latter case could arise, say, where teachers take extra resources in terms of greater free time and where the individual benefit of any incentive reward is less than their valuation of the free time.

Again, little is known about any collateral impact of accountability structures and their resulting incentives on the efficiency of resource usage. The impact will clearly vary with the magnitude of incentives, the ease of achieving desired outputs, and the alternative uses of resources.

Knowing Performance Is Poor Is Different from Knowing What to Do. Even though the purpose of accountability systems is to improve student learning, how exactly to achieve this faces several dilemmas. Accountability systems identify when things are not working well but not what corrective actions are required. In fact, an accountability system operates on the assumption that assuring good performance is something that cannot be regulated at the state level—for different schools require different solutions. The situation is further compounded by the fact that teacher quality studies suggest that the key to effective improvement may be changes in personnel, as opposed to the programmatic fixes currently focused on training and support of current teachers.[27]

The dilemma is clear. Though some schools may know how to solve their problems, others may have no idea, and past research has produced no clear indication of what precisely

[27]See, for example, Rivkin, Hanushek, and Kain, "Teachers, Schools, and Academic Achievement," Cambridge, MA: National Bureau of Economic Research, 2001.

helps students learn. Continuing research into the determinants of performance may be part of the answer, but so far such research has yet to be successful, and it is unlikely to provide any immediate guidance. This inherent and potentially serious weakness must be recognized.

Thus, a key element of the move to direct accountability is a presumption that local people, with incentives and motivation, will be best positioned to improve student outcomes. Clearly, this presumption needs to be judged over time. Evaluating this presumption should be a top priority of accountability systems.

External Validity. One of the largest issues facing accountability systems is also one of the most basic. Ideally, tests and incentives should align with a school's learning objectives. At the same time, a system that is not geared to ultimate users—higher education and the job market—cannot be very productive.

All current testing is focused on meeting an initial set of standards that are assumed to reflect the set of knowledge that adequately prepares students for their postschooling years. There is surprisingly little attempt to match this with subsequent performance. The research on this is also quite thin. There is increasing research suggesting that performance on cognitive tests is strongly related to labor market earnings, but this research has not been very careful in distinguishing among alternative performance measures (and their underlying standards of knowledge).

Conclusions

Within the past quarter of a century, the desire to improve student performance has caused policymakers to focus directly on student achievement. Although prior policy has focused almost exclusively on what's going into the educational system, recent reforms have shifted the focus to what's coming out—what we want students to know and how we can be sure they know it. The accountability systems

now being put in place are an attempt to ensure that the goals for student knowledge are actually accomplished. The simple structure of current accountability systems, however, masks how its elements interact in a complex fashion that can produce unexpected outcomes.

Current accountability systems revolve around measured student performance, even though student performance is influenced not only by students but also by parents, teachers, and schools. Concentrating on student performance is a very important and positive change in how we view schools. Nonetheless, although the movement toward performance-based systems offers the best chance for improvement, the journey has just begun.[28] The focus on improving outcomes should be applied with equal rigor to educational performance and to the accountability systems themselves. The challenge is harnessing this fundamental movement to bring about the desired changes.

[28]Eric A. Hanushek and others, *Making Schools Work: Improving Performance and Controlling Costs,* Washington, DC: Brookings Institute Press, 1994.