National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

The Confusing World of Educational Accountability

Eric A. Hanushek and Margaret E. Raymond
Hoover Institution, Stanford University, Stanford, CA 94305-6010

National Tax Journal
Vol. LIV, No. 2

INTRODUCTION

Accountability has been a watchword in education for decades—for who could be against it? It has not been a reality, however, because accountability is threatening to 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. While many states and districts are moving forward with accountability schemes, they are likely to run into real problems with design or interpretation that compromise and distort their impact. In fact, while it seems natural to measure outcomes and hold schools responsible for them, the reality is much more complicated. Achieving the beneficial effects of accountability and performance incentive schemes will require a deeper understanding of the dynamics of accountability in public schools.[1]

Considerable controversy also accompanies accountability in schools. Parents, teachers, policy makers, 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 student performance 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 origins of today's movement to enhanced accountability systems lie in the historical performance of the U.S. educational system, which is briefly reviewed to identify the policy environment. The structure and function of accountability systems used in the states are then described. Following that, there is a discussion of the difficulties with the current implementation and with concerns about the potential impacts.

A fundamental perspective of this discussion is, however, the very large potential benefits that are likely to accrue from the focus on student performance. The change to public con-

[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.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

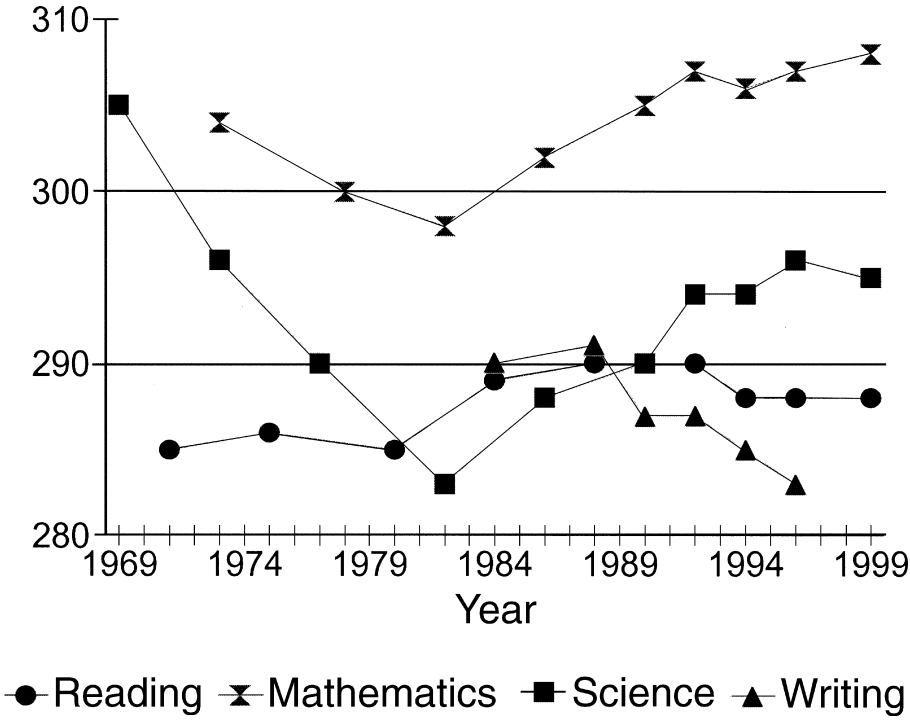
cern and attention to student outcomes is a major improvement in the area of educational policy. The problems identified in this paper are offered as contributions to improve on the important strides that have already occurred.

THE STATE OF U.S. EDUCATION

Understanding the dynamics of the U.S. education system is important both for motivating recent attention to accountability and for understanding the issues facing policy makers. The simplest description is that the performance of United States students has remained stagnant while the costs have been increasing steadily.

The performance of U.S. education is best traced by considering the performance on examinations given under the auspices of the National Assessment of Educational Progress, or NAEP. While our state–centered educational policy has precluded both national standards and national tests, the periodic NAEP exams in different grades and in different subject areas record performance of a random sample of U.S. students over the past three decades. A simple summary of performance of 17–year–olds over time on the NAEP is provided in Figure 1. This figure tracks average scores in reading, math, science, and writing.[2] The story is one of flat achievement. Reading and math scores are

Figure 1. National Assessment of Educational Progress—17 Year–Olds



[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.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

slightly higher at the end of three decades, while 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 scores and rankings on international tests, where the United States has placed at or below the middle of the international distribution since the first tests in the 1960s.[3] Second, the U.S. performance has not been for want of trying. As Table 1 shows, school resources have risen consistently over the relevant time period. Real spending per student more than tripled between 1960 and 1995, and this growth was driven by many of the arguments traditionally made for educational improvement—reduced pupil–teacher ratios and improved teacher quality in terms of educational credentials and experience.[4]

The dominant approach to policy making over much of this period has been regulation of inputs and educational process. Efforts were concentrated on the level of resources provided to schools and to specific programs and processes. This approach has been especially appealing to legislatures and courts—the places where overall fiscal decisions tend to be made—because of the ease of setting resource policy. As shown, however, in the face of substantial aggregate increases in resources, little evidence suggests that student performance has increased. Similar results are found in detailed analyses of performance across classrooms and schools (Hanushek, 1997). The few available analyses of the distribution of student performance after changes in funding distributions required by courts also have shown little evidence of narrowed variation in student results (Downes, 1992; Hanushek and Somers, 2001).

This backdrop has contributed directly to movement toward stronger accountability in schools. Because of concerns about results, attention to student achievement has gained momentum. This attention has actually been manifested in a variety of forms, discussed below, but a common theme of regulation of outcomes rather than the more traditional regulation of inputs follows the general trend of regulatory reform witnessed in other policy fields. In addition to an outcomes orientation, newer regulatory frameworks leave the means of production to the discretion of the producer, but hold them accountable for achievement. Perfor-

TABLE 1
PUBLIC SCHOOL RESOURCES IN THE UNITED STATES, 1960–95

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

[3] 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 Hanushek and Kimko (2000).

[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 (Rothstein and Miles, 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 Hanushek and Rivkin, 1997).

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

mance is monitored and communicated publicly, and the information about performance thus serves as a driver of innovation and/or competition.

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 meeting with President George Bush in 1989. As a result of this meeting, a commitment was made to a set of national educational goals. These goals included such things as "the United States should be first in the world in science and math performance by 2000." While this bit of wishful thinking later was belied by international test scores, it nonetheless underscored the movement toward measurable goals based on student outcomes.[5]

The Goals 2000 ideas blended into what is today perhaps the most acclaimed path to educational improvement: so called "standards based reform." The heart of standards based reform rests on educational goals and measured progress toward them. By these arguments, public disclosure of both provides the most reliable path to achieving them. Within this general idea, however, there are many variations and interpretations—including simultaneously setting standards for inputs or for programs and processes.

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. A basic idea behind the current school reform movement is that by measuring and reporting student achievement—particularly in reference to agreed upon outcome standards—the relevant actors will improve their performance.

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. Their 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. The bad performance history—now widely recognized—has highlighted the importance of improving student outcomes and of using resources better. Recent history has shown, moreover, that we do not know how to link programs, resources, and other inputs to student outcomes. Accordingly, regulation of inputs cannot be assumed to satisfy outcome goals. The 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. By any measure, this change in focus is an enormous stride in policy deliberations.

The resulting movement toward accountability in education has developed a clear focus on outcomes. Specifically, through testing of student performance, 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

[5] 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.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

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. Another part comes from the development of explicit incentives that will lead to innovation, efficiency, and fixes to any observed performance problems.

States have not, however, entirely bought into an exclusive focus on outcomes. Their long histories of input regulations have some carryover to current systems. Moreover, without knowledge of what does and does not contribute to outcomes, states will have a hard time diagnosing the causes of poor performance and suggesting the correct solutions; hence more rigorous attention to inputs and processes will need to accompany investments in accountability systems.

Our question is simply, "given what states are doing, what more is needed to get there?"

CURRENT PRACTICE[6]

The basic skeleton of accountability systems involves goals, standards for performance, measurement, and consequences for varying levels of performance. While states differ in significant ways, a general description of the structure of these systems is useful to support comparison of the actual plans and the ways their elements interact.

The actual details of each element have much to say about the usefulness and success of any system writ large. We introduce key elements of each aspect of the stylized accountability system, along with flagging some of the most serious design issues within each. Following that, we move to larger questions that currently are unanswered.

Goals

An accountability system begins with a set of goals about what is to be accomplished by the accountability system. While 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 measurement typically emanate from them. Since most states create their accountability practices through statute, the best opportunity for careful delineation of a state's system purpose and goals is conceptually found in the enabling legislation. Nonetheless, few states set unambiguous and measurable goals for their accountability systems. While the lofty and vague goals may be required to ensure legislative approval, such vagaries can hobble the functionality of the system by leaving real ambiguity about what is to be done by whom.

The goals and the subsequent implementation of the accountability system are typically notable for their focus of attention—whether on students, schools, or teachers. Each of these groups feels targeted, regardless of the actual mechanics—although the degree of attention to each differs significantly across states. On the surface it may seem desirable to derive multiple uses from a single system, but, as discussed below, there is a serious risk of dilution in the overall utility of the accountability system if it is built to satisfy multiple and diverging expectations without care to deal with each individually.

Standards

Standards typically present the details of what is expected. They create boundaries or domains for attention. Recent focus concentrates on "outcome standards,"

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

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

but many states also retain their historical standards pertaining to resources and processes. The typical student outcome standards delineate the extent to which students should have demonstrated 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: 36 states have standards in English or language arts, 44 in Mathematics, 43 in Science and 27 in Social Studies. Many researchers identify this movement to explicit measurement of student performance as the key element of current school reforms (cf. Elmore, Abelman, and Fuhrman, 1996).

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, which is discussed below. Tensions also exist about the relationship between standards and learning goals; one continuing example is seen in the arguments, on the one hand, that standards need to be rigorous and, on the other, that standards do not provide adequate maps for higher–order material or reasoning.

Standards are introduced in order to change behavior. The current standards based reform is explicit in the view that the development of standards will lead to better performance that can accomplish the standards. As a snapshot of the interim effect of adopting standards, a national survey asked teachers if they had altered their classroom behavior (Belden, Russonello, and Stewart, 2000). The majority of respondents indicated that the standards have necessitated a more challenging curriculum and greater attention to material that is aligned with the standards in their state. This internal view has, however, yet to be matched by consistent evidence that student performance has been affected. The possible discrepancies of teacher reports of changes and student outcomes also highlight the difference between output standards and input or process standards.

Standards have also proved controversial because they become intertwined with the goals of the schools and with the methods of instruction. With diffuse goals, differences of opinion on what and how to teach become the source of intense battles. For example, controversies over the instruction of mathematics have revolved around the importance of knowing basic math operations versus the need to have broad conceptual skills. While each is clearly important, various curricula and approaches to mathematics instruction have placed more weight on one than the other, leading to conflicts over standards. At one level, developing clear standards of what is to be known in various areas appears to be a straightforward issue amenable to professional judgment. At another, it has proved difficult and political because of poorly articulated educational goals and because of current uncertainties about the ultimate effectiveness of different approaches to teaching and learning.

Measurement

The largest portion of the conflict about statewide accountability systems surrounds the tools and means used to determine compliance with standards. Proving that the standards have been met requires some sort of measurement. This in turn requires several decisions: who to measure; what approach to use; how to create useful metrics for the dimensions of interest; and, frequently, where to set the critical value or cut–point for meeting the standard.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

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 variant of the test scores is made public.

Note that direct assessment of performance focuses on students. Consistent with the goals and standards related to learning, all 50 states test students. Other influential parties, such as governors, legislatures, parents, and state boards of educations, are currently excluded from the focus of measurement, even though they materially affect the ways schools behave and the ways students perform (Goff, 2000). Teachers are frequently tested, but this testing is designed to screen who gets into teaching and not test any elements of actual classroom performance.[7] Whether such testing furthers improvement of student performance depends on the quality of the test as a predictor of performance, something that remains uncertain.

A key focus has been construct validity, that is, whether the metrics bear a direct relationship to the material that the standards they intend to capture. Some 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.[8] Forty–five states used criterion–referenced assessment in English, 43 in Mathematics, 23 in History/Social studies (largely in middle and high schools) and 29 in Science. The quality of various state criterion–referenced in capturing 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 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.[9]

The testing techniques used by states are presented in Table 2. Every state uses at least one technique to assess students. Forty–nine states use standardized tests with a multiple choice format. Fewer states, 38, 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.

In addition, many states add in other factors such as attendance rates (nine states), drop–out rates (14 states), or patterns of course enrollment (three states) when assessing the performance of

[7] Thirty–nine states use tests for beginning teachers on content knowledge. No state has elected to test teachers periodically during their careers.

[8] 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.

[9] 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 cutpoint 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 cut–point cannot address any weaknesses or limitations in the underlying measures.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

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).

schools. Use of these latter measures appears less directly related to standards than test scores.

Derivative Measures

While most of the public attention has gone to the development of standards and of their related measures, the use of student achievement data, particularly when there are multiple objectives, is an equally important issue. The goal of the accountability system is invariably improvement of student performance, but we typically think of student achievement as representing the combined outcome of student ability and effort, of parental inputs, of teacher inputs, and of school programs and resources. Even with accurate and reliable data on student performance, the outcome statistics produced must reflect the actions of the actors.

The issue is most frequently raised in assessing the performance of teachers and schools. If we take accountability down to each of these actors, it is straightforward that none should be held responsible for bad performance by others. For example, if a teacher happens to face a class of ill–prepared students but does a terrific job of improving their performance, she should not be penalized for their initial level of preparation. Similarly, a teacher who faces a very well prepared group should get credit for her job in improving them but not for their initial preparation. The implication is that any accountability and incentives applied to teachers should focus on the teacher's addition to student learning—and would require adjusting the level of student performance for the preparation of students. Similar arguments can be made that student accountability should focus on the gains of students after allowing for any differences in the value–added of teachers.

The best way of attributing student performance to the contributions of different actors or resources is unclear. A variety of alternative approaches have been proposed and experimented with by the states. The most obvious starting measure —applied in virtually every existing accountability system—is the simple mean of all student test scores for a district or a school. This aggregate summary, however, mixes all sources of performance. Other alternatives, which are found in state reports and in academic research, include:

- Annual change in school average score (AC);
- Average of the mean individual gains in scores (AG);
- Average scores of a school relative to state average scores for students of similar background (RA);
- Regression adjusted scores to remove individual background differences (R).

The list could be extended, but these illustrate that measures of value–added of schools take many different forms with the important implication that they differ in the extent to which they reveal the underlying causes of the observed outcomes. The fundamental quest of accountability systems based on output standards often

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

introduces some conflicts: how to provide a clear view of the effect of schools on performance while maintaining desirable overall achievement goals. A very common approach in how states create incentives involves a combination of the level of score and the school change in score over time. Variations in this are, for example, employed in school reward systems in North Carolina and California. Unfortunately, as discussed below, the properties of alternative adjustment systems are not very well understood currently, although it appears that linkages to student achievement growth produce more reliable assessments (Clotfelter and Ladd, 1996; Rivkin, Hanushek, and Kain, 2001).

Finally, the ability to derive such alternative measures of performance of schools and teachers interacts with the measurement approach. The methods that appear to have superior properties involve tracking the performance changes of individual students over time. This approach can control better for ability and background differences across students that bias simple aggregates, which do not consider variations in the cohorts being assessed. Tracking individuals over time, however, 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, while subject to reliability concerns at any point in time, generally defy consideration of growth in performance over time (cf., Koretz et. al., 1993).

As discussed below, state accountability systems offer real promise for improving U.S. education. An important issue in releasing this promise is ensuring that the data of the accountability systems provide the correct signals to drive incentives.

Reporting

Report cards for schools are prepared and published in 45 states, but the calculations differ widely, making comparisons impossible.[10] In addition, 34 states also produce 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 assist the public's understanding of what the specific statistics mean, 17 states (with another six in the future) create ratings systems. 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.

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).

[10] Gormley and Weimer (1999) provide a useful analysis of alternative public reporting schemes across a variety of areas of government.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

For the states using multiple measures in their ratings, many do not reveal the exact way in which they are incorporated, so we are unable to determine the extent to which the measures produce reliable distributions that accurately correlate with school performance. The lack of computational transparency and consistency is troubling and can potentially lead to problems when consequences are attached to measured performance.

Uses and Consequences

The educational goals for student learning, combined with the standards and measurement of performance, provide the underlying performance data entering into accountability systems. However, they do not identify the uses of the accountability system.

In most states, the accountability system actually has multiple objectives—including, frequently, the creation of a measuring rod for the current outcomes, the improvement of instruction in the schools, the provision of incentives for students, schools, and teachers, and the basis for various rewards and punishments related to outcomes.

The very core of the standards and accountability movement is to induce alignment between standards, teaching and performance of students. In contrast to a regulatory approach, the underlying philosophy in the accountability approach is to yield control of the process, but to focus carefully and meaningfully on outcomes. Consequences—both positive and negative—are the fulcrum that affects leverage throughout the other parts of 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.

Consequences applied to individual students are weighted toward sanctions for substandard performance. Test scores are used as a graduation requirement in 18 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 application of any direct incentives to schools and teachers is more complex and varied than the student component. In judging schools, states have made judgments about individual school and district performance over many years. With the advent of new accountability systems and ratings (as shown in Table 3), these judgments are likely to become more systematic. 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.

At least currently, the consequences that schools face primarily incorporate rewards for good performance or significant improvement. Schools are directly rewarded in 20 states, including 16 who permit the use of reward funds to flow directly to teachers in the form of bonuses. Far fewer states report the option to impose sanctions, and even fewer states report using them if available. Only 14 states are authorized to close, reconstitute or takeover a failing school. Of those states, only four have used their option for intervention and only in 70 schools in total. Sixteen states are permitted to replace teachers or principals, but only two cases have been pursued. Only nine states permit students in pervasively poor–performing schools to enroll in other schools; widespread court challenges have delayed this option from being applied. Clearly, with the recent Florida court decision to uphold the A+ program (which permits students in schools that twice receive failing ratings under the state accountability system to

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

enroll elsewhere), the number of schools affected is likely to rise. Eleven states are authorized to revoke accreditation. However, since accreditation can be reinstated contingent on plans to improve, not on proven performance, this option cannot be considered as strong a consequence as others.

Only Texas reports using the student test scores to evaluate their teachers.

INTERNAL CONSISTENCY AND RATIONALITY

Accountability systems have swept into the states on a wave of public and political enthusiasm, but the details of current systems are not entirely understood. It will be important that a variety of the essential mechanics are refined by on–going experience. Some of the larger issues are likely to command considerable attention as accountability systems become more central to school operations.

Accountability systems rest on three legs: standards, measurement, and consequences. Yet at the most fundamental level the relationship among these is frequently ignored. Consider a simple model of student achievement such as:

$$[1] \quad A_{it}^{s} = \sum_{\tau} X_{i\tau}^{p} + X_{it}^{s} + E_{it} + \sum_{\tau} S_{it}^{p} + S_{it}^{s} + \varepsilon_{it}.$$

At any point in time, t, the achievement of student i in school s is determined by a series of nonschool factors, X, by the student's and parents' effort, E, by school inputs, S, and by a random error, ε. Importantly, achievement at any point in time, say 12th grade, reflects past inputs of school and nonschool factors—indicated by the summations across prior years, τ, and these prior experiences may have taken place in other schools. This simple relationship is meant to characterize the various components of achievement and to provide a clear way to see how various accountability systems relate to them.

Now consider the use of measured achievement for accountability and incentive purposes. The most common direct incentives built into state accountability systems revolve around student requirements. As described, about half of the states have test requirements for graduation of students on the books, and others are sure to follow. Few have been binding yet because of phase–in requirements and experimentation with tests and cut–offs, even though they have been the focus of much attention. Clearly, students and their parents have a substantial influence over performance.[11] This aspect of learning has been well documented, although the impact of differing performance requirements on student achievement is less well understood.[12]

The common approach to student incentives is to establish a cut–off score for measured achievement (A). If one simply wishes to provide information to on performance—say, to colleges or to employers—this approach works.[13] On the other hand, one purpose of such accountability is providing incentives, i.e., changing the level of effort, E, in equation [1] (Bishop, 1996). However, as is clear from equation [1], the level of score is only partially related to E and depends on the current and

[11] The close relationship of family background to student achievement was vividly driven home by the "Coleman Report" (Coleman et. al., 1966), so much so that some people interpreted that report as indicating that families were all that really mattered. Subsequent research has show substantial effects of teachers and schools (e.g., Hanushek, 1992 or Rivkin, Hanushek, and Kain, 2001).

[12] The importance of student incentives has been most thoroughly developed by John Bishop (e.g., Bishop, 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.

[13] More information would of course be provided by releasing the levels of scores and not just whether or not the score exceeds a given cut–off.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

past actions of parents and schools. Thus, incentives are muted, particularly if feasible levels of effort for an individual will not overcome deficiencies in achievement from the other sources.

This concern is part of a larger problem of linking incentives to effects. Some argue that, if students are finding it difficult or impossible to pass the required tests, pressures will be placed on school boards that will lead to their improvement. These pressures might be self–generated by school personnel who wish to do a good job, might come from school boards and parents, or might be the result of Tiebout pressures from school district choices. But again, as shown by equation [1], the current school personnel are just part of the achievement of each student.

A general principle is that accountability systems work best if they provide a direct link between outcomes and the behavior of each actor. Thus, consequences for teachers must be directly related to their contribution to the value–added outcomes of their students. If related to overall levels of student performance, the system would obviously be unfair for teachers who worked with students entering their classrooms with the largest deficits. They would be punished for something outside of their control. And, the rational teacher might well worry more about student selection than about teaching. One implication is that improper measurement can break the link between actions and consequences. Another is that measurement and reporting should be able to identify and reward the contributions of each participant.

These issues have led to the various approaches delineated, both academic and governmental, to produce reliable estimates of the value added of different actors. The previous discussion of derived 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 well understood.

Putting the common measures in terms of the components of achievement in equation [1] provides a way of assessing each:

[2] *average change* = $AC = \overline{A_t^s} - \overline{A_{t-1}^s}$

[3] *average gain* = $AG = \overline{(A_{it}^s - A_{i,t-1}^s)}$

[4] *relative average* = $RA = \overline{A_t^s} - \overline{\overline{A_t^{SES}}}$

[5] *regression adjusted* = $R = \overline{A_t^s} - \overline{X_t^s} b$.

The standard incentive argument is that incentives for schools (and teachers) should be directly related to S_t^s, the contribution of schools in equation [1]. Each of these common adjustments attempts to move the simple school average achievement closer to S.

The average change in performance, AC, compares two successive cohorts in a school. If the comparisons are made for a single grade in the school (say, the sixth grade), AC will include any average differences in cumulative nonschool factors, ΣX, for the two cohorts. In large schools with a stable population, these differences may not be too large—small and changing schools, however, can be problematic. Additionally, AC includes any differences in the cumulative school inputs. This is again not a large problem if there is limited student mobility, but increases in importance with the amount of student school changing.

A variant on AC would track the average score of specific cohorts across grades. For example, if would be possible to calculate the average fourth grade achievement in 1998 and compare that to the average fifth grade achievement in 1999. If the cohorts do not change, this would effectively eliminate differences in cumulative school and nonschool backgrounds

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

[$\Delta(\Sigma X)$ and $\Delta(\Sigma S)$] and leave only differences in effort across grades [$\Delta\Sigma$] and differences in average error terms [$\Delta\varepsilon$].

This modified approach brings the achievement measure close to the impact of schools. Yet, the left out factors raise some concerns. First, 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 example, in Texas, approximately 20 percent of the elementary and middle school students change schools each year (not counting normal structural moves). Moreover, the propensity to move is related to student background factors, with disadvantaged students being more likely to move in a year (Hanushek, Rivkin, and Kain, 2001). Thus, any biases in this measure are likely to be larger for schools serving disadvantaged populations. Similarly, mobility of teachers and principals makes it difficult to infer who is responsible for any performance changes of schools over time.

One particular concern has been differences in cohorts generated by the schools by their test taking policies. If schools exert some control over who takes tests (through formal or informal exclusion rules), they can manipulate the test scores to some extent. Note, however, that simple exclusion schemes probably have little long run effect on derived scores that operate on continuous change measures, because the gains of one year typically become problems for the second year.

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 (Kane and Staiger, 2001). This presents a dilemma, since error can be reduced by averaging over time but such averaging makes it difficult to pinpoint any performance changes. The possibility of such problems suggests particular care in the treatment of small schools.

The other approaches are also subject to question and refinement. The average gain for individual students, *AG*, perfectly eliminates differences in school and nonschool background factors. In the ideal world, students who move would be followed from their prior schools, so that everybody could be included. In the less than ideal world where this is not possible, the measures of school performance would be related just to the stable population. This selection could leave out significant elements of school performance.

The two adjusted measures (*RA* and *R*) have similar properties. The comparison to similar schools is designed to provide a performance standard based on achievement of schools in the state with a "similar" student body in terms of socio–economic background. While varying in details, the basic approach is to divide schools based on family characteristics and to compare schools within a broad category (e.g., bottom quintile). The regression adjustment approach (*R*) typically expands on this idea by estimating the expected score for a school with a given *SES* background (based on estimated parameters, β).[14] The difficulty with both of these approaches is that they begin with a comparison standard that adjusts not only for nonschool factors (*X*) but also partly for school differences (*S*). These approaches look at the relationship of achievement, *A*, and *X*, but equation [1] indicates that *A* is determined by additional systematic factors. If, for example, *X* and *S* are correlated within the state, either of the adjustment approaches will partially include school differences—pushing the accountability measure away from the desired measure.

[14] This approach typically regresses average school achievement on contemporaneous values of family background (X), ignoring both historic values and school factors.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

Different adjustment methods, such as those previously identified, lead to differing rankings (Clotfelter and Ladd, 1996). Some appear to be superior on a conceptual level—such as calculating average gains for individual students—but not much is known currently about the ways in which the alternative methods create or distort incentives for schools and teachers. Further investigation of the alternatives is an obvious priority.

A further issue, which extends the previous concerns about separating sources of performance, is the use of specialized 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.[15] Many of the issues raised about school accounting are relevant because teacher effects must still be separated from those of individual students. Nonetheless, because these approaches can more effectively use information about students and teachers over time, they offer considerable promise.[16]

The level of incentives is affected by the standards and measurement. It seems natural to many to judge performance as meeting standards or not, i.e., 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. For a school or teacher, incentives based on raising students over passing scores leads to stronger attention to students close to the cut–score—because those are the students generally most easily moved across the boundary. At the same time, it provides weaker incentives to work with students far below or far above the cutoff.[17] The problems of differential incentives become especially acute when considering heterogeneous populations. It is very difficult to set cutoffs for passing that don't leave large groups far away when the underlying student performance is very heterogeneous.

Dealing with passing scores in 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 dealing with different populations. In particular, it would be unfair to develop standards that implied large negative consequences for disadvantaged students and minority students, who have on average performed poorly on tests such as the NAEP. However, it also would severely constrain the accountability and incentive system if it had no effect on higher performing students. The conceptual issues involve the uses and interpretation of the accountability system. The system can be used simply to identify student performance and signal to others— employ-

[15] Rivkin, Hanushek, and Kain (2001) show how teacher quality can be separated from student factors by using panel data on different cohorts of students. The State of Tennessee has actually implemented an alternative approach to identifying individual teacher impacts and uses this in its internal school management (Sanders and Horn, 1994).

[16] The issue of error variance (from ε in equation [1]) comes up because classrooms generally involve relatively few students, making the problems of small samples relevant. However, by combining information across years for individual teachers this issue is lessened.

[17] 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 gaining less than one grade in achievement and a ceiling on the largest payment, led several private providers to ignore both the poorest and best performing students (Gramlich and Koshel, 1975).

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

ers, colleges, and the like—who is below the chosen cutoff for knowledge. It can also be used to provide incentives for higher performance. As Betts and Costrell (2001) demonstrate, these alternate uses lead to some unexpected outcomes when they interact with varying cutoffs in heterogeneous populations.

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. Continuous information about performance against standards permits alternative weighting of scores and allows for incentives across a wider range of the distribution.

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. One option may be to set fixed standards for diplomas or graduation but permit flexible timeframes for meeting them, looking instead at value–added progress over time.

OTHER ISSUES WITH CURRENT SYSTEMS

The move toward more complete accountability systems has introduced a wide variety of approaches. Some of the early examples introduce issues that bear on their potential effectiveness.

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 was 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 standards 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 to avoid substantial injury to individual students for the school systems in many parts of New York State, most particularly for New York City.

The primary problem in this situation is the significant heterogeneity of the population. It is difficult to devise passing standards that work well for all parts of the distribution. If meeting standards is entirely infeasible, the incentive effects are negligible.

One response, followed by New York State, is to stretch out the time period before the standards are applied. Thus, while 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.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

Symmetry of Awards and Sanctions

In moving from input regulations to output standards, students, schools, and teachers have clearly defined expectations for performance. Experience in other industries suggests, however, that to motivate behavior toward the desired target both positive and negative consequences must be defined .[18] The lure of reward may not be sufficient to overcome the inertia of habit, but likewise, the existence of only sanctions can demoralize and undermine sustained effort. Since people react differently, the designers to these systems should anticipate different patterns of behavior and incorporate these differences into the design. Having symmetry in the range of consequences creates avoidance incentives and attraction incentives. In this way the full range of current motivations is addressed. 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. One implication of this is that states must be prepared to review and revise as experience reveals better information about the underlying linkages.

In the simplest example, there is uncertainty about how schools and teachers react to the incentives introduced. For example, 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? There is a distinct trade–off between adjusting incentives and maintaining a strong set of incentives, however. 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 adjustment. 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 actors must believe that the incentives will remain in place over the long term. Balanced against this, however, is the difficulty in designing incentive structures given our current knowledge.

At a more fundamental level, methods of accumulating knowledge about how to deal with problems must be introduced. As indicated, many schools may not know what they can do to improve. While operating an output accountability system, states continue to regulate inputs through mandates about processes, teacher certification rules, and on–going teacher testing. One interpretation of their continued attention to inputs is that states do not trust the operation of schools. Alternatively, these requirements could represent the development of new hypotheses about how certain inputs relate to student outcomes. An advantage of accountability systems that require regular and detailed information about outcomes is that these underlying regulatory premises can be tested.

[18] Recent attention in the public utility field highlight the inadequacy of sole use of rewards; until penalties were included in the accountability schemes, service quality and other measured outcomes did not change. See www.fcc.gov\sq\reports2000.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

SOME OF TODAY'S BIGGER UNANSWERED CONCERNS

Other issues that go beyond the current accountability systems are possibly more important for the future.

Reactions to Bad Performance

A key issue in considering accountability systems is what to do if a student or a school is not meeting expected goals. Consider a school whose students have been judged as not meeting expected performance standards. Should resources be added or removed from this school?

The key to this fundamental question rests on information that frequently is not generated by accountability systems. Is the poor performance the result of deficient student backgrounds that simply require even larger effort to remediate? This cause argues for more resources. Or is the poor performance the result of bad teaching, bad school programs, and inefficient use of existing resources? This cause argues for different solutions, which may involve fewer resources. In particular, in the case of poor performance, one does not want to establish perverse incentives that reward doing badly.

Most accountability and incentive schemes prejudge or ignore this fundamental issue. The correct answer requires sufficient evidence to distinguish the causes of poor performance. While this is largely an implication of prior discussions about aligning results with actors, it has obvious importance to overall design issues. The current systems have not been demonstrated to be effective at this.

An easy way to put this into the previous context is to relate it back to the various efforts to separate school effects from other factors. That discussion indicated that further work was required to refine the methods of deriving accurate accountability measures, but this is also the work that is needed to identify policy directions.

Incentives and Efficiency

The motivation for attention to accountability should not be lost. Over at least the past decades, student achievement has been essentially flat. In that respect the system has not been doing worse. At the same time, there have been dramatic increases in the resources going into schools. Combined with the achievement results, a natural conclusion is that the largest problem with the current system is the inefficient use of resources—added resources are not being systematically turned into improved student outcomes.

Accountability systems developed by states seldom, however, address issues of resource usage and performance. Instead they concentrate almost exclusively on pure achievement results. They begin with outcomes and move forward to consequences without regard for the factors that create the results.

Will incentives to improve student achievement outputs naturally lead to better use of resources? The answer is not obvious. While 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 rewards 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.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

Knowing Performance is Poor is Different from Knowing What to Do

Another more subtle aspect of feasibility revolves around whether people know what to do. Accountability systems identify when things are not working well. They do not identify the corrective actions that are required. At the same time, the very assumption of introducing output incentives is that the system cannot be effectively regulated from higher levels—i.e., that the way to do better cannot be easily described and cloned. More than that, the evidence on the importance of teacher quality differences (e.g., Rivkin, Hanushek, and Kain, 2001) suggests that improvement may come from changes in personnel, as opposed to programmatic fixes that are frequently the focus of public discussion. This situation of course introduces a difficulty. While some schools may recognize the problems and know how to change, others may have no idea about how to deal with their shortcomings.

The dilemma is clear. Past research into the determinants of student performance —whether looking at teacher characteristics, specialized programs, or management and leadership—has not produced clear indications of how systematically to improve student performance. However, if current personnel don't know how to produce student performance now, can they be expected to select improvements in the future after the incentives have changed? Continuing research and evaluation may be part of the answer, but on–going research into the specific determinants of performance has yet to be very successful and is unlikely to provide any immediate guidance to school personnel. 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, provided better incentives and motivation, will be able to move to better student outcomes. Clearly, this presumption needs to be judged over time. The circumstances where this is and is not the case should be an important element of the evaluation of accountability systems.

External Validity

One of the largest issues about accountability systems is also one of the most basic. Testing and measurement people in schools have spent considerable time on issues of alignment of the testing with the curriculum. The underlying notion is that the tests and incentives must directly relate to the learning objectives embedded in the system. At the same time, a well–tuned system that was not geared to the external demands for educated people in society would not be very productive.

All testing is focused internally, and 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 motivation of improving student performance has led policy makers to focus directly on student achievement. While prior policy had focused almost exclusively on the inputs to the educational system, the most recent reform has reverted in some ways to more fundamental issues—what we want students to know and how we will recognize whether they know it.

The accountability systems being put in place represent an exciting and very positive development. The focus on ensuring

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

that the goals for student knowledge are actually accomplished cannot be underestimated.

The simple structure of current accountability systems, however, masks a complexity that underlies the outcomes that can be expected. The central component of current accountability is testing of student performance. While there is significant controversy about what materials to test and how to test them, the basic building block of accountability remains measured student achievement.

Improvement of student performance, however, occurs only through the actions of students, parents, teachers, and schools. The basic idea behind the current reform movement is that by measuring and reporting student achievement—particularly in reference to agreed upon outcome standards—the underlying actors will respond to the inherent incentives to improve their performance. The mechanisms are generally undefined, but rewards and sanctions for some or all of the actors are introduced to ensure improvement of student performance.

The structuring of current accountability systems to one degree or another tends to deal with these issues of refined measurement and attribution of results. However, the newness to education of accountability for outcomes means that the reality of current reporting and accountability systems will need refinement. In many cases we do not have adequate experience, theory, or empirical evidence yet to judge the actual implementation, but the impact of these systems will depend upon learning as they are operating.

While the movement toward systems based on student performance offers the best chance for improvement (Hanushek et. al., 1994), the journey is much closer to the beginning than to a successful conclusion. It would be very unfortunate if extravagant expectations for miracles, when not immediately met, led to abandoning the focus on outcomes and accountability. The current systems will require adjustment and modification, but that is a much more likely path to educational improvement than the more traditional jumping from one "solution" to another.

REFERENCES

Belden, Russonello, and Stewart.
Making the Grade: Teachers' Attitudes toward Academic Standards and State Testing. Washington, D.C.: Belden, Russonello, and Stewart, 2000.

Bishop, John.
"Signaling, Incentives, and School Organization in France, the Netherlands, Britain, and United States." In *Improving America's Schools: The Role of Incentives*, edited by Eric A. Hanushek and Dale W. Jorgenson, 111–45. Washington, D.C.: National Academy Press, 1996.

Clotfelter, Charles T., and Helen F. Ladd.
"Recognizing and Rewarding Success in Public Schools." In *Holding Schools Accountable: Performance–Based Reform in Education*, edited By Helen F. Ladd, 23–63. Washington, D.C.: Brookings, 1996.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James Mcpartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York.
Equality of Educational Opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Downes, Thomas A.
"Evaluating the Impact of School Finance Reform on the Provision of Public Education: The California *Case." National Tax Journal* 45 No. 4 (December, 1992): 405–19.

Education Week.
Quality Counts 2001: A Better Balance. Washington, D.C.: Education Week, January 11, 2001.

Elmore, Richard F., 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*, edited by Helen F. Ladd, 65–98. Washington, D.C.: Brookings, 1996.

National Tax Journal
Vol. 54 no. 2 (June 2001) pp. 365-384

Goff, John M.
A More Comprehensive Accountability Model. Washington, D.C.: Council for Basic Education, 2000.

Gormley, William T., Jr., and David L. Weimer. *Organizational Report Cards*. Cambridge, MA: Harvard University Press, 1999.

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

Hanushek, Eric A.
"The Trade-off between Child Quantity and Quality." *Journal of Political Economy* 100 No. 1 (February, 1992): 84–117.

Hanushek, Eric A.
"Assessing the Effects of School Resources on Student Performance: An Update." *Educational Evaluation and Policy Analysis* 19 No. 2 (Summer, 1997): 141–64.

Hanushek, Eric A., and Dennis D. Kimko.
"Schooling, Labor Force Quality, and the Growth of Nations." *American Economic Review* 90 No. 5 (December, 2000): 1184–208.

Hanushek, Eric A., and Steven G. Rivkin.
"Understanding the Twentieth–Century Growth in U.S. School Spending." *Journal of Human Resources* 32 No. 1 (Winter, 1997): 35–68.

Hanushek, Eric A., John F. Kain, and Steven G. Rivkin.
"The Costs and Benefits of Switching Schools." Hoover Institution, Stanford Unversity. Mimeo, 2001.

Hanushek, Eric A., and Julie A. Somers.
"Schooling, Inequality, and the Impact of Government." In *The Causes and Consequences of Increasing Inequality*, edited by Finis Welch, 169–99. Chicago: University of Chicago Press, 2001.

Hanushek, Eric A., with others.
Making Schools Work: Improving Performance and Controlling Costs. Washington, D.C.: Brookings Institution, 1994.

Kane, Thomas J., and Douglas O. Staiger.
Improving School Accountability Measures. NBER Working Paper No. 8156. Cambridge, MA: National Bureau of Economic Research, 2001.

Koretz, Daniel, 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.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain.
"Teachers, Schools, and Academic Achievement." NBER Working Paper No. 6691. Cambridge, MA: National Bureau of Economic Research, 2001.

Rothstein, Richard, and Karen Hawley Miles.
Where's the Money Gone? Changes in the Level and Composition of Education Spending. Washington, D.C.: Economic Policy Institute, 1995.

Sanders, William L., and Sandra P. Horn.
"The Tennessee Value–Added Assessment System (TVAAS): Mixed–Model Methodology in Educational Assessment." *Journal of Personnel Evaluation in Education* 8 No. 3 (October, 1994): 299–311.