CPRE Policy Briefs

Reporting on Issues and Research in Education Policy and Finance

September 2003
RB-38

Graduate School of Education
University of Pennsylvania

Redesigning Accountability Systems for Education

by Susan H. Fuhrman

Accountability is a topic on everyone's mind. In just about every state, schools are being held accountable for student performance under systems put into effect over the past 5-10 years. As states are providing remedies and enacting sanctions for low performance, policymakers are realizing the daunting implications of the task in front of them. In over half the states, students will have to pass a state test to graduate from high school; concerns about large numbers of failures, particularly for minority students, are mounting. The recent reauthorization of the Elementary and Secondary Education Act, the No Child Left Behind Act (NCLB) of 2001, sets new requirements for state accountability systems as a condition of federal aid for disadvantaged children. As a result, states are actively reexamining their accountability policies.

To assist in the redesign of accountability systems, the Consortium for Policy Research in Education (CPRE) and the Center for Research on Evaluation, Student Standards, and Testing (CRESST) sought to assemble knowledge from new research on emerging accountability systems. A book, *Redesigning Accountability Systems for Education*, edited by Susan H. Fuhrman and Richard F. Elmore (Teachers College Press, in press), contains chapters by leading accountability researchers. This issue of CPRE Policy Briefs summarizes the book by focusing on four questions the authors of the book address:

1) How valid are new accountability systems?
2) How fair are new accountability systems?
3) What are the effects of new accountability systems?
4) What is necessary to improve the functioning of accountability systems?

This Policy Brief reviews the many issues that states are confronting as they implement accountability systems, and provides guidance for states looking to fine-tune or redesign accountability systems to help meet policies as they were intended. Specifically, this Brief offers recommendations for improving accountability systems by enhancing the use of expert technical advice, by improving the collection and interpretation of system data, and by investing in capacity building to ensure that both students and educators have the necessary means to effectively respond to accountability systems.

Background

The accountability systems written about in the book are those established over the past 5-10 years, mostly at the state level, although a number of districts have similar systems. NCLB accountability provisions also reflect the same principles.

These systems are distinguished by their attention to school-level performance and by their inclusion of consequences for that performance. They are quite different from earlier approaches to accountability that primarily focused on district compliance with state regulations. The new systems grow out of a climate that draws strong parallels between education and business; they intend to focus schools on the bottom line. They also reflect an attempt, strong in rhetoric if not reality, by states to back off from detailed regulations

Book Release

Redesigning Accountability Systems for Education

Edited by Susan H. Fuhrman and Richard F. Elmore (In Press)

Available from Teachers College Press; Telephone: (800) 575-6566; Online orders: www.teacherscollegepress.com

Consortium for Policy Research in Education

University of Pennsylvania
Harvard University
Stanford University
University of Michigan
University of Wisconsin-Madison

about the process of education. The new systems reflect an explicit theory of action about improving student achievement that stresses the motivation of teachers, students, and administrators.

The new systems assume that, when they are operating as intended:

- *Performance, or student achievement, is the key value or goal of schooling and that constructing accountability around performance focuses attention on it*. Since the indices that are used to measure school status and progress are composed primarily of achievement measures, the systems are intended to maximize focus on those measures.
- *Performance can be accurately and authentically measured by the assessment instruments in use*. Assessments are aligned to student standards and gauge achievement of those standards in reliable and valid ways. If accountability is to hinge on performance, then the key measures used by the accountability system must correctly assess performance. Further, the new systems generally assume that school performance can be fairly assessed through testing; only in a few states do accountability systems include provisions for visiting and reviewing schools by observing teaching and learning.
- *Consequences, or stakes, motivate school personnel and students*. Not only do those subject to stakes focus more on performance, they try harder because both positive inducements (such as bonuses) and negative sanctions (such as school takeover, reconstitution, or denial of promotion or graduation) are meaningful and real.
- *Improved instruction and higher levels of performance will result*. Teachers trying harder to teach and students trying harder to learn will connect to mean better interaction around content. The assessments will also help promote good instruction by providing feedback on student performance. Following this assumption, motivation is a key to improving instruction. If teachers don't have the capacity necessary to respond to the accountability system incentives, it is assumed that the incentives are strong enough to motivate them and administrators to find it somehow, by seeking additional professional development, for example. Also, attaching consequences at the school level assumes that schools will collectively be able to fashion a response and, therefore, that they have, or will be motivated to form, some sort of internal coherence. Many accountability systems are accompanied by policies to build capacity and most have some assistance strategies embedded in the consequences for failing schools, but accountability policies focus primarily on altering the incentive structure as a means to improving instruction and performance. This primary reliance on incentives to motivate teachers and schools to do something the schools have never done before — to succeed with essentially all students — suggests that these systems make an important additional assumption or set of assumptions (i.e., that teachers already know how to succeed with all students but choose not to, or don't expect to, with some, or that at least somebody knows how to succeed so that, if motivated, others can learn how to do it too).
- *Unfortunate unintended consequences are minimal*. If the systems work as intended, the goal of higher performance will not be undermined by perverse incentives or other negative developments. For example, instruction will improve, not become narrowly focused around test-taking skills, higher hurdles for high school graduation will not increase dropping out, and holding schools accountable will not cause exclusion of special-needs students from testing or retention of students in non-tested grades.

Are the new systems working as intended? Are the assumptions borne out? We can address these questions by asking how valid and fair the systems are; by asking about their effects on motivation, instruction, and

The Consortium for Policy Research in Education (CPRE) is funded by the Institute of Education Sciences, United States Department of Education. The research reported in this Brief was funded under Grant No. R308A960003. Opinions expressed in this Brief are those of the author and do not necessarily reflect the views of the Institute of Education Sciences; the United States Department of Education; the Center for Research on Evaluation, Student Standards, and Testing; CPRE; or its institutional members.

performance; and by asking what would be needed to improve system function.

How Valid are New Accountability Systems?

This question asks whether the new accountability systems are accurately focusing on student learning as their rhetoric suggests. In other words, are assessments sensitive to instruction — are they correctly measuring student learning — and are they put to appropriate uses, given the information they provide?

With respect to assessment, a first question has to do with the adequacy and appropriateness of test content and of the cognitive demands of the test (Baker & Linn, in press). Commonly, this issue is spoken of as "alignment." Are tests measuring achievement of the knowledge and skills states expect of students? Do tests adequately cover the content in state standards in terms of both topic coverage and level of rigor? The evidence is not encouraging on this score. Achieve, an independent, bipartisan, nonprofit organization, founded by a group of governors and chief executive officers in 1996, has worked with over 20 states examining alignment. It found that state tests do cover standards (nearly all test items measure content in standards), but coverage is often superficial with tests measuring the least complex of the skills called for. Tests tend to be unbalanced, measuring some standards but not others. For example, some state high school math tests tend to focus mostly on numbers and measurement, with little emphasis on algebra and geometry. High school tests are particularly problematic, posing a relatively low level of challenge, which is in sharp contrast to fourth- and eighth-grade tests in some of the same states (Rothman, in press). Achieve only reviewed one commercially available test; at grade 10, one-quarter of the test questions did not match the state standards at all. This is in contrast to the findings about significant matches between standards and tests in states that design their own tests and raises concerns about the likely increased use of less expensive commercial tests in response to NCLB's increased testing requirements. States that use commercial tests will need to ensure that test publishers augment the examinations in ways that align with the state's standards.

Accountability systems establish levels of performance, or cut scores, in order to define a certain level of achievement as "proficient" or "basic." States use different approaches and methods to set cut scores and, sometimes, the methods are not well elaborated. The fluidity of these definitions is illustrated by the action of some states in the wake of NCLB's requirement to bring all students to "proficient" within 12 years. As of this writing, at least three states have changed their definition of "proficient," using for federal purposes a level previously called "basic" or "partially proficient." Further, measurement error is associated with any test score, and both individual students and schools can be misclassified as either "proficient" or "not proficient" simply because of random error. Research has shown that the probability of misclassification is substantial. When accountability systems require disaggregated reporting of scores, measurement and sampling error is greater for schools with large numbers of subgroups. In addition, the danger of misclassification is greater for small schools. Error rates increase as the number of students decrease. It is not uncommon for the best and worst performers to include large numbers of small schools, probably placed into these categories by error. It is not clear that policymakers know the probability of misclassification or that such information is provided to all users (Baker & Linn, in press).

Another aspect of validity has to do with whether scores represent learning or other factors. A "status" score or achievement level reflects the student or students' background as much as it does any learning that took place in the year of testing. For that reason, many states include gain scores or improvement ratings in their accountability measures. Some use both the percentage of students at a given achievement level and a gain score to rate schools. States use different models for judging improvement. For example, they can look at changes in the performance of successive years of students for the same grade or they can look at changes in performance from one grade to the next for students who were tested in both years. The problem with the first model is that different cohorts of students can have very different characteristics. In addition, in areas with a lot of mobility, the turnover of students in a

school could be responsible for performance changes, not learning. The second model is very appealing because it holds schools accountable for the value they add; it controls for student background by controlling for student achievement. Even though this approach may not completely eliminate the effect of background factors, such as student access to help at home during the school year, it comes closer than the other approach (Linn, in press).

Validity also concerns whether measures are put to appropriate use. Many testing experts would fault today's accountability systems on several grounds. Consequences are often applied on the basis of a single measure of mastery, rather than using multiple measures that tap into different ways of demonstrating competence in a content domain. Policymakers often say they are using multiple measures when they provide multiple opportunities to take the same test, but that is not the same as having multiple assessments. Also, the chances of misclassification raise doubts about the application of harsh consequences based on a single test administration. In addition to giving students multiple opportunities to take tests that count for graduation or promotion, some states are averaging scores for schools over a period of years.

Furthermore, the use of a test score to impose a reward or a sanction presumes that the individuals whose actions produce those scores — teachers and students — have the wherewithal and know-how to do their job. For a score to be valid, teachers have to understand and act on the import of the accountability information. Because school accountability systems focus almost exclusively on outcomes, they produce little in the way of reliable information about classroom practice. Nor do typical schools have other mechanisms for collecting and sharing information about instruction across classrooms. Hence, school personnel rarely have enough data to figure out what factors produce given outcomes and design a remedy; they lack sufficient data for attribution (O'Day, in press). And importantly, for accountability systems to be valid, teachers must have the capacity to teach students the knowledge and skills to be assessed. A major issue cited by authors in the book is lack of instructional capacity — of materials and teacher knowledge and skill — and therefore of opportunity to learn. If accountability theory suggests that by providing strong incentives, teachers will muster abilities they lacked before the incentives were imposed, then the theory must be faulted. Capacity does not magically appear, as will be seen when effects are examined.

How Fair are New Accountability Systems?

Fairness has a lot to do with validity and is hard to single out. It is not a valid use of a test to allocate consequences for teachers based on the scores of their students if variations in those scores are heavily influenced by factors other than quality of instruction or instructional effort. So, if low student socioeconomic status depresses scores in a school, the average score is neither a fair nor valid measure of that school's instructional efforts. And with respect to students, applying consequences according to their test scores, if they have lacked the opportunity to learn the material, raises both validity and fairness questions. But in this section we take up some aspects of fairness not discussed above: the inclusion of students with disabilities and low levels of English proficiency, the disparities in achievement among subgroups, and the uneven application of consequences.

States are required by federal law to include students with disabilities and limited English proficient (LEP) students in their assessments and accountability systems, and NCLB is significantly more directive about inclusion than past policies. However, in recent years, some 36 states were cited by the federal government for problems with including students with disabilities, and 33 were cited for problems with including LEP students (Thurlow, in press). Sometimes states provide adequate accommodations for students (such as longer time to take a test), but do not include scores for all such accommodated students in their accountability systems. Exclusion of scores is even more prevalent for students who use alternate assessments or "nonstandard" accommodations, accommodations the student's Individualized Education Program team deems important even though they are not on the state list. As Thurlow (in press) puts it, "Unless states

figure out a way to include students with non-approved accommodations in accountability systems, they may create incentives to designate students with non-approved accommodations solely as a way of excluding them."

Even though federal law regulates inclusion in accountability systems related to Title I, it is silent with regard to inclusion of disabled or LEP students in state graduation and promotion testing. Policymakers may give such students alternative diplomas, but aside from the well-established GED (General Educational Development Diploma), the value of such alternatives in terms of a student's future educational or employment opportunities is uncertain (Heubert, in press). Disabled, LEP, and minority students fail state graduation tests at much higher rates than other students. Even in Texas, which according to some studies has made significant progress in reducing the achievement gap on its high school test between Whites and other students, the failure rates for Hispanics and African American students was more than double that of Whites as of 1998 (Carnoy, Loeb, & Smith, 2001; Natriello & Pallas, 2001 as cited in Heubert, in press). In states with more rigorous high school assessments, both the disparities and the failure rates, with some as much as 60-90%, are higher. Mounting evidence about the greater prevalence of underprepared and misassigned teachers in high-poverty schools (Ingersoll & Jerald, 2002) suggests that opportunity to learn is not evenly distributed. At the same time, it is becoming harder for students to bring legal action regarding these tests. In 2001, the Supreme Court decided that private individuals could no longer bring "disparate impact" cases under Title VI of the 1964 Civil Rights Law, which had been a powerful tool in the past (Heubert, in press).

While some students seem more at risk from new accountability systems than others, it is worth noting that students generally face more consequences than adults under new state accountability systems. Stakes are seriously imbalanced, applying more harshly to students than to schools and the adults that work in them. The adults are somewhat sheltered by the fact that a school is a collective of individuals; consequences are diffused throughout the organization rather than falling on specific individuals, but students bear the brunt of consequences as individuals (Elmore, in press; O'Day, in press; Siskin, in press). Stakes fall unambiguously on students, who, unlike the adults who are supposed to be providing them with the opportunity to learn, do not have the means to defend themselves politically. If they are represented at all in debates about accountability, it is by adults who have their own interests to protect (Elmore, in press). In addition, states seem to be moving ahead in the application of stakes on low-performing students — withholding promotion or graduation — while dramatically withdrawing from applying the consequences their policies require for low school performance. Because they lack the capacity to conduct in-depth reviews and to provide assistance, states are typically targeting for action many fewer schools than are eligible for remedies on performance criteria (Fuhrman, Goertz, & Duffy, in press).

Given the mixed record of including all children in accountability policies, the disparate impact of these systems on different groups of students, and the uneven application of stakes, it would be hard to argue that new accountability systems are currently fair, although improvement on these factors may come over time. This is certainly what policymakers are promising. The more overarching issue of fairness with respect to the new policies is one we have mentioned before: Do students have the opportunity to learn the material on which they are being assessed. As Elmore (in press) points out, "In a society where educational attainment is heavily related to future income, retention in grade, denial of diplomas, and dropping out have consequences that are extremely serious for students." It is unethical to punish students for not learning content they have not been taught. What do we know about opportunity to learn? For that, we turn to evidence about the effects of new accountability policies.

What are the Effects of New Accountability Policies?

A central point is that the effects of accountability policies vary. New accountability systems certainly get the attention of teachers and other school personnel. Teach-

ers and principals report significant effects of assessments on curriculum and instruction and studies have shown that they allocate their time according to the centrality of subjects in the testing system. In other words, assessment policies are motivating and lead to modifications in practice (Herman, in press). But *how* teachers actually respond to the signal, *how* they modify curriculum and instruction, differs quite a bit from school to school and even from student to teacher.

A number of studies find that curriculum and instruction become narrowed as a result of increased focus on state assessments. Non-tested subjects are given short shrift and teachers use the test format as a model for instruction, so when multiple-choice items dominate in the assessment, they are included in teacher worksheets as well. In addition, teachers report spending significant amounts of time in test preparation, more so in schools serving high-poverty children (Firestone, Camilli, Yurecko, Monfils, & Mayrowetz, 2000; and Herman & Golan, 1993 as cited in Herman, in press). On the other hand, attention to the assessment can mean adding to the curriculum, depending on the nature of the assessment. Researchers found teachers including more problem-solving tasks and writing in states like Maryland and Kentucky (Firestone, Mayrowetz, & Fairman, 1998; Stecher & Barron, 1999; and Stecher, Barron, Kaganoff, & Goodwin, 1998 as cited in Herman, in press). And some teachers in all states studied rose to the challenge and tried to enhance instruction to meet new standards.

What accounts for this variation? Some of the difference has to do with the nature of the assessments; more sophisticated assessments with open-ended items are modeled in practice just as less sophisticated tests are associated with worksheets and drill-type activities (Herman, in press). But, in large measure, the variation reflects differences in capacity among schools; in the knowledge, skill levels, and belief systems of teachers; in the ability of the school to fashion a collective response to external accountability; and in the effectiveness of leadership. Accountability systems do not by themselves appear to mobilize new capacity; schools' responses to them depend heavily on the capacity they already have. As Elmore (in press) puts it, "The best predictor of how a school will respond to the introduction of stakes at Time 1 is its organizational culture and capacity at Time 0…" As pressure increases, low-capacity schools may add academic content and remediation, but without deliberate capacity building, they are unlikely to make large improvements in their core instructional capacity. Schools with more attention to and capacity for academic success often respond to accountability pressure in ways that increase their academic focus and coherence (Elmore, in press).

Even in high schools, where generally standards reforms seem to have had the least effect to date and achievement gains have yet to be seen, varied responses to accountability systems are seen. High schools are being asked to do what they have never done before — bring all children to common high standards, instead of differentiating academic content. They have difficulty focusing in on a few academic subjects — the ones most likely to be tested — since this threatens the importance of faculty in many other departments. As we have seen, students, many of whom come to high school far behind in their academic progress, are increasingly the targets of accountability pressures as individuals. Yet, some high schools respond more constructively than others. Schools that are more academically focused to begin with face the challenge of providing their academic programs to all, not just most students. But that is less daunting than the situation of schools without serious academic focus; they must now invent it. To some teachers in such schools, who often find less than half of their students graduating, the challenge of preparing students most at risk seems impossible (Siskin, in press). When such very low-capacity schools respond to accountability systems by focusing more on performance, they may be on a long-term improvement trajectory, but they may also be complying in a pro forma way, without much deep capacity building (DeBray, Parson, & Avila, 2003).

Despite the fact that capacity-related variation is the predominant finding in studies of classroom effects of accountability policies, there are some overall trends in achievement data. Looking at eighth-grade mathematics scores on the National Assessment of Educa-

tional Progress (NAEP) from 1996-2000, Carnoy and Loeb (in press) found significantly larger gains in states with strong accountability systems (those with significant consequences for schools and students), across all racial and ethnic groups and particularly for African Americans. While fourth-grade results were not as strong, the relationship between forceful accountability and African American performance was noted there as well. Although there has been some national debate about accountability tests leading to increased retention in ninth grade, and students dropping out, Carnoy and Loeb did not find such a relationship.[1] However, they also did not note any positive effect of accountability systems on student attainment. The new systems are not holding students in high school at greater rates, nor are they leading to greater rates of college-going. Since postsecondary behavior, particularly college attendance, has substantial long-term effects on income and life opportunities, we can only hope that score increases in lower grades eventually presage not only better student performance but also better student attainment at higher grades. It is important to note that there is considerable variation from state to state in NAEP scores, even among states with strong accountability systems. Strong accountability systems send a signal, but as shown, they do not in themselves provide the capacity necessary for students and schools to respond constructively.

What is Necessary to Improve Accountability Systems?

Accountability systems need not be set in stone. They can be refined and improved over time. As of this writing, states have planned changes to their systems in response to NCLB that they will now have to design in detail and implement. This process could provide opportunities for improvement.

Many observers have recommended significant changes in existing accountability systems, such as increasing the use of multiple measures or assuring that adults bear consequences before students. CPRE and CRESST have developed standards for accountability systems (see sidebar on page 8) to help policymakers develop more valid, fair, and effective systems (Baker, Linn, Herman, Koretz, & Elmore, 2001). *Redesigning Accountability Systems for Education* includes many other specific recommendations about improving accountability systems (Elmore, in press; Herman, in press; Heubert, in press; O'Day, in press). Several themes run through those recommendations.

First, *technical information about assessment and accountability systems must be brought to bear when policymakers deliberate accountability systems*. Policymakers need to know the error terms of assessments, for example, so they can determine the chances of misclassifying students or schools based on a test administration. They need to know how validly the assessment aligns with their standards, using measures of alignment in addition to coverage, how validity could be improved by including additional measures, and the trade-offs among various means of setting cut scores. If they set requirements for schools to make certain amounts of progress, they need to know if those requirements are feasible, given past performance and likely gains. They also need to know the advantages and challenges of using value-added accountability models as opposed to other models. Certainly, accountability systems are deeply political, with much consideration of possible winners and losers coming into decisions about how to structure them. But if policymakers want to advance their overall aim of improved performance, they need solid technical information from independent, credible sources — from experts — in addition to those with a vested interest in promoting a particular assessment.

Second, *additional information about the education system is necessary to interpret accountability system performance data*. At least three kinds of information greatly enhance the ability of users to make sense of and act on performance data produced by an accountability system. Enhanced information would also make possible a broader array of measures of the health of the education system, hopefully alleviating the intense focus on assessments.

[1] *In an earlier study of Texas, Carnoy, Loeb, and Smith (2001) found that an increase in ninth-grade retention in Texas pre-dated the TAAS (Texas Assessment of Academic Skills) system, though it may have stemmed from earlier increases in high school graduation requirements.*

CPRE/CRESST Standards for Accountability Systems

Standards on System Components

- Accountability systems should employ different types of data from multiple sources.
- The weighting of elements in the system, different test content, and different information sources should be made explicit.
- Accountability systems should include data elements that allow for interpretations of student, institution, and administrative performance.
- Accountability expectations should be made public and understandable for all participants in the system.
- Accountability systems should include the performance of all students, including subgroups that historically have been difficult to assess.

Testing Standards

- Decisions about individual students should not be made on the basis of a single test.
- Multiple test forms should be used when there are repeated administrations of an assessment.
- The validity of measures that have been administered as part of an accountability system should be documented for the various purposes of the system.
- If tests are to help improve system performance, data should be provided illustrating that the results are modifiable by quality instruction and student effort.
- If test data are used as a basis of rewards or sanctions, evidence of technical quality of the measures and error rates associated with misclassification of individuals or institutions should be published.
- Evidence of test validity for students with different language backgrounds should be made available publicly.
- Evidence of test validity for children with disabilities should be made available publicly.
- If tests are claimed to measure content and performance standards, evidence of the relationship to particular standards or sets of standards should be provided.

Stakes

- Stakes for accountability systems should apply to adults and students.
- Incentives and sanctions should be coordinated for adults and students to support system goals.
- Appeal procedures should be available to contest rewards and sanctions.
- Stakes for results and their phase-in schedule should be made explicit at the outset of the implementation of the system.
- Accountability systems should begin with broad, diffuse stakes and move to specific consequences for individuals and institutions as the system aligns.

Public Reporting Formats

- System results should be made broadly available to the media, with sufficient time for reasonable analysis and with clear explanations of legitimate and potential illegitimate interpretations of results.
- Reports to districts and schools should promote appropriate interpretation and use of results by including multiple indicators of performance, error estimates, and performance by subgroup.

Evaluation

- Longitudinal studies should be planned, implemented, and reported, evaluating effects of the accountability program. Minimally, questions should determine the degree to which the system: builds capacity of staff; affects resource allocation; supports high-quality instruction; promotes student equity access to education; minimizes corruption; affects teacher quality, recruitment, and retention; and produces unanticipated outcomes.
- The validity of test-based inferences should be subject to ongoing evaluation. In particular, evaluation should address: aggregate gains in performance over time and impact on identifiable student and personnel groups.

- More data about classroom-level curriculum and instruction would help school users figure out why test scores are at certain levels and decide what to do about it. Without information about practice, schools are limited in designing remedies for poor performance. As O'Day (in press) points out, professional accountability systems that include opportunities for peer exchange about practice provide greater knowledge for action than bureaucratic systems that include only information about results.
- More knowledge about the state of opportunity to learn would help policymakers design fairer systems. Knowing the extent to which students are truly being taught the material to be assessed would help policymakers determine realistic progress goals and assess the fairness of consequences. Knowing the variation in opportunity to learn would help policymakers channel additional resources and assistance to needy schools. Several states worried about high failure rates on high school exit exams undertook studies of opportunity to learn that were instrumental in setting timelines for the initiation of consequences (Fuhrman, Goertz, & Duffy, in press).
- Evaluations of accountability systems are essential. Good evaluations would indicate whether students are being appropriately included in assessments, whether assessments have disparate impacts on various groups, whether classroom practice is changing in response to assessment (in ways both intended and not intended), whether it provides remedies for poor performance work, and a host of other equally critical questions. Evaluations will show whether teachers have the ability to do the expected job and whether that capacity is fairly distributed.

Third, *capacity building is essential*. Deliberate interventions to improve teacher knowledge and skill, provide extra assistance to students at risk of failure, and to build school communities capable of responding to performance pressure are necessary. Further, states and districts need added capacity if they are to assist schools and intervene in instruction. Without investments of this type in capacity, improvements related to accountability systems are likely to be short lived and superficial, and inequities are likely to increase. Policymakers have worked hard on the motivation side of the equation in developing accountability policies; they must work equally hard on providing educators and students the wherewithal to respond to the new incentives.

Finally, to enhance the use of expert technical advice, to improve information gathering, and to invest in capacity, policymakers need *political stamina*. Accountability systems have become such important cornerstones of state policy that some policymakers are afraid to modify them, worried that opponents will seize the opportunity of revision to undermine the whole system. Their concern has mounted as backlash to accountability policies has gained force. However, the opposition includes not only those philosophically against state-directed testing and consequences for performance, but many who would be supporters in principle but are concerned about such issues as unequal opportunity to learn, disparate impacts, reliance on single measures, and harsh consequences for students. Some of the latter group would be willing to come to the table and discuss ways to improve accountability systems, making it politically possible to modify these systems without risking their complete undoing. This was the case in several states that modified their high school exit exam policies over the past several years. Continued leadership and business support, willingness to commission and attend to research about the state of opportunity to learn, and readiness to compromise on specific issues like test content and effective dates in order to maintain the basic program permitted refinements to occur (Fuhrman, Goertz, & Duffy, in press). Policymakers must take advantage of lessons from experience with new accountability systems and use that knowledge to change and improve the systems over time.

About the Author

Susan Fuhrman is the Dean, and the George and Diane Weiss Professor of Education at the University of Pennsylvania's Graduate School of Education in addition to being Chair of CPRE's Management Committee. Fuhrman has written widely on education

policy and finance. Her research interests include state policy design, accountability, deregulation, intergovernmental relationships, and standards-based reform.

References

Baker, E. L., & Linn, R. L. (in press). Validity issues for accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Baker, E. L., Linn, R. L., Herman, J. L., Koretz, D., & Elmore, R. F. (2001, April). *Holding accountability systems accountable: Research-based standards*. Symposium presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Carnoy, M., & Loeb, S. (in press). Does external accountability affect student outcomes? A cross-state analysis. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Carnoy, M., Loeb, S., & Smith, T. (2001). *Do higher state test scores in Texas make for better high school outcomes*? (CPRE Research Report No. RR-047). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

DeBray, E., Parson, G., & Avila, S. (2003). Internal alignment and external pressure: High school responses in four state contexts. *The new accountability: High schools and high-stakes testing*. New York: Routledge.

Elmore, R. F. (in press). Conclusion: The problem of stakes in performance-based accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Fuhrman, S. H., Goertz, M. E., & Duffy, M. C. (in press). "Slow down, you move too fast": The politics of making changes in high-stakes accountability policies for students. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Herman, J. L. (in press). The effects of testing on instruction. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Heubert, J. P. (in press). High-stakes testing in a changing environment: Disparate impact, opportunity to learn, and current legal protections. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Ingersoll, R. M., & Jerald, C. (2002). *All talk, no action: Putting an end to out-of-field teaching*. Washington, DC: The Education Trust.

Linn, R. L. (in press). Accountability models. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

O'Day, J. A. (in press). Complexity, accountability, and school improvement. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Rothman, R. (in press). Benchmarking and alignment of state standards and assessments. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Siskin, L. S. (in press). The challenge of the high schools. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Thurlow, M. L. (in press). Biting the bullet: Including special-needs students in accountability systems. In S. H. Fuhrman & R. F. Elmore (Eds.), *Redesigning accountability systems for education*. New York: Teachers College Press.

Recent CPRE Publications

The following is a list of selected publications reporting on research conducted by the Consortium for Policy Research in Education. Research reports are available for a nominal fee; technical reports are available free-of-charge. For further information, please see below, call (215) 573-0700, ext. 1, or visit us on the world wide web at www.cpre.org

Technical Reports

Systemic Reform in Practice: Merck Institute for Science Education March 2003

The Impact of Standards-based Reform in Duval County, Florida, 1999-2002

Jonathan Supovitz and Brooke Snyder Taylor, May 2003 (No charge)

The Heart of the Matter: The Coaching Model in America's Choice Schools

Susan Poglinco, Amy Bach, Kate Hovde, Sheila Rosenblum, Marisa Saunders, and Jonathan Supovitz, May 2003 (No charge)

Research Reports

The Merck Institute for Science Education: A Successful Intermediary for Educational Reform

Tom Corcoran, March 2003, RR-052 ($5.00)

Teacher Leadership as a Strategy for Instructional Improvement: The Case of the Merck Institute for Science Education

Kate Riordan, March 2003, RR-053 ($5.00)

Prices include book-rate postage and handling. Make checks payable to Trustees of the University of Pennsylvania. Sorry, we cannot accept returns, credit card orders, or purchase orders. Sales tax is not applicable. To obtain copies, write:

CPRE Publications
Graduate School of Education
University of Pennsylvania
3440 Market Street, Suite 560
Philadelphia, PA 19104-3325

Books by CPRE Researchers

Books must be ordered directly from the publisher indicated.

The New Accountability: High Schools and High-stakes Testing
Martin Carnoy, Richard F. Elmore, and Leslie S. Siskin (Eds.), 2003, $19.95
Available from Routledge: www.routledge.com

Who Controls Teachers' Work? Power and Accountability in America's Schools
Richard M. Ingersoll, 2003, $39.95
Available from Harvard University Press: www.hup.harvard.edu/catalog/INGWHO.html

All Else Equal: Are Public and Private Schools Different?
Luis Benveniste, Martin Carnoy, and Richard Rothstein, 2003, $19.95
Available from Routledge: www.routledge.com

California Dreaming: Reforming Mathematics Education
Suzanne Wilson, 2003, $29.95
Available from Yale University Press: www.yale.edu/yup/books/094329.htm

Nondiscrimination Statement

The University of Pennsylvania values diversity and seeks talented students, faculty, and staff from diverse backgrounds. The University of Pennsylvania does not discriminate on the basis of race, sex, sexual orientation, religion, color, national or ethnic origin, age, disability, or status as a Vietnam era veteran or disabled veteran in the administration of educational policies, programs, or activities; admissions policies, scholarships, or loan awards; and athletic or University-administered programs or employment. Questions or complaints regarding this policy should be directed to Executive Director, Office of Affirmative Action, 1133 Blockley Hall, Philadelphia, PA 19104-6021 or 215-898-6993 (Voice) or 215-898-7803 (TDD).

About CPRE

The Consortium for Policy Research in Education (CPRE) studies alternative approaches to education reform in order to determine how state and local policies can promote student learning. Currently, CPRE's work is focusing on accountability policies, efforts to build capacity at various levels within the education system, methods of allocating resources and compensating teachers, governance changes like charters and mayoral takeovers, finance, student and teacher standards, and student incentives. The results of this research are shared with policymakers, educators, and other interested individuals and organizations in order to promote improvements in policy design and implementation.

CPRE unites five of the nation's leading research institutions to improve elementary and secondary education through research on policy, finance, school reform, and school governance. Members of CPRE are the University of Pennsylvania, Harvard University, Stanford University, the University of Michigan, and the University of Wisconsin-Madison.

CPRE Policy Briefs are published by CPRE. To learn more about CPRE research or publications, please call 215-573-0700 or access CPRE publications at www.cpre.org; www.wcer.wisc.edu/cpre/; or www.sii.soe.umich.edu.

CPRE Policy Briefs

Graduate School of Education
University of Pennsylvania
3440 Market Street, Suite 560
Philadelphia, PA 19104-3325

NON PROFIT
U.S. POSTAGE
PAID
PERMIT NO. 2563
PHILADELPHIA, PA