



# UNWARRANTED INTRUSION

*Isn't it ironic  
that Republicans  
sponsored the single  
largest—and  
potentially the  
single most  
damaging—  
expansion of federal  
power over the  
nation's education  
system in history?*

by RICHARD F. ELMORE

Inside the Washington, D.C., beltway, the reauthorization of the Elementary and Secondary Education Act (ESEA) is seen as either a sea change in federal education policy or a half-measure designed to demonstrate the political leadership's willingness to “do something” on education. On one side are supporters of the legislation who point to its substantial tightening of school accountability; its granting of more flexibility to states and school districts in the use of federal funds; and its commitment to applying sanctions to and providing aid for failing schools. On the other side are those who argue that the bill doesn't go far enough. Supporters of school choice believe that opportunities for real reform were lost when conservatives and New Democrats failed to persuade their colleagues and the president's advisors to include vouchers as part of the reform package.

In other words, there is no genuine opposition in Wash-

ington to accountability rules that simply fail to understand the institutional realities of accountability in states, districts, and schools. And the law's provisions are considerably at odds with the technical realities of test-based accountability. Never, I think, in the history of federal education policy has the disconnect between policy and practice been so evident, and possibly never so dangerous. What's particularly strange and ironic is that conservative Republicans control the White House and the House of Representatives, and they sponsored the single largest—and the single most damaging—expansion of federal power over the nation's education system in history.

The federal government is mandating a single test-based accountability system for all states—a system currently operating in fewer than half the states. The federal government is requiring annual testing at every grade level, and requiring states to disaggregate their test scores by racial and socioeco-

conomic backgrounds—a system currently operating in only a handful of states and one that is fraught with technical difficulties. The federal government is mandating a single definition of adequate yearly progress, the amount by which schools must increase their test scores in order to avoid some sort of sanction—an issue that in the past has been decided jointly by states and the federal government. And the federal government

in the mid-1980s by the National Governors Association, led by Bill Clinton, then the governor of Arkansas. It took the form of what was then called the “horse trade”—states would grant schools and districts more flexibility in making decisions about what and how to teach in return for more accountability for academic performance. This idea became the central theory of today’s accountability reforms. It was an appealing idea in

## The accountability movement meant that politicians could take credit for improving schools without committing themselves to serious increases in funding.

has set a single target date by which all students must exceed a state-defined proficiency level—an issue that in the past has been left almost entirely to states and localities.

Thus the federal government is now accelerating the worst trend of the current accountability movement: that performance-based accountability has come to mean testing, and testing alone. It doesn’t have to. In fact, in the early stages of the current accountability movement, reformers had an expansive view of performance that included, in addition to tests, portfolios of students’ work, teachers’ evaluations of their students, student-initiated projects, and formal exhibitions of students’ work. The comparative appeal of standardized tests is easy to see: they are relatively inexpensive to administer; can be mandated relatively easily; can be rapidly implemented; and deliver clear, visible results. However, relying only on standardized tests simply dodges the complicated questions of what tests actually measure and of how schools and students react when tests are the sole yardstick of performance.

If this shift in federal policy were based on the accumulated wisdom gained from experiences with accountability in states, districts, and schools, or if it were based on clear design principles that had some basis in practice, it might be worth the risk. In fact, however, this shift is based on little more than policy talk among people who know hardly anything about the institutional realities of accountability and even less about the problems of improving instruction in schools.

### The Implementation Problem

The idea of performance-based accountability was introduced

principle: governors and state legislators could take credit for improving schools without committing themselves to serious increases in funding.

The movement got a major boost in 1994, when Title I—the flagship federal compensatory education program—was amended to require states to create performance-based accountability systems for schools. The vision behind the 1994 amendments was that Title I would complement and accelerate the trend that began at the state level; they required states to develop academic standards, assessments based on the standards, and progress goals for schools and school districts, all within ambitious timetables. The merger of state and federal accountability policies—or alignment, as it was called—was supposed to occur by the year 2000. By the end of the decade, it was difficult to find more than one or two states lacking some form of statewide testing program and public release of the results. However, in all but a few states the basic architecture of performance-based accountability systems remained relatively crude and underdeveloped. In those few states where the idea had been developed most extensively—Texas and Kentucky, for example—the systems worked well enough, according to the testimonials of their sponsors, to legitimate the idea that they were successful in general. Even in these states, however, there were legitimate criticisms of the accountability system’s actual effect on academic performance and drop-out rates.

Nevertheless, by the late 1990s, it was abundantly clear that the states had fallen well short of what the crafters of the 1994 Title I amendments had envisioned. It was also clear that the federal government possessed very little leverage with which to force them along. States varied vastly in their admin-

istrative capacities to implement performance-based accountability systems. Many did not have testing systems in place that met the federal requirements. Most did not have the capacity to administer and monitor testing programs of the scale required to meet the federal goals. More important, creating accountability systems at the state level is essentially a political act, not an administrative one, and the federal government's harmless knuckle-rapping was hardly going to overcome the intransigence of a state legislature or governor. Indeed, the ability of the U.S. Department of Education to monitor and enforce compliance with the 1994 law was limited; budget cuts whittled away at the department's Title I staff just as their responsibilities were increasing. As is always the case, the department's senior political appointees were reluctant to make life too difficult for governors and chief state school officers, who are among their key political constituencies. The rub: by 2000, the target date for full compliance, fewer than half the states had met the requirements. In this environment, it came as no surprise to learn that by the year 2000 many schools with Title I-eligible students were simply unaware of the program's major policy shift in 1994.

This situation should have signaled to the Bush administration and Congress that there were complex issues of institutional capacity at the state and local level that could not be brushed aside by simply tightening the existing law's requirements. If more than half the states were unable or unwilling to comply with the requirements of the previous, less stringent, more forgiving law, why would one expect *all* the states to comply with a much more stringent and exacting law? Part of the problem is political. Even though virtually all the states have joined the accountability bandwagon, for many states it is largely a symbolic act. The basic designs of the systems are still primitive; the authority of state education officials to oversee school districts is still limited in many cases; and the political consequences of imposing large-scale, statewide testing in states with strong traditions of local control are risky. Consequently, support for accountability among state legislators and governors is often highly volatile.

Another part of the problem is administrative. Mounting a statewide testing system is a task beyond the capacity of most state departments of education. Those that have embarked on large-scale testing are stretched to their limits just managing test-development work or monitoring testing contractors. Many states are not doing this work particularly well; others still don't know what the work entails. Still another part of the problem is technical. Standardized tests inevitably become highly politicized and, in the course of the debate, the limits

of testing are subjected to public scrutiny. Many policymakers enter the accountability debate not knowing much about testing, and they often discover, much to their chagrin, that what they don't know can hurt them. Many legislators, for example, are surprised to hear that off-the-shelf standardized tests may not validly measure the content specified in state-mandated standards and that norm-referenced tests (tests that deliberately create a normal distribution around a mean) may not be effective in measuring changes in performance.



### The Capacity Gap

The working theory behind test-based accountability is seemingly—perhaps fatally—simple. Students take tests that measure their academic performance in various subject areas. The results trigger certain consequences for students and schools—rewards, in the case of high performance, and sanctions for poor performance. Having stakes attached to test

scores is supposed to create incentives for students and teachers to work harder and for school and district administrators to do a better job of monitoring their performance. If students, teachers, or schools are chronically low performing, presumably something more must be done—students must be denied diplomas or held back a grade; teachers or principals must be sanctioned or dismissed; and failing schools must be fixed or simply closed. The threat of such measures is supposed to be enough to motivate students and schools to ever-higher levels of performance.

This may have the ring of truth, but it is in fact a naïve, highly schematic, and oversimplified view of what it takes to improve student learning. The work that my colleagues and I have done on accountability suggests that internal accountability precedes external accountability. That is, school personnel must share a coherent, explicit set of norms and expectations about what a good school looks like before they can use signals from the outside to improve student learning. Giving test results to an incoherent, atomized, badly run school doesn't automatically make it a better school. The ability of a school to make improvements has to do with the beliefs, norms, expectations, and practices that people in the organization share, not with the kind of information they receive about their performance. Low-performing schools aren't coherent enough to respond to external demands for accountability.

The work of turning a school around entails improving the knowledge and skills of teachers—changing their knowledge of content and how to teach it—and helping them to understand where their students are in their academic development. Low-performing schools, and the people who work in them,

don't know what to do. If they did, they would be doing it already. You can't improve a school's performance, or the performance of any teacher or student in it, without increasing the investment in teachers' knowledge, pedagogical skills, and understanding of students. This work can be influenced by an external accountability system, but it cannot be done by that system. Test scores don't tell us much of anything about these important domains; they provide a composite, undifferentiated signal about students' responses to a problem.

Test-based accountability without substantial investments in capacity—internal accountability and instructional improvement in schools—is unlikely to elicit better performance from low-performing students and schools. Furthermore, the increased pressure of test-based accountability, without substantial investments in capacity, is likely to aggravate the existing inequalities between low-performing and high-performing schools and students. Most high-performing schools simply reflect the social capital of their students; they are primarily schools with students of high socioeconomic status. Most low-performing schools also reflect the composition of their student populations. Performance-based accountability systems reward schools that work against the association between performance and socioeconomic status. However, most high-performing schools elicit higher performance by relying on the social capital of their students and families rather than on the internal capacity of the schools themselves. Most low-performing schools cannot rely on the social capital of students and families and instead must rely on their organizational capacity. Hence, with little or no investment in capacity, low-performing schools get worse relative to high-performing schools.

Some changes in the new law provide relatively unrestricted money that states can use to enhance capacity in schools if they choose to. However, neither state nor federal policy is currently addressing the capacity issue with anything like the intensity applied to the test-based accountability issue. So an enormous distortion is occurring in the relationship between accountability and capacity, a distortion that is being amplified rather than dampened by federal policy.

## Abusing Tests

During the cold war, just about anyone who raised questions about the distribution of wealth in America was branded a Communist, thus chilling debates over social justice. Debate in the realm of education reform is being similarly chilled. Critics who suggest that there might be problems with the ways in which tests are being used for accountability purposes have been

essentially marginalized. They're smeared with accusations of being against accountability of any kind and of being apologists for a broken system. The idea that the performance of students and schools can be accurately and reliably measured by test scores is an article of faith in test-based accountability systems. Consequently, tests are being misused and abused in ways that will eventually undermine the credibility of performance-based accountability systems.

Probably the most serious problem lies in the use of test scores to make decisions about students' academic progress—decisions about whether they can advance to the next grade or graduate from high school. The American Psychological Association's guidelines for test use, as well as the consensus of professional judgment in the field of educational testing and measurement, specifically prohibit basing any consequential judgment about an individual student on a single test score. The primary reason for this principle is technical, not ethical. Test scores have a significant margin of error associated with them. That margin of error increases as the number of cases decreases; individual scores are typically much less reliable than aggregates of many individual scores. The best that can be said about an individual test score is that it falls within a range that is described by its coefficient of reliability. Unless the range is extremely small, which it isn't for any standardized test, the likelihood of error is high.

The solution to this problem is to use multiple measures of a student's performance when making consequential decisions. But this solution is more expensive because it introduces a new level of technical complexity into the system. For instance, say that high-school graduation was based on a composite of grades, test scores, and portfolios of students' work. Developing such a composite would not only be a challenging technical feat; it would also introduce a certain amount of human judgment into the system. Policymakers tend to distrust the professionals who make such judgments.

A similar problem arises at the school level. Under Title I, schools are expected to meet their annual yearly progress goals. This involves calculating a school's annual gain in test scores from one year to the next. Title I also requires disaggregating these scores by students' ethnic and economic backgrounds. But research has shown (see Thomas Kane, Douglas Staiger, and Jeffrey Geppert, "Randomly Accountable," in this issue) that these measures are highly unreliable for schools the size of a typical elementary school, and they are particularly unreliable for even smaller groups of students. Schools are often misclassified as low or high performing purely because of random variation in their test scores, unrelated to any educational factor.



The standards and accountability movement is in danger of being transformed into the testing and accountability movement. States without the human and financial resources to select, administer, and monitor tests are now being forced to begin testing at all grade levels. This is the surest way to guarantee that the test will become the content. Instead of creating academic standards that drive the design of a standards-based assessment,

current law repeats all of the strategic errors of the previous ESEA reauthorization, only this time at a higher level of federal intervention. The prognosis is not good. The best we can hope for is that the capacity problems of states and localities will become more visible as a political issue at the state and federal levels, triggering responses that will help schools overcome the obstacles they face in improving the quality and

# The standards and accountability movement is in danger of being transformed into the testing and accountability movement.

low-capacity states will simply select a test based on its expense and ease of administration. Thus the criticism that many state assessments are invalid because they fail to test the curriculum that is being taught. The more the issue of validity is submerged in the political debate on accountability, the more likely that charges of “teaching to the test” will be essentially accurate. A test with no external anchor in standards or expectations about student learning becomes a curriculum in itself, which trivializes the whole idea of performance-based accountability.

## Prognosis

The idea of performance-based accountability plays to the greatest weaknesses of the American education system. After World War II, most industrialized countries nationalized their education systems, but not America. Just the idea that students in Louisiana should be held to the same academic and performance standards as those in New York was enough to inspire heated political debate for decades. One consequence of leaving decisions about content and performance to states and localities for so long is that they never developed the institutional capacity to monitor the improvement of teaching and learning in schools, to support the development of new knowledge and skill in teachers and administrators, and to develop measures of performance that are useful to educators and the public.

The difficult, uneven, and protracted slog toward clearer expectations and supports for learning has barely begun in most states and localities. The history of federal involvement in that long endeavor is at best mixed and at worst a failure. The

intensity of teaching and learning. Likewise, it is to be hoped that the technical failures of testing will gain higher visibility and trigger a response that focuses more on the assessment of student learning and less on the administration of tests. The worst that can happen is that test-based accountability will widen the gap between schools serving the well-off and the poor, thereby confirming, at least in the public’s mind, that expecting high levels of learning from all children is unrealistic.

As with many policy innovations, performance-based accountability in education is mutating into a caricature of itself. Never has this been clearer than in the reauthorization of Title I. The idea of giving schools and school districts greater flexibility in return for greater accountability for student performance—the original principle behind the “horse trade” of the 1980s—makes a great deal of sense. What we have discovered, however, is that accountability for performance requires substantial investments in organizational capacity: state departments of education need the capacity to select, implement, and monitor sound measures of performance; schools need support in developing internal coherence and instructional capacity; schools and districts need help in creating reasonable, diverse ways of assessing student learning; and teachers need support in acquiring the knowledge and skill required to reach larger numbers of students with more demanding content. As performance-based accountability becomes test-based accountability, these critical issues recede, and a sensible policy becomes a nightmare.

—Richard F. Elmore is a professor of educational leadership at the Harvard Graduate School of Education and a senior research fellow at the Consortium for Policy Research in Education.