Redesigning Accountability Systems for Education

EDITED BY

SUSAN H. FUHRMAN
RICHARD F. ELMORE

2004

Teachers College, Columbia University
New York and London

Chapter 12

Conclusion: The Problem of Stakes in Performance-Based Accountability Systems

Richard F. Elmore

PERFORMANCE-BASED ACCOUNTABILITY: A WORK IN PROGRESS

If nothing else, the chapters in this book demonstrate that performance-based accountability is much more a work in progress than a finished product. The current message of policy makers and advocates, fearing retrenchment on reforms to which they are attached, is "stay the course." But stay the course with what? As with any policy idea, performance-based accountability, at its best, is a skeletal design—a set of highly provisional ideas about what needs fixing in American education and how it should be fixed—which is played out in a complex institutional, political, and organizational arena. The test of this policy's success is not whether it survives "intact" in this arena, but whether it is robust enough, both in its initial design and in its myriad adaptations to specific problems and contexts, to influence behavior and values in a powerful way.

There is abundant evidence that this policy is more robust than any other in the field of education over at least the past 40 years. Performance-based accountability continues to dominate the policy agenda in states and localities as it has for the past decade-plus—a remarkable accomplishment in a political environment where reform agendas typically have shifted from

year to year. With certain important exceptions, there has been a general increase in the clarity and utility of content standards over time, as well as an increase in the degree of alignment between tests and content standards (see Rothman, this volume).

There is also evidence that the fundamental message that content and performance standards should influence classroom practice has reached teachers in elementary and high schools (see Herman and Siskin, this volume), although not always in the form that policy makers intended. The idea of treating special education students and English-language learners as part of the same opportunity and accountability structure as other students is clearly embodied in policy at the federal and state levels, if not in practice at the local and school levels (Thurlow, this volume). There is a growing body of evidence that student performance, for African Americans and Whites, if not for Hispanics, is increasing, especially in strong reform states, while retention and progression are not adversely affected (Carnoy & Loeb, this volume). The recent reauthorization of the federal Title I program, No Child Left Behind, is an unprecedented use of federal money and authority to promote what has been, up to the present, primarily a state reform agenda, signaling an even longer-term commitment to performance-based accountability.

Another sign of whether performance-based accountability is a robust policy idea is whether policy makers are astute enough to recognize when it is necessary to make changes in policy design in response to new information about the policy's effects. As Fuhrman, Goertz, and Duffy (this volume) argue, there is abundant evidence that states at least are trying to manage the issue of stakes for students so as to maintain the political momentum of the reform while adjusting its specific provisions to the realities of implementation.

It is unlikely then that this reform will recede in the foreseeable future. If anything, political pressure for school performance will increase. This said, it is also clear that the reform's weaknesses and gaps will become increasingly apparent the further and deeper it extends into the complex institutional structure of public education. The issues here are both technical and organizational.

On the technical side, it is evident that what policy makers and the informed public *think* performance-based accountability is, differs considerably from what it *actually is*. In political discourse, it is common to hear both opponents and advocates speak as if test results were the metric of success in performance-based accountability. As Baker and Linn, and Herman (this volume) demonstrate, the idea of equating student learning with test performance is suspect, both in terms of the technical characteristics of tests and the incentive effects of testing on instruction. The key issue

here—probably regarded as excessively fussy and technical by reform advocates—is that no test, no matter how sound, can do more than *sample* what students actually know in a given domain, and even at that, the conclusions one draws from test results about both student and school performance are subject to severe limits on reliability. Using tests as the exclusive measure of performance for accountability purposes can distort conclusions about what students actually know, by substituting knowledge in the sample for knowledge in the domain, and can influence instruction and school organization in counterproductive ways by focusing attention on measures of improvement that do not necessarily represent evidence of strong learning.

The antidotes for these misuses of tests are obvious but difficult to focus on in the rattling din of largely ill-informed debates about testing: strong curriculum-embedded assessments of student learning that are immediately available to teachers as they engage in instruction; strong content knowledge on the part of teachers and administrators so they can distinguish between the sample that the test measures and the domain that defines what students are expected to know; knowledgeable use of tests, with a full awareness of their technical limits; multiple measures of instructional quality and student performance, with no high-stakes decisions based on a single measure; and, above all, adherence to clear principles of test use and design for accountability systems of the type outlined by Baker and Linn (this volume). Most of this advice is currently ignored, or attended to only marginally, in the design and implementation of state accountability systems. A big part of the problem in the technical arena is that policy makers and reform advocates often think they know more about testing than they do, and as a consequence they think they are advocating for and implementing policies that are, in fact, quite different from what they think. Danger lies here.

On the organizational side, it is clear that the complexities of improving schools in the face of performance-based accountability are more apparent to practitioners and researchers than they are to policy makers and reform advocates. As both Herman and O'Day argue in this volume, the fundamental purpose of standards-based reform is not to improve test scores for students and schools, nor is it to get teachers to comply with external directives about what to teach—these are indicators of success, not success itself. The purpose of standards-based accountability is to increase students' access to academic content and to improve the quality of teaching and learning in schools. Depending on the test and the initial performance of students, it is possible, within limits, to increase test scores without significantly increasing either students' access to academic content or the quality of teaching and learning. And the evidence accumulates that the schools

that most need improvements in access, teaching, and learning are often the ones that focus most on test scores and least on deeper improvements. This result is a classic example of what organizational sociologists call goal displacement: The challenging goal set by policy makers—in this case, improvement in access and learning—is displaced in favor of the easier, more feasible goal of teaching test items.

O'Day demonstrates that improving the academic culture of a school—particularly a low-performing school—is much less a problem of complying with the dictates of accountability policy than it is literally one of building an organization around a fundamentally new idea of itself, and this process is multilayered, extending from the individual, to the collegial group, to the school, to the system in which the school resides. Siskin (this volume) argues persuasively that this process of reconstructing schools is dramatically more complex for high schools than it is for elementary schools, and that the task is much more urgent for high schools because they are, by definition, the end of the line and the final reckoning for students.

While policy makers pay lip service to the problems of organizational capacity in schools and school systems responding to performance-based accountability, there is little evidence that states and localities have worked out the actual processes by which schools will become more coherent, instructionally focused organizations. The potential and actual disconnect between the testing side of performance-based accountability systems and the capacity-building side can only become more apparent—and more dangerous—as these systems work their way into schools and classrooms.

A central issue—if not *the* central issue—that joins these multiple concerns about the future of performance-based accountability, is stakes. At some point in the process of assessing students' and schools' performance, the assessing stops and the stakes fall. Accountability without stakes of some kind is a shadow game. In this concluding chapter, I will address the issue of stakes as an organizing theme in accountability policy.

THE PROBLEM OF STAKES: POLITICS, POLICY, AND PUBLIC ETHICS

Performance-based accountability systems operate on the theory that measuring performance, when coupled with rewards and sanctions—one version of what I will refer to here as stakes—will cause schools and the individuals who work in them, including students, teachers, and administrators, to work harder and perform at higher levels (for similar treatments of the theory of action behind standards-based accountability, see Fuhrman, Baker & Linn, Rothman, and Herman, this volume). The idea is appealingly simple: design an incentive structure that rewards students for engaging

their energy in learning academic content at high levels, teachers for teaching a broad range of students more effectively, and schools for organizing themselves to manage instruction more effectively. This idea has achieved considerable social and political credibility with the spread of standards-based, or performance-based, accountability systems at the state and local levels. It recently has become the centerpiece of federal policy with the passage of the No Child Left Behind Act—the revised Title I of the Elementary and Secondary Education Act—which, among other things, requires states to engage in annual testing for individual students in grades 3–8. It sets in place rewards and sanctions based on state-prescribed formulas for annual increments in school performance. And it requires states to disaggregate student performance data by school, based on student demographics.

An important part of the working theory of these policies is that performance-based accountability is a necessary condition for large-scale improvements in student learning and school quality, and addressing the so-called "achievement gap" between poor, minority students and others. Absent a strong and coherent message, carried through the stakes these systems embody, schools will do what they want to do for students, or what they think it is possible to do, without necessarily paying attention to what they *might* be able to do if they were working at higher levels of effectiveness.

Another part of the working theory—less explicit than the former—is that students can be motivated to invest more in their own learning by being given direct feedback on their academic performance, benchmarked against statewide standards, and by bearing consequences, ranging from retention in grade to withholding diplomas, for failure to meet those standards.

Performance-based accountability systems are, to say the least, works in progress. Their designs are still schematic and, in many respects, underspecified. This reality is often lost in the highly charged political debate over the particularities of such systems. As noted above, it is difficult for policy advocates—or opponents, for that matter—to acknowledge that there are many things we do not know about the essential elements of accountability systems. Indeed, there are many things we can't possibly know except by experimenting and observing the results of these systems on the ground.

Nowhere is this question of what we don't know more apparent than in the issue of stakes. State policies require proficiency levels for grade promotion and graduation for students, for example, without any empirical evidence or any defensible theory about how much it is feasible to expect students to learn over a given period of time or what types of instruction have to be in place in order for students to meet expected rates of improvement (see Linn, this volume). Likewise, state policies set expected levels of improvement in schools without any evidence or theory about how schools

actually respond to external pressure for student performance, and whether the ways in which they respond do or do not benefit students (see O'Day and Siskin, this volume). In addition, the tests on which stakes are based are fallible and limited measures; the statements they make about student and school performance carry margins of error for both students and schools, making clear judgments about performance difficult (see Baker & Linn, this volume). These limits of tests are overlooked routinely in current accountability policies.

State accountability policies are essentially political constructs; they represent consensus positions among key political actors about what it is reasonable and essential to expect students and educators to do about academic learning. These policies carry the authority of law. But they are also highly provisional social experiments. Most of the knowledge required to make them work more effectively—to meet the goals that policy makers want to accomplish and deliver the benefits that they promise to individuals—can be acquired only through observing how the policies actually work and developing more elaborate and complex understandings of how students and educators actually respond to the incentives they carry.

Acknowledging what we do and don't know about performance-based accountability carries an ethical, as well as a political, responsibility. If we actually don't understand the underlying parameters of a policy and therefore cannot predict their effects, is it ethical to use the policy to deliver life-altering consequences for individuals? We know, for example, with about as much certainty as it is possible to know anything in social science, that school attainment affects future income, and that attainment is related to cognitive skills that have value in the workplace (Murnane & Levy, 1993; Murnane, Willett, & Levy, 1995). The more years of schooling one acquires, the higher one's income. We also know that graduating from high school carries a substantial income premium, as does any participation in postsecondary education after high school. We know with some reasonable degree of certainty that retention in grade substantially increases the likelihood that one will fail to complete high school. Retention in grade once significantly increases the likelihood of dropping out; retention in grade twice makes it more likely than not that one will fail to complete high school.

Is it ethical, in these circumstances, to deny grade promotion and/or high school graduation to students based on policies that embody highly uncertain theories about the effect of incentives on the behavior of students and educators? In order to make a powerful ethical case for policies like this, one has to argue at least one of three positions: (1) the collective good that follows from the policy exceeds the sum of the individual costs entailed in the policy, and the collective good has been politically determined to be worth pursuing in its own right; (2) the individuals who are hurt by the

policy are, in some sense, responsible for their own fate—they have, in effect, chosen the consequences they are bearing; or (3) those who are hurt by the policy, but not responsible for its consequences, can in some way be compensated by the winners for the damage they have borne.

Another, more slippery kind of ethical problem grows out of the question of whether organizations, or collectivities, actually can be held accountable for their impact on the life chances of individuals. A central premise of performance-based accountability policies is that they hold schools accountable for the performance of individual students. As we shall see in more detail later, the school as an organized entity can be a very elusive construction. Many schools in high-poverty neighborhoods, for example, have highly unstable student enrollments, high teacher turnover, and significant administrative turnover. These schools are the primary targets of the most punitive provisions of most accountability policies. Yet, in what sense are they "organizations," and in what sense can they be held collectively "accountable" for their impact on students? The central fact of their existence is that they have little or no binding capacity to act as organizations, and since many of the people who are present in the organization at time 1 are not present at time 2, it is questionable what or whom one is holding accountable for what. One possible answer is that it is not the school itself but the sponsoring organization—the school system—that is accountable. But, of course, the whole point of accountability for performance is that it is supposed to be located in the place where the work actually is performed, not in some distant place. What does it mean, then, to say that we hold a school accountable for its impact on students, when the membership of the organization is unstable and its very capacity to make binding choices as an organization is questionable? Can we discharge our public responsibility in any meaningful ethical way by charging manifestly incompetent, or incapacitated, organizations to be accountable for their impact on students when they are organizations in name only? Under what conditions does it become plausible to assume that school is actually a school for purposes of accountability?

Another problem arises out of the knowledge and competence of educators in schools. Is it ethical to hold individuals—in this case, educators—accountable for doing things they don't know how to do and can't be expected to do without considerable increase in their own knowledge and skill? It is plausible to assume that educators actually know how to substantially improve student performance, but that they are for some obscure reason withholding this knowledge because they have been insufficiently motivated or rewarded by the existing incentive structure?

The idea that teachers and administrators actually would refrain from doing something they know would contribute to student learning because

they are insufficiently motivated or rewarded seems highly implausible. The more likely possibility, and the one that emerges from research on accountability in this volume and in other places, is that educators literally do not know what to do. That is, they don't possess the knowledge and skill necessary to produce the kind of learning necessary to meet the requirements of performance-based accountability systems, and, more important, the accountability systems themselves don't provide the knowledge and skill necessary to do the work. Whose responsibility is it to provide this knowledge and skill? Is it the responsibility of educators themselves to somehow find out what to do and then do it? Or is it a problem of collective responsibility? If it is collective, in whom does it reside? If it resides outside the school, who has the incentive to provide it and to raise the resources necessary to provide it? More important, can people in schools be held accountable for their effects on student learning if they haven't been provided with the opportunity to acquire the new knowledge and skill necessary to produce the performance that is expected of them?

There is a major ethical problem in the politics of stakes applied to students. When stakes are applied to educators—to teachers, administrators, and the organizations in which they work, as well as locally elected officials who are responsible for governing schools—they are applied to adult individuals who have the means to defend themselves politically against the consequences of the actions levied against them. They can, and do, engage in political action to shape and mitigate the impact of the policies on them.

Students are, by virtue of their age and status, unable to act on their own interests with the same political force and authority as adults. They are represented in the debate on stakes, if they are represented at all, largely by adults, who claim to speak for students' interests, but who have their own individual and organizational interests that usually supersede the interests of the students whom they claim to represent. The claim, "We're doing this for the students," usually means that there is a more or less bald appeal to some interest other than students following close behind. Insofar as students bear the consequences of performance-based accountability policies, then, they bear them as an indirectly represented party to the political debate that shapes the consequences for them.

It is not coincidental that policy makers speak of students largely as passive participants in accountability systems—people to whom accountability provisions are addressed but who are seen as having no active role in determining the nature of these provisions—and that the politics of accountability are dominated by institutional interests—school systems, professional organizations, private-sector advocacy groups—who claim to speak both for their own interests and for the interests of students. In a pluralist

democracy, which rewards the capacity to mobilize and voice political interests, being an unorganized and unrepresented interest is a serious liability. When other organized interests all claim to speak for someone, it is safe to say that none of them do. How do we exercise responsibility for the consequences of policies that fall on individuals who are unrepresented in the processes by which those policies are made and implemented?

Finally, it is important to acknowledge that performance-based accountability systems don't actually *create* stakes for students and educators; it is more accurate to say that they rearrange and redefine stakes. This is a political and ethical point that often is conveniently overlooked by vocal opponents of performance-based accountability systems. They often argue as if there were no stakes for students before the advent of formal accountability systems, which is, of course, manifestly untrue. Students who went to low-quality schools before formal accountability came into play got the same low-quality instruction the day before the systems went into effect as they got the day after, and they were adversely affected by that instruction—that is, there were "stakes" attached to being poorly educated. That these stakes were largely buried and invisible makes them no less real or consequential for the individuals involved. Students in low-quality schools pay a high price for being there. Likewise, educators working in low-quality or mediocre schools also could be said to bear the consequences of neglect—being chronically unsuccessful with students can hardly be said to be satisfying work. Accountability systems don't so much create stakes, then, as rearrange and redefine them, in some instances making the socially and politically sanctioned aims of schooling more explicit and locating responsibility more clearly in specific individuals. It is simply not accurate to say that this rearrangement and redefinition of stakes puts stakes where there were none before.

What is most notable about accountability policies, as they presently exist, is their avoidance of these issues in the details of their design and implementation. As states face the possibility of denying high school diplomas based on graduation exams, retaining students in grade based on proficiency tests, and closing failing schools, the problem of who is actually responsible for student failure has become deeply politicized. Opponents of "high-stakes" testing argue that performance-based accountability systems are inherently unfair to students and teachers, conveniently ignoring the fact that these systems exist because of the manifest failure of many schools to provide adequate learning for these same students in the past. Supporters of performance-based accountability systems respond to their critics by, on the one hand, arguing that it is important to stay the course with stakes in order to demonstrate the gravity of the problems facing schools and students, while on the other hand, publicly and privately exploring ways to reduce or alter the impact of stakes on students and schools (see Fuhrman, Goertz, & Duffy, this volume). Neither side of this debate seems comfortable publicly acknowledging that there may be much that we collectively don't know about how to design and implement accountability policies, that it is important to try to learn something about how these policies actually work, and that the level of uncertainty that surrounds the issue of stakes carries with it the ethical responsibility to be temperate in our actions.

The Theory and Practice of Stakes: Reasoning Out from the Instructional Core

Stated as a problem of policy design, the central issue is *on whom* stakes should fall and *with what consequences* in order to cause the level of *improvement* or *performance* that policy makers want. Stated as a problem of a theory of action, the central issue is the relationship between the allocation and intensity of stakes, on the one hand, and individual and organizational responses, on the other. That is, what kinds of stakes are likely to evoke what kinds of responses in which parties under what conditions?

The key gaps in existing accountability policies lie in the interstices of these questions. What behavior and resources are stakes supposed to mobilize? From whom? What is a good result? What are the preconditions that lead to a good result? And if these preconditions don't exist, how can they be mobilized?

Since the nominal purpose of accountability systems is to increase the quality of academic learning in schools and, hence, to increase student performance, it is hard to imagine a theory of how stakes work in schools that doesn't involve a theory of the instructional core and its relationship to the setting in which it sits. "Something" is supposed to happen as a result of schools and the individuals in them coming to terms with their performance and the stakes that are attached to it. What that "something" is, is largely unspecified in accountability systems.

Student, teacher, content

Many of us who work on issues of policy and its relationship to learning have been deeply influenced by David Hawkins's important formulation of the instructional core as the relationship between the "I" (the teacher), the "Thou" (the student), and the "It" (the content). Hawkins (1974) argues:

> No child . . . can gain competence and knowledge, or know [her]self as competent and as a knower, save through communication with others involved with

> [her] in [her] enterprises. Without a Thou, there is no I evolving. Without an It there is no content for the context, no figure and no heat, but only an affair of mirrors confronting each other. (p. 52)

Becoming a competent learner, Hawkins continues, involves a gradual process of intentional emancipation of the student from the teacher as the mediator of content, accompanied by the development the knowledge, skill, and understanding necessary to become one's own teacher. This process depends heavily, Hawkins argues, on the capacity of educators, and eventually students, to make dispassionate and clear judgments about the extent and depth of their learning in the context of specific body of knowledge.

> The child's overt involvement in a rather self-directed way, using the big muscles and not just the small ones, is most important to the teacher in providing an input of information wide in range and variety. . . . [T]he first act of teaching . . . the first goal, necessary to all others, is to encourage this kind of engrossment. The child comes alive for the teacher as well as the teacher for the child. They have a common theme for discussion, they are involved together in the world. . . . I remember being very impressed by the way some people, in an encounter with a young child would seem automatically to gain acceptance while other people, in apparently very friendly encounters with the same child, would produce real withdrawal and, if they persisted, fear and even terror. Such was the well-meaning adult who wanted to befriend the child—I and Thou—in vacuum. It's traumatic, and I think we all know what it feels like. I came to realize (I learned with a good teacher) that one of the very important factors in this kind of situation is that there be some third thing which is of interest to the child *and* to the adult, in which they can join in outward projection. Only this creates a possible stable bond of communication, of shared concern. (Hawkins, 1974, pp. 55, 57–58)

And, one might say, of common understanding. The purpose of stakes—of *any* incentive designed to affect academic performance—is to mobilize commitment, energy, and knowledge around the student's and teacher's mutual engagement in the content. To the degree that the student and teacher are in concert around this task, rather than in conflict, the level of engagement is likely to be high; to the degree that the student and teacher are in conflict, the level of engagement is likely to be low.

The level of engagement depends in turn on the degree of competence that the teacher and student bring to their work. Being a good teacher means making one's own learning—both as a fact and as a process—manifest in relation to a particular body of content; it means being on display, in some sense, as a learner and modeling this process for students. Being a good student means being a good apprentice to the learning that is manifest

in the practice of the teacher, and, over time, assuming increasing control over that process oneself. The relationship of the student and teacher is disciplined by the presence of challenging content; mastery of content is, in a sense, the standard by which teacher and student judge whether the relationship is about learning, and the degree to which they are learning, as opposed to a personal relationship or a relationship with some other instrumental purpose.

The success of this triangular relationship depends on building a sense of efficacy or agency on the part of the teacher and student. People, in general, enjoy doing what they perceive themselves to be good at, and avoid doing that which they perceive themselves to be unsuccessful at. Low efficacy elicits low engagement; high efficacy elicits high engagement. A successful incentive structure, then, is one that draws the student and the teacher into situations in which they build efficacy and agency. As this happens, many of the rewards of academic work come from the work itself—the developing sense of efficacy and influence over the conditions of one's own learning—and fewer come from the external rewards and sanctions that accompany the work.

In this relationship, the teacher's sense of efficacy comes from the observed effects of her work with the student; that is, the teacher's agency is manifested in what the student produces by way of evidence of learning for the teacher in the moment. Immediate feedback from the student's learning is the most proximate source of motivation and indication of efficacy for the teacher. Getting this feedback requires: (1) that the teacher be competent enough in the content and pedagogy required to engage the student, (2) that the student be willing to engage the teacher at least to the extent that she provides some evidence of learning, and (3) that there is some common means for the teacher and student to understand the joint product of their work.

The student's sense of efficacy comes from a kind of willing suspension of disbelief in which the student agrees to engage the teacher around the content on the—often largely unmet—expectation that she will learn something that will have value, either intrinsically or in relation to some goal the student wants to reach. A good part of creating efficacy and agency in students consists of not just knowing how to teach but also understanding what it is that might have value to the student and making that value explicit in the relationship. When teachers say that some students are "easier" to teach than others, what they are observing is that some students come into the relationship equipped with a set of understandings that lead them to value certain things that the teacher regards as important—or at least they are compliant enough to suspend their disbelief in the lack of value. Some students—probably most students at risk of academic failure—come

equipped with no such understandings and, possibly because of their previous academic experience, are unwilling to suspend their disbelief and may be actively resistant to what they regard as chronically unsuccessful teaching.

A student's academic competence—and therefore her ability to extract efficacy, agency, and value around learning with a given teacher—is a joint product of what the student knows and believes as a consequence of *prior* teaching, as well as the learning that grows out of *present* teaching. When judgments about the effectiveness of teaching are based on student performance at a single point in time, these judgments send very mixed signals to individual teachers and cloud the relationship between the student's learning and the teacher's sense of efficacy. What exactly is the teacher responsible for? The student's performance at a given moment? The learning that the teacher adds to the student's performance as a consequence of their interaction? Or some compound of the two? If the teacher is not responsible for the learning of the student that occurred, or didn't, before the student arrived in her classroom, who is? Holding prior teachers responsible for current levels of learning has value possibly for the present students of those teachers, but no value at all for the student in her present circumstances, since she can't recoup learning that failed to occur in the past.

At some point, the incentives that power the relationship between the teacher and the student lose their traction, failures become cumulative, and we have to invent organizational or collective incentives to minimize the likelihood that unsuccessful teachers will pass their failures on to others with impunity. We will come to this issue in a moment.

Reasoning out from the academic core allows us to think about stakes in terms of their capacity to mobilize and engage students and teachers in the common work of understanding content and in the larger task of passing the work of learning from the teacher to the student. As a general rule, incentives that draw the teacher and student together around the content are more likely to produce higher levels of academic learning than those that don't. Incentives that increase the level of competence of teachers as teachers (in the presence of content) and the level of competence of students as students (likewise) are more likely to produce higher levels of learning than those that don't. Incentives that reinforce the importance of academic content as the mediator of the relationship between students and teachers are more likely to result in higher levels of academic learning than those that don't. And incentives that focus on the learning that occurs between the teacher and student in the present, rather than those that hold the present teacher responsible for the student's past learning, are more likely to engage teachers.

Schools and school systems

It also seems evident that teachers and students who are incompetent at the work of learning have very strong incentives to displace responsibility, efficacy, and agency away from themselves and onto others. If the work of being a teacher and a student lacks meaning, if teachers and students lack the prerequisite knowledge and skill to engage one another in useful ways, and if the conditions of the work are such that it is not clear what the expectations are for what will be learned and how, then it makes sense for teachers and students to blame one another for failure, for teachers to blame previous teachers, and for educators in general to blame the communities and families from which students come. This is a world in which everyone except oneself is responsible for what happens. Low performance breeds low sense of efficacy, which in turn breeds low efficacy, and so on. How this changes is a subject we will return to shortly.

We get to the problem of organizations, as noted above, by confronting the problem that, other things being equal, ineffective teachers can easily pass their failures on to others, and the success of teachers and students at time 2 is heavily mediated by the success of the same students with other teachers at time 1. It is impossible, in other words, to solve the problem of increasing the performance of teachers and students in one classroom without also solving that problem in schools and school systems more generally.

American schools, on average, are notorious for being perilously close to organizations in name only. In the modal school, there is very little interaction among teachers around academic work. Most teachers think of their teaching practice as highly individual and idiosyncratic, and the data support the conclusion that there is more variation in instructional practice and student performance among classrooms in the United States than in any other industrialized country. Content is largely textbook-driven; textbooks emphasize topical coverage rather than understanding, continuity, and depth.

As students advance from elementary to secondary schools, they confront increasingly complex and unintegrated organizations. In many middle schools and most high schools, the subject-matter department is the dominant organizational unit above the classroom, for management purposes, not the school as a whole. Schools at this level are largely organizational fictions, at least in terms of the way they affect the actual work of teachers and students around content. In these organizations, adults function with relative autonomy in classrooms, with minimum oversight on curriculum from the departmental level, and virtually no influence on academic work from the school or system level. Students are the main source of continuity

in these schools since they are the only members of the organization who are required to travel across internal boundaries in order to get their work done. Adults take little or no responsibility for the continuity and coherence of students' experience from one part of the organization to another. Learning, from the students' perspective, is a composite of discrete, often idiosyncratic, experiences accumulated into patterns that may or may not represent progress through a body of knowledge (see Siskin, this volume).

Local school districts, in their modal state, are similarly fictitious as organizations engaged in the propagation of learning. The leadership patterns and policy agendas of school districts are chronically unstable, reflecting the electoral cycles of local school boards and the weak incentives for retention of school superintendents. Systems tend to move restlessly from one policy and administrative initiative to another, with no direct connection to the academic core.

It is difficult to imagine an organizational form that is any *less* adapted to the demands of consistent, high-level engagement of students and teachers around content in the ways described above. Schools, in their modal form, are designed to buffer teachers from virtually any interference in the academic core; schools, and the people who work in them, have limited to no capacity either to influence or to improve instructional practice. Districts, which have nominal responsibility for the improvement of instruction, tend, if they have any capacity at all, to reinforce patterns of volunteerism, idiosyncrasy, and instability of goals in the way they deliver assistance to teachers and schools.

It is absolutely essential to understand that when policies lay down stakes on incoherent organizations, the stakes themselves do not cause the organizations to become more coherent and effective. The stakes are mediated and refracted by the organizations on which they fall. Stakes, if they work at all, do so by mobilizing resources, capacities, knowledge, and competencies that, by definition, are not present in the organizations and individuals whom they are intended to affect (see Herman, Siskin, and O'Day, this volume). If the schools had these assets in advance of the stakes, they presumably would not need the stakes to mobilize them. In this context, stakes make no sense as policy instruments unless they are joined in some systematic way with assistance that is designed to create the organizational assets that are required to respond to the stakes. In the absence of this kind of assistance, most schools and systems will respond within the constraints of their existing assets, which are, by definition, inadequate to respond to the task.

This view accords with the developing knowledge we have of how schools respond to external accountability systems that carry stakes of various kinds. These accountability systems produce a range of responses,

rather than a single type of response common across all schools and school systems (see Siskin, this volume; also see Abelmann & Elmore, 1999). 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; how a school looks at time 2, other things being equal, after the application of stakes at time 1, will be some incremental departure from how it looked at time 0. Hence, schools that have a weak instructional core in academic subjects, and low rates of student access and success in those subjects, tend to respond to new accountability requirements by "gaming" the system—teaching test items, rather than changing access, content, and pedagogy in academic subjects; by focusing on students who are at the margins of the performance levels in the accountability system, rather than the lowest-performing students or all students; and by encouraging certain students to be absent on test days. With increasing pressure, these schools might add academic content and remediation, but, other things being equal, don't tend to make large improvements in their core instructional capacity. Schools that have higher initial capacity in the instructional core—greater access to more-demanding academic content, more attention to success in those domains, clearer expectations for student academic performance, and so on—tend to respond to the external pressure of stakes, even moderate to low stakes, with organizational improvements that give increasing focus and coherence to their existing capacities. Hence, stakes work, if they work at all, by mobilizing and expanding capacities in high-capacity schools and creating potential demand for capacities outside the organization in low-capacity schools. In the latter case, if there are no capacities to bring to the organization, there is little reason to expect the organization to do anything other than make incremental adjustments to already unsuccessful practices.

At the individual and collective levels, stakes work by mobilizing capacities in the service of higher-quality instruction and performance. At the individual level, high-quality instruction requires high mutual engagement of teachers and students in content. To the degree that teachers and students bring the skills and knowledge necessary to be successful in this relationship, they are able to create learning that has mutual value, either intrinsic or instrumental. The stakes, and the external goal they represent, may be a way of focusing this knowledge and skill; they do not create knowledge and skill where none existed before. At the collective level, stakes are refracted through school organization before they reach teachers and students. American schools are modally not well constructed to focus external stakes into a productive relationship with the instructional core; in fact, they are mainly built to diffuse these influences. Hence, when stakes are applied across a number of schools, they produce a range of responses

related to the schools' internal capacities around initial capacities of the organizations they affect.

DESIGNING ACCOUNTABILITY POLICIES

As noted earlier, the goal of accountability systems is improvement in access and learning, not simply reward and punishment for performance. Improvement can be measured, in part, by the assessments used in accountability systems, but these systems do not, in themselves, provide the capacities to improve. Improvement requires high levels of engagement among teachers and students around demanding content, and engagement increases with efficacy and agency. The resources that enable this engagement vary by teacher, by school, and by school system. Stakes for schools and students are effective in promoting improvement insofar as they enable and reinforce this engagement. Several design principles follow from this analysis.

The first is, in some sense, the most obvious and the most difficult to incorporate into existing state accountability policies. It is that individual and collective *stakes should be based on defensible, empirically based theories about what it is possible to accomplish on measured performance within a given period of time*. State accountability systems, and now federal policy, have set in place systems that allocate rewards and sanctions on the basis of schools' progress toward performance goals at rates that are arbitrarily defined, and in some cases probably educationally and psychometrically impossible. The evidence that supports these goals and rates of improvement, insofar as it exists at all, comes from the systems themselves, not from any external assessment, or benchmark, of what it is possible to achieve. Since the accountability systems are corruptible—indeed, there are strong incentives embedded in the accountability systems themselves to make rates of improvement look better than they actually are—they should not be the sole basis for determining what it is possible to achieve. Schools and the individuals in them should not be held accountable for producing results that are educationally or psychometrically impossible.

The kind of research that is necessary to establish external benchmarks for rates of improvement in accountability systems inevitably would raise the issue of the level of resources and capacity necessary to produce results. Having an external benchmark for rates of improvement, then, becomes a way of raising policy issues about capacity.

Demand for an empirically based benchmark for rates of improvement also would force states and localities to develop, in practice, a more specific

working theory of how the accountability system is supposed to produce improvement in student learning. No accountability system currently has such a working theory at any level of specificity that is useful as a guide to action for school administrators and teachers. There are no direct connections, for example, between the assessments that are used to judge the performance of schools in accountability systems and the formative assessments that schools actually would need to have in place in order to judge whether they are making progress with students on a day-to-day, month-to-month basis between testing points. There are no ways of assessing the skill and knowledge requirements of teachers necessary to meet the expectations for teaching practice that will result in the learning that the accountability systems require, or the resources for professional development required to achieve these expectations. There is no clear understanding of how to make curriculum content and alignment decisions that actually support teachers and students in learning that is engaging and has value to them, while at the same time meeting the expectations for performance. Accountability systems are, at this point, policy constructs in search of a theory of action.

The second design principle stems from the accountability standards outlined by Baker and Linn (this volume)—that *stakes should be based on valid, reliable, and accurate information about student and school performance*. Empirically grounded theories of performance and improvement require measures of performance that are sensitive to instruction, that are broadly based enough to represent useful information about the content domains they sample, and that are used in ways that are appropriate to their technical characteristics. It seems reasonable to expect also that the tests should be aligned with clear and understandable content standards (Rothman, this volume), that these content standards should be connected to performance standards that are likewise clear and adaptable to differences among students (see Thurlow and Heubert, this volume), and that the performance measures should be verifiable against other measures (see Carnoy & Loeb, this volume). A central standard in the framework outlined by Baker and Linn is the idea that no decision that has a major impact on a student should be made on the basis of a single measure, nor should students be judged based on a single opportunity to demonstrate performance. A critical part of the theory of stakes is to engage students progressively in their own learning. This process requires accurate and fair assessments of students' knowledge and skill, but it also requires that the incentive structure reward students for persistence, effort, and engagement, not simply for a single performance. Accountability systems, and the stakes they entail, also should work to solidify the relationship between the teacher and the student in the presence of the content—encouraging sus-

tained engagement and ownership of academic work by teachers and students, rather than single events.

The third design principle is that *students should not be held accountable for learning content they have not been taught*. As an ethical and political matter, students are both the clients and the unrepresented constituents of accountability systems. The institutional interests of teachers, administrators, and school systems are well represented in the politics of accountability. Insofar as there are stakes levied against these interests, they have ways of defending themselves politically. Students do not have ways of defending themselves, except by relying on other institutions and individuals who have conflicting interests, or, disastrously, by withdrawing from a system to which they have not consented. 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.

As a practical matter, it makes no sense whatsoever to levy consequences on students for failing to demonstrate knowledge or mastery of content they have not been taught. At present, there are no safeguards in any state accountability system or in federal policy that would establish whether students actually have received instruction in the content that is contained in the tests that they are expected to pass. If schools and school systems were required to specify when, where, and how students were to receive instruction in the content they are expected to master—not in general, as in curriculum guides and course titles, but the actual event in which the instruction took place and its effect on student learning at that time—the question of whether student failure is a consequence of the student's lack of engagement or the failure of the system would become clearer.

Under the current design of accountability systems, student stakes, where they exist, fall unambiguously on individual students, but stakes for educators are highly diffused throughout the organizational structures in which they work. Stakes seldom, if ever, fall with equal severity on individual adults, or on the organizations in which they work. So it is relatively easy for accountability systems, in the absence of countervailing pressures, to ratchet up stakes on students—the unrepresented constituency—and to allow stakes for institutions and educators to become increasingly diffuse. The discipline of having to account for the actual instruction students receive and its initial effects creates a countervailing pressure for schools to pay attention to whether they are actually discharging their responsibility for instruction. If it is impossible to establish whether students actually have been taught the content on which they are being tested, then it is the institutions that should be accountable for the failure rather than the students.

Requiring schools and school systems to account for what they teach, to whom, how, and when also focuses attention on the prerequisite knowledge necessary for students to meet performance standards and the issues of teacher knowledge and skill embedded in academic performance. Are students failing to meet performance standards because they have failed to attend school? Because they have been taught the content in a watered-down form that is not consistent with what they are expected to know? Because they cannot read at a high enough level to master the content? Because the teacher doesn't actually understand the content at a high enough level to meet the standards required of the student?

This principle presumably would require schools and school systems to develop much more individualized ways of understanding why students succeed and fail at academic work, and a much more detailed understanding of individual teachers' contributions to student learning. Most school systems currently have no way of following students' academic progress through the grade structure. Accountability systems operate at much too high a level of abstraction to make this possible; single test scores for individual students over time are neither very reliable nor very effective ways of diagnosing student learning. Tracking student performance over time requires formative assessments that are closer to the ground and more connected to the curriculum as it actually is taught.

The fourth principle is that *schools should be accountable for the value they add to student learning, not the effects of prior instruction; schools systems should be accountable for the cumulative learning of students over their career in the system*. Stakes, insofar as they fall on schools, should be assessed on the basis of what the school actually does or doesn't contribute to student learning, not on the basis of what the student has or hasn't learned in other schools. When a student enters ninth grade reading at the fifth-grade level, the school in which that student resides, and the teachers who engage that student, should be accountable for the increment in learning that occurs in their domain, against a reasonable standard of the progress that student should be able to make with high-quality instruction. The *system* is responsible for the fact that the student is reading at the fifth-grade level, and the system is therefore responsible for remediating the prior deficit that the student brings to ninth grade. To the degree that remediation occurs in the context of a specific school, then it should carry additional resources and these resources should carry additional accountability.

Exactly what it means for collectivities—such as schools and school systems—to be "accountable" for the value they add to student learning is, as we have seen, highly problematic. Collectivities have ways of diffusing stakes; the more pathological the organization, the more likely it is to deflect and diffuse the stakes that are leveled against it. Whereas stakes

focused on individual students cannot be displaced, those focused on organizations can be easily diffused. This problem is endemic to the design of accountability systems, but it can be mitigated by more specific and targeted reporting of performance data. School-level results should be reported, for example, by focusing on evidence of student growth as a consequence of instruction in that school. System-level results should focus on proportions of students failing to meet performance targets on a cohort basis, in the system as a whole, and proportions of students attending low-performing schools. The point is to focus on the type of data, their frequency, and the problems that schools and school systems should be expected to do something about. Existing accountability systems, for the most part, send very confusing, and therefore relatively ineffective, signals about student performance—they do not differentiate between what a school adds to student performance and what that student brings to the school, they do not distinguish between what the school should be expected to do about student learning and what the system is responsible for, and they do not report information in ways that concentrate stakes on those who bear responsibility.

The fifth principle is *the reciprocity of accountability and capacity—for each increment in performance I require of you, I have an equal and reciprocal responsibility to provide you with the capacity to produce that performance*. Accountability systems do not produce performance; they mobilize incentives, engagement, agency, and capacity that produce performance. Accountability systems do not, for the most part, reflect any systematic coordination of capacity and accountability, nor do they reflect any clear understanding of what capacities are required to meet expectations for performance and where the responsibility for enhancing those capacities lies. A more specific and coherent theory of action for accountability systems would help. For example, in order to meet performance expectations in a given content domain, teachers would have to reach a certain level of mastery of that content themselves, they would have to know how to engage students from a variety of starting points in that content, and they would require access to materials and formative assessments that would support their teaching, and the modeling of their own learning for students, in that content domain. Whose responsibility is it to ensure that these conditions are met? If it is the state that initiates the accountability requirement, then it is the state's responsibility to ensure that the capacities are in place to meet those requirements. But who actually provides the support that increases capacity, is a more complex question of comparative advantage. Districts have to have some presence, since they are the systems in which schools operate and they, as noted above, are responsible for performance problems that spill over school boundaries. Private providers might be more efficient in responding to demands for additional capacity,

but they are unlikely to operate effectively in relation to accountability systems if they are not disciplined in some sense by state and local strategies. It is also not clear whether the knowledge and competence actually exist to provide the level of support to schools and teachers necessary to meet the demands of teaching required by performance standards. Are schools actually accountable for their own performance if they clearly can't provide the level of instruction necessary to meet standards, but no one, including the jurisdiction that initiates the accountability system, is able to provide them with support necessary to meet the expected level of instruction? To hammer on low-capacity, low-performing organizations, without providing investments in capacity, in effect encourages them to engage in practices that are not consistent with the goals of the accountability system.

Another difficult incentive problem lies in the domain of resources. There is substantial evidence now from school systems that have launched large-scale improvement processes that there are resources for investing in capacity at the school level that lie within the existing budgets of those organizations. To lay new resources on school systems, without requiring them to reallocate their own resources toward improvements in capacity, is to reinforce their own prior inefficiencies. So the principle of reciprocity also requires states—as the main agents of accountability—to orchestrate their own policies around capacity building, with the requirement that local school systems, and schools, should have to use their own resources first. The strategies for doing this are not well worked out.

The central fact of accountability systems as they presently exist is that they are political artifacts crafted out of relatively superficial and underspecified ideas to meet the demands of political action. They are not well-worked-out practical systems. They require substantial investments in developing working models of what it is possible to produce by way of performance on what sort of timeline, what capacities are required in order to produce these expectations, how those capacities will be provided, and what the impact of stakes, as they presently are constituted, will be on the actual improvement of instruction. These questions are central to the long-term success of performance-based accountability systems.

REFERENCES

Abelmann, C., & Elmore, R. F., with Even, J., Kenyon, S., & Marshall, J. (1999). *When accountability knocks, will anyone answer?* (CPRE Research Report No. RR-42). Philadelphia: Consortium for Policy Research in Education, University of Pennsylvania.

Hawkins, D. (1974). I, thou, and it. In *The informed visions: Essays on learning and human nature* (48–62). New York: Agathon Books.

Murnane, R., & Levy, F. (1993). Why today's high-school-educated males earn less than their fathers did: The problem and an assessment of responses. *Harvard Educational Review*, *63*(1), 1–19.

Murnane, R., Willett, J., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *The Review of Economics and Statistics*, *77*(2), 251–266.