Educational Evaluation and Policy Analysis
Fall, Vol. 7, No. 3, 1985, pp. 261-279

The Design of School Accountability Systems

Guy Benveniste
University of California, Berkeley

Accountability became fashionable in the 1960s. In the 1970s, there was a heated discussion about the pros and cons of accountability. We discussed the need for accountability (Lennon, 1971; Lessinger 1970; Lessinger & Tyler, 1971; McDonald & Forehand, 1973). We also discussed problems and limitations (Bacon, 1978; Barro, 1970; Browdy, 1977; Duncan, 1971; Guthrie, 1979; Olmsted, 1972; Ornstein & Talmage, 1973; Spencer & Wiley, 1981). During the 1960s and 1970s state accountability systems began to emerge. For example, Michigan, Florida, and New York instituted statewide accountability systems. These varied in philosophy, emphasis, and effectiveness. What is new today is a widespread recognition that increased state funding of public education will inevitably trigger new calls for state accountability. Faced with demands for improvements in the schools, legislatures are increasingly attempting to affect policy and outcomes by instituting statewide controls on the schools.

Current trends are to rely on testing, more specifically on standardized true-false testing. There are many reasons for this tendency. Standardized tests are available and useful. They have increasingly been used to point to the successes and failures of American education. In the last decades they have become the best known and the most documented tool to evaluate the state of American schools. Because the tests are not directly linked to the curriculum, they can be used across the board to evaluate many different students or many different schools. They provide an egalitarian tool for centralized control without obliging the state to intervene in the daily affairs of the schools. Or, at least, they seem to.

Centralized control, however, inevitably results in some form of intervention. The current call for statewide accountability systems linked with economic incentives means that standardized testing is to take new importance in the schools. For example, in California, most if not all the proposed new school-based accountability systems under discussion in the legislature use the schoolwide scores of the California Assessment Program (CAP) to provide school comparisons that would be used in a statewide incentive scheme whereby schools achieving high scores would be rewarded financially. It is difficult to imagine that tests that are useful to diagnose students' educational needs and potentialities could be used to evaluate schools without distorting what happens in the schools. The fact that the tests are not linked to the curriculum makes them uniquely useful instruments to assess students or even schools. But the moment financial incentives are linked with the tests they are bound to yield goal

I have been helped by several close colleagues and students: Charles Benson, James Guthrie, Bernard Gifford, Michael Kirst, David Losk, Karen Nelson, Rick Pratt, Ama Torrance, and David Stern.

Dayna Davis, Jean Thompson, and Judy Snow have handled the script. The research was conducted with a grant from PACE (Policy Analysis for California Education, a joint project at Berkeley and Stanford, financed by the Hewlett Foundation).

Authors are always guilty for all omissions, errors, and lack of judgment. This is the case once again.

displacement: the tests will become goals; the curriculum, the teachers, and even the textbooks will start to look like the tests. The main argument of this paper is that central control is inevitable. But central control cannot operate without central responsibility. If the state needs to assess achievement, it will have to address the question of state definitions of a desirable curriculum and will have to involve educators in these definitions.

This paper reviews many of the problems associated with accountability. It goes beyond to suggest a set of principles for statewide accountability systems.

This paper argues that teachers are important. We begin by describing teacher performances we would all like to encourage. We want to establish a consensual nonpolemical view of the teacher's role, a starting point for discussing how to design accountability systems. We then proceed to discuss accountability and its uses: (a) to inform (i.e., provide feedback), (b) to reorient action, and (c) to justify action. This leads us to a more detailed discussion of how accountability actually works. We examine the importance of establishing a linkage with teacher rewards or sanctions and the greater importance of rewards over sanctions in motivating teachers. We come to the inevitable conclusion that the teaching profession, as presently structured, does not provide sufficient incentives. Accountability with little incentive leads to little change. We recognize also that excessive use of accountability, i.e., excessive use of testing of one kind or another, tends to lower the status of the profession. We believe that accountability systems should be parsimonious. They should enhance the quality of life among teachers and not require excessive paperwork.

We distinguish between top-down and bottom-up accountability and suggest a role for both. We discuss the kinds of measures we might want to collect and the uses we might put them to. We argue that standardized achievement testing is unsuitable in schoolwide accountability for four main reasons: (a) when such testing is tied to economic incentives it inevitably leads to goal displacement, even when this is not the intent; (b) such testing, when used in schoolwide scores, masks real issues such as student turnover; (c) standardized testing does not establish a minimum standard and does not tell us how to reward schools trying to deal with difficult students; and (d) standardized testing is not designed to provide sufficient incentives. Accountability systems should provide positive encouragement but standardized tests, by definition, discourage half the population.

We end the paper by presenting a set of principles and the outline of a statewide accountability system that would rely on both top-down and bottom-up accountability.

A Consensual View of the Teacher's Role

Education has many committed schools of thought, and no consensual view of good education exists. There are many accounts and reports on the subject and yet, at the extremes, we still have those who believe that good teaching requires discipline, drill, and practice, and those who believe that understanding requires careful tailoring of material to the specific characteristics of the child (Glaser, 1984). What we do know about learning theory suggests that good teaching has to be adaptive because learners learn in different ways. Therefore, there is no single best way to teach, nor is there any single best way to learn. Teaching and learning are adaptive and they are both uncommonly complex tasks. Good teachers are good because they have learned how complex teaching really is and they use differentiated strategies to achieve learning gains.

Given these facts, we can identify certain characteristics about teaching that are self-evident. For example, the importance of improving the professional competence of teachers, of making the profession more attractive, and so on. Let us list a number of characteristics that should create little or no dissension:

We expect teachers to act as professionals. We expect them to be highly adaptive and innovative; to have a calling and a sense of mission; not to fear to learn and to keep improving their professional skills. This means that they exercise professional discretion and know how to

design learning experiences to fit the varying needs of learners.

We expect teachers to be task-oriented, to enjoy their work, to be committed to the teaching endeavor. Given many different abilities and interests among the schoolchildren they happen to encounter, we expect them to be wise and involved, and to do their best for each pupil.

We expect teachers to teach. We do not really want them to do other tasks and we wish them to resist nonteaching tasks. We are against the encroaching bureaucratization of the schools, which results in more time spent filing forms, preparing plans and reports, and, generally, documenting procedures and outcomes. We therefore want to be parsimonious in designing accountability schemes.

We expect teachers to cooperate with other teachers, administrators, the parents of the children in their classes, and with others in and out of school. We expect them to cooperate with all those whose work makes a difference to the learning task including the teachers in all the feeder schools that form part of the continuous process that begins with preschool and extends through elementary, junior, and senior high school. In short, we like to think that teachers work as a team and that they make choices and decisions that enhance the capability of the team.

We expect teachers to put in time and effort. We know that the quality of education seems to be related to the amount of exposure learners have to instruction.

We expect teachers to conduct learning experiences in an orderly fashion. While we know that adaptation and innovation are important, we also recognize the need for predictability, consistency, and order.

We expect teachers to be confident in their work, to have a sense of accomplishment. Good learning will not happen when those who teach sense their inadequacy, feel overloaded, or are under excessive pressure.

There is nothing unusual in this list, and the reader will think of other important expectations we have omitted. Obviously we have not explicitly stated that we expect teachers to be knowledgeable, and when and where no one knows, to be talented. Indeed, we certainly expect teachers to know what they teach and we expect them to know how to teach what they know. If they are to make wise decisions on how to structure the learning experience to fit the great variety of learners' needs, they surely need to be well trained. The case has been made elsewhere that the training of teachers is in need of much more rigor and much more effort (Stoddart, Losk, & Benson, 1984). We assume just as much when we assert that teachers need discretion, task orientation, and confidence.

Poorly trained teachers, who are not knowledgeable, need to be controlled. Accountability schemes can be designed for mediocre and bad teachers, and they can be designed to encourage good teachers. If one assumes teachers are ill-prepared and incapable of making wise choices, one attempts to limit their discretion. Controls are intended to cope with their weaknesses. Controls, however, can also mean that good teachers are hampered and are treated as if they were no better than the bad ones. This, actually, is a serious problem, and we know enough about the learning process to realize that routines, however well-intentioned, do not necessarily even help bad teachers. Moreover, routines divert attention from the more fundamental issues. What is needed is better prepared, more competent, and more self-confident teachers.

Similarly, those who allocate state resources to the public schools are hard pressed to understand why resources should go indiscriminately to good and bad schools. They are hard pressed to understand why it is not easy to measure what learning takes place in the schools, why we have such a hard time understanding why some children seem to do well in school and why others do poorly. They ask for justifications and for accountability. They ask for measures of accomplishment. As a consequence, today we see that more and more reliance is placed on achievement testing of pupils. We also find that increasing use is made of tests that measure the collective achievements of all the pupils of given schools in certain domains. Some of this testing also seeks to assess value-added learning. By this we mean we seek to measure what skills each pupil had al-

ready acquired at the beginning, say, of the school year, and what skills were added by the end of the year.

These achievement tests can be an important source of information to teachers and administrators. They provide them with individualized information and diagnosis about each pupil. This information can be used to design differentiated teaching strategies. Some tests also provide teachers and administrators with a profile of skill acquisition on a schoolwide basis.

However useful, standardized achievement tests are also intrusive measures. They are intrusive because they are used very frequently and can assume greater importance than they deserve. If parents, pupils, teachers, or administrators come to believe that it is important to achieve high scores, the tests are no longer used as diagnostic instruments; they become a goal in themselves. Much has been said about teaching to the test, and one can argue that this is not desirable because such tests are not designed for this purpose, they are necessarily limited in scope and do not capture all that it is relevant to teach. More importantly, such testing can be manipulated and data falsified. Some teachers and administrators, and even some pupils, may come to believe that it appears to be to their advantage to show high rates of learning during the year. When tests are given twice in the year, it is easy to find ways to do poorly in the first fall test and do as well as possible in the second spring test, thus achieving high annual gains. These gains, however, are only to be lost once the test is taken the next fall, and teachers or pupils do poorly again. Even schoolwide assessments can be manipulated to improve results.

These are not new insights. Much has been done to improve schoolwide assessment. For example, the California Assessment Program (CAP) assesses reading, language, and mathematics in grades 3, 6, and 12. The program is being expanded to grade 8 and to other subject areas. Matrix sampling of pupils is used to assess how well a given school is doing in a number of areas deemed important. Matrix sampling means that pupils only take a portion of each test and scores refer only to the school as a whole. Standardized scores are obtained for each school, and schools are also provided detailed information about the achievement of their pupils in each area so that they can know where they are doing reasonably well and where further effort is needed.

Criterion-referenced tests differ from conventional or norm-referenced achievement tests in that they select specific skills that students should master. The CAP tests, as presently used, cannot be used for individual student diagnosis. Nevertheless, matrix sampling and schoolwide assessments take much less time to administer, are less intrusive on individual teacher performance, and still permit schoolwide assessments.

We shall discuss these tests at greater length later and suggest further improvements. For the moment let us keep in mind that testing for diagnostic purpose is not the same as testing to see whether pupils have mastered a portion of the curriculum.

Accountability in Perspective

Accountability has three main functions: to inform, to reorient action, and to justify what is done.

Accountability serves to inform. For example, it serves to transmit information to the public about what schools are doing or to transmit information to the schools about what the public wants. At more mundane levels, testing in the schools can also help teachers design better programs, and rankings of schools may help parents choose better districts in which to live. When we think of this informative function, we do not mention rewards and sanctions. Information is nonthreatening, designed to help schools, teachers, pupils, and public better understand each other. In this instance adaptation or adjustment takes place naturally.

Accountability serves to reorient action. For example, it may serve to induce teachers or schools to improve on certain tasks and programs. At this point, we need to talk about positive rewards and penalties. It is not enough to transmit information to be heard. A legislature may want to achieve results, give additional resources, or set penalties to achieve compliance. We can design accountability sys-

tems which sample and measure action, compare the measure with a norm, and reward or penalize accordingly. We can design the system to affect individual teachers, groups of teachers, schools, districts, or other populations. If the linkage between the sample measure and rewards is well understood and strong, and if the rewards or penalties are sufficient and effective, individual or group action will be modified.

Accountability serves to justify what is done. It can become a protective strategy. For example, we can design an accountability system that sets desirable norms that we are already meeting. We use the scheme to justify ourselves. In general, accountability is not thought to serve to justify the status quo; but in practice, particularly when measures can be manipulated, accountability can also serve as a defensive strategy in conflicts between schools and public. Thus accountability becomes part of the problem, making it that much more difficult to achieve needed reform. This does not mean that all accountability schemes are automatically used to justify undesirable practice. When accountability measures stress what is relevant and cannot easily be manipulated, they do not hide errors. When accountability deals with irrelevant or hard-to-measure issues, opportunities for obscurations are greater, and may serve only to justify the enterprise.

"Good" vs. "Bad" Accountability

Let us now focus on the use of accountability to redirect action. What are good and bad accountability?

"Good" accountability measures what is important and can also be measured. It does not attempt to appraise when the measures may distort teacher behavior in undesirable directions. This is crucial. Good accountability is not more accountability. Good accountability is the careful selection of specific measures that are or can be available, and that measure what is significant. If we invent an accountability measure and reorient teacher behavior in the wrong direction, we have bad accountability.

Good accountability is tied to positive rewards in preference to penalties. Teachers are human, and human beings respond better to positive rewards. Positive rewards are scarce in education, so the design of good accountability systems has to be tied to increasing the supply of rewards. In education we have to do this with two main considerations in mind: (a) creating an incentive structure within the teaching profession, and (b) designing accountability systems that enhance the status of the profession. These two considerations are linked, and we will discuss them at greater length later.

Good accountability provides information that can readily translate into new patterns of action. It therefore measures what can be altered and not what is beyond teachers' and schooling's ability to change. Since it measures what is important and can be altered, it tends to be supported by teachers. Good accountability incites to less falsification because teachers believe in the importance of the measure. For example, unless there were strong economic incentives to do so, we would not expect teachers to falsify their reporting on how much time they have to spend on nonteaching tasks. Most good teachers resent being taken away from teaching and would prefer to document what happens in hope that the problem can be remedied.

"Bad" accountability is costly. It takes too much time away from teaching. Bad accountability measures what is difficult to measure and provides little information linkage between what is measured and how teachers might redirect their efforts. Bad accountability relies heavily on negative sanctions. It keeps reinforcing the sense of failure that prevails in American education today. It provides considerable information about what is wrong, and little about what is right or what can be done to improve the endeavor. Bad accountability leads to data falsification, which, in turn results in lowered professional ethics, in a lowered sense of achievement, and, most importantly, in false information which is used to protect the status quo.

Bad Accountability and Bureaucratization

Bad accountability is the result of poor design. The underlying assumption behind bad accountability is that teachers

are poorly trained, lazy, and prejudiced. However, instead of attempting to identify a remedy for inadequate training or for the absence of incentives that leads to demoralization, bad accountability reinforces bureaucratization by creating greater uncertainty. In an uncertain environment where it is unclear how teacher behavior might improve the accountability scorecard, a second logical bureaucratic defense is to invent rules and regulations as protective justifications: "How can you blame me for these low scores? I followed the lesson plan to the letter. . . ." Thus, bad accountability engenders more bureaucratization in the schools. Teachers have less discretion, they are less able to adapt to the varying needs of their pupils, to innovate, or to take risks, and more inclined to embrace current fads. So, once again, we find that bad accountability becomes part of the problem.

Bad accountability has further undesirable consequences: it demoralizes teachers. It makes teaching an unattractive profession. It not only reduces discretion, but it also loads teachers with considerable nonteaching tasks. Teachers are burned out because teaching is difficult, teachers sense they are overloaded with large classes, they are told they are inadequate, and, above all, they know that they have to play bureaucratic games to get by. Instead of receiving support and encouragement, they become involved in fads and routines that justify failures and upgrade their accountability scorecard.

Why does bad accountability arise in the first place? It arises because accountability can be used for undesirable purposes. It is a natural bureaucratic defensive strategy. Bad accountability provides defensive explanations for teachers and administrators. It gives the appearance of control and management when no control exists because no incentive leverage exists. It gives the impression of attending to problems, but problems are not attended to because they require real solutions. Bad accountability arises because it is often easier to appear to do something than actually to solve problems. Bad accountability has more to do with appearances than with reality.

Accountability and Measurement

Accountability involves sampling, measuring, comparing results with a norm, and—if we intend to obtain real change—activating positive rewards or negative sanctions.

What should we measure? In practice, we tend to adopt measures of what seems to be relevant, what can be measured at a reasonable cost, and—given the difficulties involved—what is already being measured. There is a natural and quite justifiable propensity to want to measure pupil achievement. However, since it seems difficult to create statewide examinations that reflect the varied curricula of school districts, since it is difficult to reach a consensus about what kind of knowledge all school leavers should have, and since it is expensive to administer and properly evaluate examinations that use problems and large essay questions—as is practiced in many European countries—we fall back on standardized true and false tests which are designed to measure certain kinds of achievement.

These tests are standardized, which means that the questions are tested on small samples of pupils and are made more or less difficult until the population taking the tests is distributed "normally." This means that half of those taking the tests will be doing better than average and half will be doing more poorly than average. Very few will be doing very well, very few will be doing very poorly, and the median and mean will be at the top of the curve.

In general, standardized tests do not tell us whether pupils know what the curriculum intends them to know. They tell us that our pupils are doing better or less well than other pupils, without reminding us that this is to be expected since that is what these tests are designed to do. The tests only give us comparative information about the ability of pupils to understand and answer selected questions.

To be sure, standardized tests can be used over the years and score improvements or losses can be observed. These changes may be due to better or worse education. They may also be due to many other factors: there may be cultural, social, or economic shifts in the population

taking the tests, the children may be better or less adapted to taking tests, they may have experiences that allow them to better understand questions, or they may be more or less motivated to answer them. In any case, since the tests are not linked to the curriculum, we really do not have a sense of what is a desirable score. Moreover, higher scores cannot continually be higher unless the tests no longer differentiate. Therefore if we train our pupils to take the test, and if they do better, the distribution will change. But if the test is restandardized, if the questions are redesigned so that the population will again distribute normally, the same differences will again reappear.

Some of these problems are addressed in standardized criterion-referenced tests designed to measure comprehension in specified skills and subject matter areas. Criterion-referenced tests, however, are still deficient. For example, they often employ true-false answers limiting their coverage of relevant skills. (Interestingly, all true-false testing inevitably downgrades the ability to write essays, yet writing is often a most important skill in higher education and at the higher levels of business and government.) Also, when used in a so-called matrix sample or when scores are aggregated, schoolwide measures do not tell us whether we are testing the same children. In some schools, turnover of students—new students coming in during the year and students leaving to attend other schools or dropping out—is a very high percentage of total enrollment. Therefore, school score variation may have little to do with student exposure to teaching. It would be preferable to use a measure of student achievement which could be allocated retroactively to all classes and schools attended. Lastly, when criterion-referenced tests are not linked to the curriculum, they provide comparative results which do not tell us whether the outcomes are due to the instruction or to other factors. In that situation higher scores may seem desirable, but we still lack a definition of desirable levels of comprehension. We still do not have a minimum standard around which we can judge the performance of schools and pupils.

The impact of standardized testing on the schools is reminiscent of Alice in Wonderland: one has to run to stay in the same place. But it is not even clear that those who run, run in the right direction.

Objective or Subjective Measures

Accountability can be based on objective or subjective measures. In practice, objective measures are often quantitative. They include scores on tests, or any objective data that can be converted into numbers. Subjective measures are more often qualitative, as when we evaluate a school climate on the basis of the subjective perceptions of participants without attempting to quantify these perceptions. Objective measures are useful when we know exactly what we want to measure, when the measures are valid and reliable, and when the measures do not have unforeseen consequences such as displacing or distorting teacher behaviors that are important. For example, if we measure the number of days in the school year or the number of hours teachers spend teaching instead of filling forms, or the number of students in the classrooms, or the number of homework assignments and the time spent on them, our measures (hours, days, months, pupil-teacher ratios) coincide with our concerns. If we attempt to measure student learning, our measures no longer coincide exactly. We invent a proxy measure such as an achievement test, and the achievement test is supposed to approximate some concept of student learning. However, as we have seen, the test only measures certain dimensions of the learning process.

Unfortunately, it is not easy to obtain good measures of time spent in certain activities. Such objective measures are not readily available. It is necessary to depend on self-reporting and on subjective perceptions. As a result, objective measures such as testing seem to be among the few that are readily available. Given the scope limitations of conventional testing, it follows that goal displacement can be a serious liability in accountability. Moreover, it is easier to manipulate or falsify proxy measures that are not there for everyone to see and verify. Motivation to falsify is greater when the measures are not considered to be valid or useful. These are some of the

problems associated with standardized achievement tests.

Objective measures are not necessarily always preferable to subjective measures. They need to be used with care. This means that we need to understand how teachers perceive them, to what extent they understand what they measure, and to what extent they can interpret the measures in terms of their action. Subjective measures must also be used with caution because they, too, tend to be amenable to manipulation and distortion when they are tied to incentives. For example, if we use subjective evaluations of something called "school climate" in a state accountability system, our measures may tell us more about what those who report think we should hear or want to hear than what is actually happening. In general, subjective measures are better used in complex in-depth evaluations where many measures are used. They are better used in site visits and in other in-depth peer evaluations of school performance.

Input, Process, and Output Measures

Accountability schemes focus on inputs, process, or outputs. They sometimes focus on all or on some of these dimensions. Generally, if we have a well-defined goal, if we have a strong theory about how to achieve it, if we understand the process, and if we know what goes into the process, we can design a rigorous accountability system that depends on all three dimensions. This is the case with electric power plants. The goal of generating electricity is well understood. The output is readily measured in kilowatt hours, and the process is well understood and measured in terms of boiler pressure, steam and condenser temperatures, and generator load. The inputs are measured in gallons of fuel oil, and accountability is readily achieved by determining overall plant efficiency. But education is not electricity generation. We have much less powerful theories about what works and what does not. In addition, we need to understand how our measures affect the schools and select only measures that have desirable consequences.

Output Measures

Output measures work best when we know and agree about what we want to achieve, when the measures are valid and reliable and when they have few unforeseen consequences. If we all agree that the schools should place students in college or in gainful employment, we can certainly obtain specific measures of the proportion of the graduating classes that is accepted in institutions of higher education or placed in gainful employment. But we need to be careful and take into account what the schools contribute to such outputs. If we reward schools for placing large proportions of their students in college or in jobs, schools will naturally seek to enroll those students who already have a high chance of succeeding, namely students coming from more advanced backgrounds. But we might be able to correct our control system and reward schools with weighted rewards that take into account school differences. For example, the rewards might take into account the social and economic environment of each school. The problem is that we do not have much experience with such systems, and we should not attempt to use such controls unless we can design a weighted scheme, test it, and determine whether it is effective.

Output measures do not work well when several simultaneous goals are pursued, and some of these goals cannot be easily measured while others can. When incentives are tied to measurements, the accountability system distorts outputs by overemphasizing those that can be measured and downgrading those that cannot. We have already alluded to these problems of goal displacement. Interestingly, while the issue is often mentioned, it is also often disregarded.

For example, the instructions on the California Assessment Program state that the specific content of the tests must not be used to determine curriculum: "It would be contrary to the purpose of the test if curricula were modified to parallel the contents of the test. To do so would conflict with both proper educational and testing practices" (CAP 1983, p.i). Yet several proposals have been made in the California legislature to use CAP testing in a

statewide accountability system tied to economic incentives.

As long as CAP is loosely tied to incentives or sanctions we can assume that goal displacement effects are going to be slight. However, if we implement a strong and effective accountability scheme in which output measures are closely linked to economic incentives, then we can certainly expect goal displacement. If the public schools in the state of California or in other states were to receive significant economic advantages for achieving high CAP scores or high scores in equivalent tests, the tests would become a goal in themselves.

Standardized testing works well for diagnostic purposes because the tests are curriculum-free and can therefore be used across many districts. But state accountability with incentives means central controls. Central controls imply responsibility. If the state uses standardized tests, it will de facto be imposing new definitions of the curriculum. Teachers, curricula and even textbooks will begin to look like the tests. If central control is desired, this requires that new examinations linked to the curriculum be used.

Process Measures

When we measure and control outputs, we say, in effect, "Look here, we want you to place large percentages of your students in college, but we do not care how you do it." When we measure and control process, we say something different. We reduce discretion. We say, "Do it this way." Process measures assume that we know how the task should be done, and we insist that it should be done that way. Process measures reduce discretion. Process measures and process controls work best when we have strong theories explaining how to perform the task, when we know what works and what does not. Obviously there are some things we do believe about teaching and these are amenable to process measures and process controls. We believe that hours spent teaching and hours spent by students learning make a difference. We know that class size and homework are perceived by teachers to make a difference. We know that some order in the classroom, and lack of disruptions, make a difference. But we do not know exactly what style of teaching is preferable for all teachers and all learners. In fact, we know that each learner learns differently, and that teachers need considerable discretion. We do not know which is the best curriculum nor do we know which is the best textbook. We do know that different learners and different teachers do best in different ways, ways that are suited to their unique learning and teaching talent. Much has been said about the importance of certain process characteristics. Time spent learning is a significant variable, and attempts to measure it can be made. Other variables are less well understood. Teacher use of lesson plans, characteristics of the supervision and leadership of the principal, or something called "school climate" all seem to be relevant and important. However, we are much less clear as to what works when, and we are much less able to devise good measures.

Since process measures and controls reduce discretion, they must only be used (a) when we are convinced that we know what works, and (b) when we can devise valid and reliable measures. We repeat again: one problem with some process measures is they are based on self-reporting and are therefore prone to falsification if the measures are tied to strong incentives.

Input Measures

Given the many problems we have described, it is not surprising that input measures remain most important. The question is whether we can be more systematic in collecting them.

Input measures, as the name implies, are measures of what goes into the task to achieve results. When we look at a budget we look at an input measure. We say, in effect, "Here are the resources, are these adequate to achieve results?" When we say that teachers should be better prepared and when we list their qualifications we also use input measures. When we speak about the ethos and norms of the profession, about the values and commitment of teachers, we talk about input variables that may be difficult to measure but are no less important.

Input measures and controls work best when the task to be performed is complex,

when many different goals are pursued and not easily measured, when the process has to be varied and adaptive, and when considerable discretion is needed to meet varied task needs. In short, it happens that teaching and learning are the kind of human activities that are most suited to input measures and controls.

Much more attention could be paid to input measures and controls that demonstrate that well-trained teachers are employed in the public schools. More incentives could be given to those schools that are able to upgrade the qualifications of their teaching staff, more incentives could be given to attract good teachers to difficult schools, and more incentives could be given for attracting qualified teachers in important subject areas.

Positive Rewards vs. Negative Sanctions

We have said that accountability schemes are used to inform, reorient, and justify action. One can inform by providing facts and figures. To reorient action, accountability needs to be linked with positive rewards or with negative sanctions. It is generally recognized that positive rewards are a stronger motivation of action than negative sanctions. Unfortunately, in a world of scarce resources, the availability of positive rewards is far less than the availability of negative sanctions. Consequently, we tend to invent accountability systems that, more often than not, rely heavily on negative sanctions. This is the case in education, where the use of negative sanctions dominates efforts to control the schools.

The reader will have to excuse us for mentioning standardized testing again, but there is no better evidence of the use of negative sanctions in education than the use of such tests. As mentioned previously, standardized testing is designed so that the population taking the test will distribute as close as possible to a normal distribution. When the mean and median coincide, it implies that half of those tested will do less well than average, and the other half will do better. We design the test, and therefore design our principal accountability system in education, to tell half of the population that they are doing poorly; only half are encouraged to know they are above average. We do not treat other human activities that way. We do not do this in higher education. We do not ask our colleges and universities to tell half our students they are below average, and we certainly do not fail half our students. Colleges and universities may have suffered from grade inflation, but grade inflation may also have to do with designing incentives for good work.

Here is a more striking example. Beauty and charm are probably distributed normally in the population. But we do not expect to improve marital relations by measuring where our partners fit in this distribution. We do not wake up in the morning and say, "Darling, you only score in the tenth percentile on the beauty and charm scale and I want you to try and improve yourself." We do not expect marital relations to thrive with this kind of measure. We say instead, "Darling, you are so charming, please get me some coffee. . . ."

The education systems of Europe and elsewhere do not use standardized testing to the extent seen in the United States. Certainly they do not use such testing for control purposes. They use, instead, examinations based in part on essay-type questions and problems. These examinations are closely aligned to the curriculum. The grading strategy does not automatically specify that half of the examination takers will be below the norm and, therefore, implicitly not meet expectations. They set a minimum standard to define who passes and who fails. Choosing the standard allows the examination designers to determine what knowledge is important. It also allows them to relate the level of difficulty with desirable targets of passes and fails. Thus they are able to build incentives into the examinations. They can also set targets for improvement and use the examinations to increase expectations. But these decisions are made by a professional corps of teachers familiar with school reality.

One does not encourage better learning or better teaching by overreliance on negative clues. Most noneducational organizations and institutions that use rewards and sanctions tend to use negative sanctions for only a small portion of the populations they control. They usually use negative sanctions for the lower 10 or 20%

of the target population, and use differentiated encouragement for the remainder. There is no better evidence of this than the reported lessons from America's best-run private corporations. The authors of *In Search of Excellence* point to the importance of incentives and support in successful American corporations. When norms are set for achievement expectations, they are invariably set so that most can succeed. Those who succeed best, the "champions," are constantly encouraged and supported (Peters & Waterman, 1982). These successful corporations even know how to tolerate failure, but more importantly, they rely on their people, they infuse a spirit of success based on a constant affirmation of excellence that defines success in ways that are achievable. They train their people well and expect them to exercise judgment:

> The sole way that company can work is to place its faith in its 2,000 well-trained, perfectly socialized young engineers who are sent to the ends of the earth for months—like the Roman general—and left only with [the firm's] philosophy and this extensive training to guide them. [A leading executive] summed up the problem when he said, "Substituting rules for judgment starts a self-defeating cycle since judgment can only be developed by using it." (Peters & Waterman, 1982, 277–278)

These companies certainly do not use standardized tests and normal distributions to judge success and excellence. They use well-understood standards that are considered to be important, and they also select these standards to create incentives through rewards. The standards are not self-defeating; the companies select them so as to encourage greater effort by making success visible and understood.

These companies also reward success by promoting their champions. Contrast again with our schools. Teachers have no significant career path. The profession is undifferentiated. All teachers do the same work whether they have just graduated from a school of education or have acquired years of experience. Given the vagaries of district financing, they do not have much job security. The only way to have access to higher salaries and to have influence on school decisionmaking is to exit teaching and become an administrator. Thus, most educational accountability systems simply flap in the wind. They use bad measures and are not linked to any incentives. They are only linked to teachers' perceptions of the uncertainties and demise of the profession.

Good accountability systems in education would have to start with a career structure for teachers that provides visible opportunities for advancement, and can be harnessed to provide leverage incentives for teacher achievement. For example, interesting recommendations along these lines were made in *Some Reflections on the Honorable Profession of Teaching* (Stoddart, Losk, & Benson, 1984). These authors recommended restructuring of teacher training, licensing through state examinations, and the creation of new career paths within the profession so that teachers might start as interns, become junior teachers, and move on to become professional teachers with the best becoming specialized teachers and mentor teachers. Similarly, it would seem quite reasonable to design accountability systems that identify the few schools that are in serious trouble so that they might be assisted, and reward and encourage all other schools so that they might further improve. Moreover, some schools might undertake collective research with institutions of higher education and even provide technical assistance to less successful schools. Thus an incentive structure could also be established among schools.

Individual vs. Group Accountability

If we want teachers to work as a team, we need to design accountability systems that reinforce group work instead of individual work. The basic performance unit of the educational system is the individual school. This is not a new idea: as Benson (1972) wrote, "All testing, auditing, information gathering, and incentive distributions should be organized around schools rather than school districts or individual classrooms" (p. 47).

An accountability scheme designed around schools also provides the opportunity to pursue a strategy based on the concept of centers of excellence and the creation of a school incentive structure.

We need also to pay far more attention to the difficult schools. Given insufficient economic incentives, teachers pursue other benefits. One of these benefits, which acts as an incentive in teaching, is to locate in a better school. These tend to be the schools that attract students with more homogeneous upper SES backgrounds. If there is no incentive for staying in low SES urban schools, these schools will have a greater share of mediocre or bad teachers. Accountability systems can be designed to reward efforts in the more difficult schools at the same time they reward efforts in the better schools. In other words, school accountability incentives can take into account the SES background together with the racial and linguistic diversity of students. They can create incentives that attract better trained teachers into the more difficult schools and can reward those schools that successfully upgrade the qualifications of their teachers. Steps in that direction are being taken in some school districts. Statewide accountability systems will expand and reinforce these efforts. But single schools are not the only relevant unit. Students go to various schools; they start in preschool and move on to elementary, junior, and senior high schools. Often these schools are in different districts, yet significant numbers of student flow from one school to the other. It is also desirable to foster cooperation among feeder schools. There may be many ways to do this; one possibility we will discuss later might be to attribute scores and economic incentives in a final school graduation examination not only to the high school but to all the schools attended by the graduate.

Incentives and Status

Status derives from perceptions. How teachers perceive their jobs affects the status of teachers. How others feel about teachers also affects their status. To be sure, salaries, and other emoluments, affect how teachers and others feel about teaching; teacher's salaries will continue to be low relative to other occupations, thus making the profession less attractive. But salaries are only one feature of the attractiveness of teaching.

In Europe, teachers generally enjoy far more status than in the United States. There are many cultural reasons for this, but one factor is that in several European countries, the careers of secondary school teachers are linked to those in higher education. Once one has obtained the higher education degree needed to teach, it is possible to teach both in secondary schools and in universities. Even if few can do it, some secondary school teachers can gradually rise in the ladder and ultimately be promoted to university appointments. The fact that this is possible increases the status of teachers. There are other factors at work. For example, these systems, including the British, have an upper cadre of school inspectors who play a special role in the control and promotion of teachers and in the design of curriculum. The inspectorate is generally recruited from the ranks of the profession. Thus, European systems, in contrast to the American pattern, seem to have far more diversified teacher career structures and more opportunities for teachers to play differentiated roles. We can safely assume this is one explanation of their relative status.

This does not mean we can adopt these European models, but it does remind us that status is enhanced when careers are perceived to be selective (not everyone can practice) and the career has a diversified and increasingly selective hierarchy (not everyone can climb, but some do).

Furthermore, status is also related to levels of discretion and responsibility. The more we control teachers, the more we invent means of reducing their discretion, the more we reduce their status. For example, when we impose a "teacher-proof" curriculum, we also tell parents and others that we do not think our teachers are any good and in so doing we reduce their status. The medical professions illustrate the extent to which status is associated with discretion, the ability to make choices and decide outcomes. Doctors have much status, in part, because they have much discretion. Nurses have much less status and also much less discretion. This suggests that we must be careful to protect teachers' discretion not only because it is good pedagogy but also because discretion enhances status.

Accountability is linked to status in two ways. First, we need a status structure as incentive to make accountability work better. Second, accountability systems can enhance or reduce the status of the profession.

In general, in the world of work, accountability is differentiated by levels of discretion. In most work situations, those who begin on the job are far more controlled than those with more experience. In this way, accountability contributes directly to the status system: when you start in some work situations, you punch a card—this is a control of how you spend your time. As you climb the status ladder, control on time spent is gradually relaxed. You know you reach higher levels because you have greater responsibility, and greater trust is invested in you. We do not have a differentiated accountability structure in education today. We could have one, if teachers were involved in setting normative standards for themselves and for their pupils. Even if few teachers could reach the upper echelons and responsibilities of their professions, the status ladder would exist and act as an incentive leverage.

More importantly, the more we routinize teaching and the more we impose procedural rules, the more we reduce discretion and downgrade the overall status of the profession. There is nothing wrong in routines when the task at hand is repetitive and predictable. But when one imposes routines on tasks that are highly variable—and teaching is such a task because each child is highly differentiated—one simultaneously downgrades the status of the profession and hampers the ability to perform. This in turn further lowers the perceived status of teachers.

One lesson we can glean from Japanese management practices underscores this point. Japanese managers trust their employees. They do not evaluate them as often as American firms are prone to do. They use evaluations at important stages in the career. But they are parsimonious. They know that constant evaluation reduces discretion and status and therefore reduces opportunities for innovations, creativity, or risk taking.

Americans are abusing standardized testing in the schools. Testing can enhance the teachers' ability to perform. It helps them diagnose their students or their own teaching. But such testing is ill-suited to accountability systems with schoolwide rewards or sanctions. If we want to control students then we should do it only at important stages in the student's career, and we should use examinations that (a) are linked to the curriculum, (b) establish a minumum that defines pass and fail, and (c) provide rankings based on accomplishments. Similarly, teachers should never be evaluated on the basis of their pupils' performance on standardized tests. They should be evaluated much less frequently, but the evaluations should be thorough. In-depth peer evaluations using many objective and subjective measures could be made not only by principals but also by top teachers coming from different schools.

Top-Down or Bottom-Up Accountability

The unit school can, in most circumstances, become the basic performance unit for new accountability systems. In the last decades, the number of accounting and accountability requirements has multiplied as most federal and state programs legislate reporting, documenting, and other management controls. Efforts are constantly needed to rationalize overlapping control requirements and to centralize their administration. Central school district administration normally has the responsibility for processing all accountability reporting requirements. Computerized management information systems permit handling of large data bases and decentralized school-based input terminals provide rapid reporting systems tied to the district information management systems. Similarly, integration is needed at state level. Decisions about data gathering and distribution cannot be taken arbitrarily and require coordination and integration. However, the usual organizational arrangement is for data to be collected by many different offices and agencies in the state government with no single central body responsible for deciding what data to collect and how to distribute it. The creation of statewide educational accountability systems inevitably requires coordination. This means placing responsibility in a single central

body charged with setting policy for school accountability.

Producing schoolwide accountability data and making it available provides new information to interested parents and pressure groups. One consequence of measuring and making information available is that it provides knowledge to political actors who are genuinely interested in what happens in the schools. These political actors, in turn, begin to exercise greater pressure for improvements. If the information and measures are readily understandable, the action of these political actors can be purposeful and effective. They may act at the local, state, or even national level. It is not only important to collect data. It is also important to know for whom it is intended.

Accountability systems can be designed to operate top-down or bottom-up. A top-down design is one that provides centralized incentives. For example, the state may decide to provide additional financial incentives to schools that reach selected well-defined standards. A bottom-up design may still centralize data gathering—so that all schools may be required to gather and disclose certain kinds of data—but this information is not linked to central incentives. The information is made available to grassroots interests and to others in the hope they will find ways to remedy deficiencies.

There is not a rule that says that top-down or bottom-up accountability is better. Obviously the center has considerable prestige and legitimacy, and in selected instances central top-down directives do provide the leadership needed to initiate reforms. But top-down accountability means centralization, and while some centralization is warranted, we know enough about the diversity of schools to know to proceed cautiously (which does not mean one should not proceed). There exist many top-down opportunities to enhance the status of the profession and of the schools.

Centralized schoolwide accountability data provide both better information about schools and opportunities for creating new incentives. Competitions, prizes, demonstrations, and other events can be organized. Successful schools can sponsor activities while acquiring visibility and status. One can imagine statewide accountability systems that organize schools in different categories involving some schools in helping others, giving to some schools enlarged responsibilities and tasks. Similarly, one can imagine a much more selective and differentiated corps of teachers with some teachers involved in the evaluation of other teachers and in the elaboration of statewide examinations, and some involved in development and research programs linked with universities and research centers. Top-down accountability need not be downgraded, but it is a complex activity that goes far beyond tying incentives to schoolwide scores.

Bottom-up accountability is particularly important where (a) implementation depends on local participation and support, (b) problems are diverse and peculiar to local conditions, and (c) measures need to be interpreted in light of local conditions. Bottom-up accountability is decentralized accountability. The American school has long benefitted from a unique system of decentralized governance. Statewide accountability is a move to centralization, but the design can be flexible. It can centralize in some promising areas and decentralize in others.

Outline of an Accountability System

Given all these considerations, what might be the elements of a schoolwide accountability system? Our discussion now moves ahead into illustrative examples.

A Classification of Schools

All schools are not alike, and, unfortunately, a few schools are very deficient. Most schools probably do reasonably well and could be encouraged to do more. Some schools are close to universities and research organizations. These schools have the capability of engaging in more research and development, not only in response to researchers' preoccupations but, more importantly, in response to school-felt problems. One historical deficiency of American educational research is that it depends too much on researchers' definitions of problems. It would be desirable to rate schools on their ability to take a greater leadership role in defin-

ing educational research priorities so as to involve them in more cooperative research. Other schools may be less inclined to initiate much research and yet, because of their sophistication, experience, and successes, they have the capability of exporting their experience to other schools. These schools could become involved in technical assistance to deficient schools.

Our School-Based Accountability Scheme could be designed to rate schools in five classes:

1. Below State Requirements.
2. Meets State Requirements.
3. Improving School: a school involved in a development program designed to increase its performance.
4. Research School: a school involved in a research-oriented development program designed to increase its performance.
5. Mentor School: a school that may be involved in research, development, or technical assistance to other schools.

The ratings would take time to develop. They would have to be sensitive to a number of variables such as urban/rural location, student turnover, and SES and linguistic composition of student body. Scaling and scoring would be within categories so that schools would compare with similar schools, as is suggested in some of the recent Californian proposals for accountability (Honig, 1984).

Top-Down Incentives

Additional resources and prestige would be provided by state agencies. One can imagine statewide competitions in certain domains, prizes, ceremonies, and other status-giving activities. Top-ranking schools would have access to added resources, schools seeking to upgrade their ratings would have access to technical assistance. Schools that failed to meet state minimum requirements might have to agree to a program of state guidance, assistance, and self-help. Schools that never meet minimum requirements might have to be reorganized or merged, and, if all remedies failed, they might simply have to be closed. The system would strengthen centralized control; however, incentives would be used not only to incite toward higher performances but also to involve some schools—and therefore some teachers—in new, different activities they are not now accustomed to.

Bottom-Up Accountability

We would want to have ratings on several dimensions, and the statewide ratings would have to be limited to fairly reliable measures. The design of our accountability system would not link measures to state incentives when these require local interpretation and when they might encourage excessive falsification. In other words, in the system, some measurements would be used for bottom-up accountability. Here the state would still take leadership in asking that data be collected and might provide technical assistance to school boards or other local groups in interpreting the results. What might these measurements be? It is too early to say. But in general they would be measurements that are difficult to obtain. For example, we would include measures of time spent on homework in this group. Definitions of time spent in homework are not easily arrived at, and data across schools might not be comparable. Moreover, it is not clear how schools, parents, and students would react if economic incentives were tied to such data. On the other hand, it might be useful to be able to compare how much time the children of various school districts seem to spend on homework. Local parent groups, school boards, and others might make better decisions if such information were available to them.

Parsimonious Measurements

Some measurements could be routinized—for example, data on qualifications of teachers or data on length of the school year. But testing data would have to be used with parsimony. If teachers had to take the equivalent of a State Bar Examination very early in their career, that examination could be used not only to select teachers but to rate schools also. Similarly, curriculum-linked student examinations would have to be developed, or where they already exist, would be used in the ratings. But the examinations would be few, and would have to be adapted to district or school differences.

A List of Measurements

What might we measure? We suggest (a) teacher preparation and achievement, (b) teacher use of time, (c) student learning time in selected subject areas, (d) order and consistency, (e) parent and community support, and (f) selected student outcomes.

Measures of teacher preparation and achievement. These controls would be based principally on teacher preparation and teacher promotions. They could include weighted averages of number of teachers credentialed; results of any statewide professional teacher examinations; percentage of teachers in each subject area; percentage of teachers in various levels (i.e., when states adopt career structures for teachers, one would want to know how many interns, junior teachers, professional teachers, specialized teachers, or mentor teachers were employed in each school).

This is input accountability. Our purpose would be to better assess where well and less well prepared teachers go and to be able to compare teacher configurations on a school-by-school basis. Norms might ultimately be established. Programs could provide incentives to assist those schools and teachers desiring (a) to upgrade their training and qualifications, and (b) to distribute skills in each school. Specialized teachers in mentor schools would be able, and expected, to carry on programs of technical assistance and training in other schools. Mentor and Research Schools would join other schools in upgrading efforts. State resources would be used to encourage talented teachers to go to difficult schools.

Measures of teacher time. The purpose of these measures would be to provide incentive for expanding effective teaching time and for the debureaucratization of the schools. Teachers would self-report approximate time spent on teaching and nonteaching tasks. Efforts to reduce paperwork might be reported. We would also want to develop some measures of teacher-principal interaction. We would want to be able to compare teacher/principal ratios and how time is spent in supervision.

These process controls could provide bottom-up accountability with school-by-school comparisons. Some of them (i.e., reduction of paperwork) might be used in statewide competitions. In time, norms for acceptable levels of nonteaching time could be established. Desirable teaching time distributions and the expansion of teaching functions (i.e., participation in teaching improvements instead of spending time on reporting) might again provide norms for comparisons and improvement.

Measures of student learning time. These might be approximated by course enrollment data, turnover rates, pupil/teacher ratios, school day activities, length of school year, and out-of-school learning time. The purpose would also be to design reporting and incentives to encourage greater student flows into selected domains such as reading, math, science, history, literature, art, and ethics. Also data on average time spent on homework, counseling, and remedial work might be obtained.

Again, these might best be used for bottom-up accountability. The purpose would be to control how students spend their time and to increase time spent on certain subject areas. The cost of reporting might limit what could be done, yet much more might be achieved so as to be able to make useful school-by-school comparisons.

Both of these process measures (time spent teaching, time spent learning) may be difficult to obtain. Yet new school management information systems should be able to generate such data. Time, by itself, does not tell us too much about the quality of teaching or learning which takes place. But combined with teacher qualification measures, time accountability can begin to tell us more about who does what and when, and provide ideas for remedies and ways of handling deficiencies.

Measures of order and consistency. Our purpose here would be to measure and identify problems of truancy, absenteeism, vandalism, and disruptions in the schools. We would also want measures of student turnover. An accounting scheme could be readily established to provide a list of schools requiring priority attention and help. We would also want to develop and use measures of student cooperative

behavior in school. There have been interesting programs designed to foment student cooperation and interest. It would seem desirable to build such programs into any bottom-up accountability scheme. State prizes and other encouragements might also be provided (Wynne, 1984).

Measures of parent and community support. These would include school worker volunteer hours, parental volunteer hours, total dollar resources contributed by individuals and private organizations, other income and contributions. Our purpose would be to assess, publicize, and encourage community support of the schools. Measures of volunteer support would also be supplemented with information on school-by-school funding, thus providing comparative information, school by school, on the distribution of local, state, and federal funds. A statewide school-by-school unit cost tabulation would provide new insights on the way resources are allocated.

These input measures could be tied to a set of state incentives to further private contributions particularly to schools serving communities with few private resources. School-by-school information about resource availability and the contributions of involved communities could lead to greater efforts toward fomenting parent and community involvement.

Student ability and outcome measures. The reader must have already become aware that we believe that standardized achievement testing should be used only for diagnostic purposes. The results of testing should be provided to teachers to assist them in planning their teaching strategies.

Standardized criterion-referenced testing would be used to diagnose and advise schools as to apparent deficiencies. These programs should and could be expanded to cover more subject areas. However, because of the nature of these tests we do not believe that incentive schemes should be directly tied to such school scores. As mentioned earlier, such testing is beneficial as long as the tests do not influence the curriculum. If strong incentives or sanctions were to be linked to the tests, we can safely assume the curriculum would naturally adapt itself to the narrower objective of improving student scores.

Some student outcome measures would be used. In our design, we would start by asking teachers to design a single statewide examination which would set minimum requirements and evaluate higher accomplishments. In California such an examination would replace district-generated minimum standard testing under AB3408. Some results from this new state examination might have to be evaluated differentially for each category of schools. Or we might find that some portions of this examination would be differentiated and adapted to the needs of pupils with different cultural and linguistic backgrounds. For example, we might have different portions of the examination for high academic achievers and for vocationally oriented students. We might correct or take into account whether English is a first, second, or later language. We might design different portions to fit what is desirable preparation for college or work. Certain sections of the examination might be optional. We might centralize certain portions of the examinations and decentralize others. Obviously, the opportunities for implementation are many, and much work would have to go into the elaboration and testing of such examinations. We would expect teachers to play a dominant role in this process. We would also expect them to play a dominant role in administering and grading the examinations.

We would also want to use essay-type questions and problems in these examinations. The examinations would be aligned to the curriculum, which suggests that some portions might use the familiar format of current testing and others might rely on other formats. In any case, such examinations should be conceived, administered, scored, and evaluated by an elite corps of mentor and specialized teachers who would be given the necessary time to carry out the task.

The scores of all graduating students and of those failing could also be used in weighted incentive schemes that would allocate results across all the schools that had been attended by each student. Our purpose would be to create new incentives for greater collaboration between

high schools and their feeder schools. A student with high scores would provide credit both to the high school, the junior high, and the elementary schools attended. To be sure, there would be a time lag before an elementary or junior high school might be credited. However, we suspect that if schools knew such a scheme existed, they would respond and some collaboration would be obtained once the scheme is established.

In California, this minimum requirement examination could also be combined with an expanded use of the Golden State Examination under AB 813. This latter examination could continue to be taken on a voluntary basis, but the scores could be used in a weighted measure with those of the state minimum requirement examination.

These statewide examinations might also be supplemented with additional information about numbers and proportions of students completing programs, dropout rates, rate of admission in various levels of postsecondary education, and rate of placement in gainful employment.

We might also want to measure student inputs, namely student ability. For example we could measure student IQ and contrast overall school student ability with student outcome measures. This would help us understand which schools are more successful than others in helping students with different abilities.

We would not use aptitude tests such as SAT scores in our top-down accountability schemes, since aptitude tests measure student characteristics that are not necessarily attributable to schooling. Similarly, we would not use advanced placement tests, university reading and mathematic diagnostic tests, or grades achieved in college in our accountability system. One reason we would reject some of these measures has to do with our concern with teachers. We would want the top-down accountability system to enhance the profession by giving it more responsibility. Therefore, when selecting measures, we would prefer measures that give more responsibility to teachers and less to the institutions of higher education. This is why we would not advocate rewarding schools on the basis of the performance of their students in institutions of higher education. We would prefer to see a cadre of school teachers acquire responsibility for certifying their outputs. We have little doubt they, and others, would pay close attention to the match of their assessment with those of other institutions. We would therefore collect such data but use it in bottom-up accountability.

Conclusions

The main point of this paper is that more accountability is not necessarily better. Better accountability means that we are more concerned with attracting good people to teaching and more concerned in making teaching a desirable profession. Therefore, better accountability means finding ways of making teachers, students, and the community more responsible and more committed to the task. It means that we are concerned with using accountability to increase the status of teachers, and similarly, we are concerned with increasing the value of a school diploma. We describe a course of action that would increase the central controls of the state on public education. This turn of events is as it should be. It reflects the increasing role the state plays in financing the public schools. Accountability here means accountability to those who are responsible and who provide most of the needed resources. We describe a style of accountability that is highly flexible, leaving more discretion to teachers, to schools, and to local initiative. Thus, both centralization and decentralization are pursued simultaneously.

Our values and tastes will change. What is important today may pale tomorrow and what is not important today may be perceived as such later. No accountability scheme can be permanent. Moreover, our knowledge of schools, or of teaching and learning, will improve as time passes. We will gradually know more as we develop new measures. More importantly, we cannot know beforehand whether this scheme or others will benefit the schools. The message is clear: we need to proceed with caution.

Proposals of this nature will require considerable discussion before any implementation can take place. Statewide ex-

aminations of teachers and pupils represent a radical departure from current practice. But current practice is not sanctified and cannot be expected to meet the needs of a changing environment. These proposals are, by themselves, indicative of new trends in California and elsewhere in the country.

References

BACON, W. (1978). *Public accountability and the schooling system*. New York: Harper and Row.

BARRO, S. M. (1970). An approach to developing accountability measures for the public schools. *Phi Delta Kappan, 52*, 196–205.

BENSON, C. ET AL. (1972, June). *Final Report to the Senate Select Committee on School District Finance, Vol. I*. Sacramento.

BROWDY, H. S. (1977). The demand for accountability: Can society exercise control over education? *Education and Urban Society, 9*, 235–250.

California Assessment Program. (1983). *Examiner's Manual 1983*. California State Department of Education.

DUNCAN, M. G. (1971). An assessment of accountability: The state of the art. *Educational Technology, 11*, 27–30.

GLASER, R. (1984). Education and thinking. *American Psychologist, 39* (2), 93–104.

GUTHRIE, J. W. (1979). Educational accountability. *Proceedings of the Academy of Political Science, 33*, 24–32.

HONIG, W. (1984, February). *Education reform: Next steps*. Unpublished manuscript, California Department of Education, Sacramento.

LENNON, R. T. (1971). To perform and to account. *Journal of Research and Development in Education, 5*, 3–14.

LESSINGER, L. M. (1970). *Every kid a winner: Accountability in education*. Palo Alto, CA: SRA, Inc.

LESSINGER, L. M., & TYLER, R. W. (1971). *Accountability in education*. Worthington, OH: Charles A. Jones Publishing Company.

MCDONALD, F. J., & FOREHAND, G. A. (1973). A design for accountability in education. *New York University Educational Quarterly, 4*, 7–16.

OLMSTED, R. (1972). Review of *Every Kid a Winner*. *Harvard Educational Review, 42*, 425–429.

ORNSTEIN, A. C., & TALMAGE, H. (1973). The rhetoric and the realities of accountability. *Today's Education, 62*, 70–80.

PETERS, T. J., & WATERMAN, R. H., JR. (1982) *In search of excellence: Lessons from America's best run companies*. New York: Harper and Row Publishers.

SPENCER, B. D., & WILEY, D. E. (1981). The sense and the nonsense of school effectiveness. *Journal of Policy Analysis and Management, 1*, 43–52.

STODDART, T., LOSK, D. J., BENSON, C. S. (1984). *Some reflections on the honorable profession of teaching*. Berkeley: University of California.

WYNNE, E. A. (1984). School award programs: Evaluation as a component in incentive systems. *Educational Evaluation and Policy Analysis, 6*, 85–93.

Author

GUY BENVENISTE, Professor of Education, School of Education, University of California, Berkeley, CA 94720. *Specializations*: Organizational analysis, policy research.