Data-Based Accountability in Education

James W. Guthrie
Michael W. Kirst
(Editors)
June 1984

84-4

This is a PACE Project sponsored paper. PACE, Policy Analysis for California Education, is a joint undertaking located at the University of California, Berkeley and Stanford University. Its Directors are James W. Guthrie and Michael W. Kirst. PACE is funded by The William and Flora Hewlett Foundation. However, the analyses and recommendations contained in this paper are not necessarily endorsed either by the Hewlett Foundation or the PACE directors.

Data-Based Accountability in Education

TABLE OF CONTENTS

Foreword

Providing Incentives and Constructing Measures for California School Performance

For more than half a century California has been moving toward a greater role for state government in educational policy formation. This trend was substantially intensified by the June 6, 1978 passage of Proposition 13. Enactment of this initiative, which drastically reduced local property tax revenues and restricted local governments' ability to raise future revenue, effectively eliminated local discretion over school funding and provided California with a de facto system of state financing for public schools.

Enactment of the Educational Reform Act of 1983, Senate Bill 813, signaled yet another stage in the escalation of pedagogical policy making to the state level. This massive reform effort, supported widely by state legislators and executive branch officers, business leaders, and a variety of public representatives, establishes mandates and provides incentives for approximately eighty kindergarten- through-12th-grade education reforms. It is among the broadest and most forceful efforts ever attempted by a state to enhance the productivity of its schools.

Following the development of a system of state financing and the state's efforts at achieving greater productivity, there is now developing an intense effort to measure the consequences of state dollars and state reform efforts. Policy makers are asking what is it we are getting for our money? Do all the new reforms make any difference? Are California schools enhancing their productivity? Will California's pupils enable the state to succeed in an ever more intense economic competition with other states and nations?

Until recently California has been ill equipped to answer questions such as the foregoing. The state's education data collection and analysis efforts were oriented primarily towards questions of school finance equity and the distribution and effectiveness of categorical aid programs. Even this data gathering, at least until recently, has not been systematic, coordinated across agencies and levels of education, or designed to inform local planning and evaluation. Moreover, little thought had been given in the past to using regularly collected

data for purposes of measuring school and school district productivity. Lastly, only modest thought had been given to the ways in which the state might actively reward the most productive schools and districts.

The policy papers included in this package address various facets of the above mentioned topics. Guy Benveniste of the University of California, Berkeley, School of Education explores the underlying issue of accountability and describes the implications of different types of accountability measures. In "New Directions for State Education Information Systems," Michael Kirst of Stanford University's School of Education argues for a state "Information Czar" who would coordinate and integrate the various "data streams" that are currently collected and disseminated in a fragmented fashion. An argument for identifying and rewarding merit schools, rather than merit teachers, is presented by Walter I. Garms, of the University of Rochester. Garms discusses methods of measuring merit and specific indicators of merit and argues that schools need freedom to manipulate resources to achieve desired results.

Gene Dawson of the School of Education at Berkeley describes how data are collected for the California Basic Educational Data System (CBEDS) and offers suggestions for improving their reliability. Edward Haertel of Stanford University discusses general problems of measuring the effects of reform, and analyzes three proposed indicators of quality: SAT test scores, course enrollments, and hours of homework or number of writing assignments completed. Finally, David Stern, of the University of California at Berkeley, further explores the merit school concept and discusses issues related specifically to California's new "quality indicators" program.

Taken together, these papers constitute a significant contribution to our understanding of the complex issues involved in measuring the performance of schools and should be a valuable source of guidance as policy on accountability measures is formed.

THE DESIGN OF SCHOOL ACCOUNTABILITY SYSTEMS

by

Guy Benveniste

Professor

School of Education

University of California, Berkeley

May 1984

CONTENT

Highlights

"Accountability" in education should focus on schools as the relevant performance unit and not on individual students, or teachers. It would also be desirable to encourage greater collaboration between preschools, elementary, junior and high schools.

Accountability works best when it relies on incentives to redirect action and performance. Unfortunately the teaching profession in the United States lacks a career structure and consequently there exist very few incentives that can be used to reorient teacher behavior. The design of new accountability systems has to be conceived jointly with the restructuring of the teaching profession. Accountability should be designed so as to increase the status of the professsion.

Accountability implies external control. External control can be exercised through inputs, processes or outputs. Instruction is a complex professional activity requiring considerable professional judgement and discretion. While some output and process controls are warranted in education, we need to emphasize input accountability. This means that the time has come to reexamine the training and licensing of teachers. Teachers should acquire a master's degree early in their carrer. They should take a state examination similar to the state bar examination given to lawyers. Such an examination should emphasize both theory and practice.

Accountability systems need to be designed around measures that provide direct information to teachers or to parents and administrators as to how to rectify their behavior. This means that accountability systems need put more emphasis on variables that are directly controlled by parents, teachers or administrators and less on performances or outcomes teachers or others cannot, or do not know how to affect. For example, we urge that more emphasis be given to measures of time spent teaching and learning certain subject matters. School by school data should be collected regarding parent and community support, participation of students in school cooperative activities and levels of order and consistency. Some emphasis might also be given to ultimate outcomes such as how many students are placed in college or how many find productive employment. Some of this data is not easily obtained. More importantly, some of this data would probably be manipulated if strong economic incentives were tied to them. Yet such data would be most useful to local

communities, school boards, parent organizations and other actors interested in the schools. State wide accountability systems need not be limited to "top down" accountability. Bottom up accountability is also important. The state may ask that selected data be collected for top down accountability and design incentives to encourage improvements. But, at the same time, the state may consciously refrain from including other data in such incentive schemes and rely instead on bottom up accountability.

Current practices of standardized testing of achievement in certain subject matters, while necessary for diagnostic purposes, should be sharply downgraded in future accountability schemes. The use of criterion referenced tests such as those used in the California Assessment Program may be expanded, but should not be directly linked to schoolwide rewards and sanctions.

Current practices of standardized achievement testing are unsuitable in schoolwide accountability for four main reasons: 1) when such testing is tied to incentives it leads to serious goal displacement. The tests become a goal but the tests are not linked to the curriculum; 2) such testing, when used in school wide scores, masks real issues such as student turnover; 3) standardized testing does not establish a minimum standard and does not tell us how to reward schools trying to deal with difficult students, and 4) standardized testing acts as a dis-incentive, because the tests are designed so that helf the population taking them are bound to do less well than average.

It would seem far more preferable to rely on better testing. This would require the much less frequent use of far broader examinations closely tied to the curriculum. These examinations would also set a minimum standard. The teaching profession should be heavily involved in designing these examinations. As a start, the state could adopt a single statewide graduation examination, to test minimum achievement at completion of high school. But the examination should measure more than minimum achievement. It should also measure higher accomplishments.

The results of such examinations and other measures could be used in a School Base Accountability System. Such a system should be designed to involve schools in self-improvement and in helping other schools improve. The overall measures could lead to a ranking of schools in five categories: 1) Below State Requirements, 2) Meets State Requirements, 3) Improving School (involved in a development program), 4) Research School (involved in cooperative research with an institution of higher education), and 5) Mentor School (provides assistance to other schools). Schools that do not meet state requirements should receive special guidance and assistance, and research and mentor schools should receive additional funding to foment more exchanges of

good and successful practice.

Such schemes should be tested during an initial five year period to see whether they seem to lead to actual improvements. In due time they might lead to important reforms in education.

Outline of Paper

Accountability can be part of the solution. That is, accountability can lead to school improvement. However, accountability can also be part of the problem. It can lead to gross distortions. Teaching is a complex task. Our best understanding of the learning process suggests that good teaching has to be tailored to the particular intellectual structure of each learner. This suggests the need for considerable teacher discretion. It also suggests that there is no easy way to measure effective teaching because there is no easy way to assess how different teaching strategies contribute to actual value added learning. It is difficult to know how instruction affects the extent the learner has been able to move from point A to point B. Yet we tend to believe we can assess learning achievement and therefore assess teaching effectiveness. Since we believe we know how to assess results we are prone to institute accountability schemes based on achievement testing. Some of these, in fact, do not work well, but we tend to use them more and more. One purpose of this paper is to explain how testing might be improved so that some of its deficiencies be minimized.

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 non-polemical view of the teacher's role, a starting point for discussing how to design accountability schemes. We then proceed to discuss accountability. How accountability can be used : 1) to inform (i.e.,provide feedback); 2) to re-orient action, and 3) 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 incentives 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 schemes should be parsimonious. They should enhance the quality of life among teachers and not require excessive paper work. Given these realities, we present a set of design considerations for controls that enhance the profession.

The Design of Accountability Systems

A Consensual View of the Teacher's Role

Education has many committed schools of thought and there does not exist a consensual view of good education. 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 school children they happen to encounter, we expect them to be wise, 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 non-teaching tasks. We

We expect teachers to teach. We do not really want them to do other tasks and we wish them to resist non-teaching 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 a continuous process beginning with pre school and extending 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, and 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 indescriminately 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 there is more and more reliance 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 already 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 administrator. 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 these tests are used very frequently and can assume greater importance than they deserve. If parents, pupils, teachers or their administrators come to believe that it is important to achieve high scores, the tests are no longer used as a diagnostic instrument; 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 is relevant to teach. More importantly, such testing can be manipulated and data falsified. Some teachers, 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 three, six and twelve. The program is being expanded to grade eight 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 of 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 school wide assessments take much less time to administer, are less intrusive on individual teacher performance and still permit school wide 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 if pupils have mastered a portion of the curriculum.

Accountability in Perspective

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

-Accountability serves to inform. For example, to transmit imformation 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 non-threatening, designed to help schools, teachers, pupils, and public better understand each other. In this instance adaptation or adjustment takes place

-Accountability serves to re-orient action. For example, to induce teachers or the 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, want to give additional resources, or set penalties to achieve compliance. We can design accountability systems 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 pitting schools and public. Thus accountability becomes part of the problem, it makes 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 obscurifications 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 behaviour 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 measure what is significant. If we invent an accountability measure and re-orient teacher behaviour in the wrong direction, we have bad

Good accountability is tied to positive rewards in preference to penalties. Teachers are human, and human beings respond better to positive rewards. In education, positive rewards are scarce 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: 1) the creation of an incentive structure within the teaching profession and 2) designing accountabilisy 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 non-teaching 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 informative 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 acountability 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 assumptions 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 lead to demoralization, bad

accountability reinforces bureaucratization by creating greater uncertainty. In an uncertain environment where it is unclear how teacher behaviour might improve the accountability score card, 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, less able to innovate, 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 non-teaching 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 score card.

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 there exists no incentive leverage. It gives the impression of attending to problems, but problems are not attended because they require real solutions. Bad accountability arises because it is often easier to appear to do something than actually solving 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 curriculum of school districts, since it is difficult to reach a concensus 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 they 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 less 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 this 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: 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, and 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 re-standardized, if the questions are re-designed so that the population will again distribute normally, the same differences will again re-appear.

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 which limits the 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 dropouts - is a very high percentage of total enrollment. Therefore, school score variation has 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 objective measures do not have unforeseen consequences such as displacing or distorting teacher behaviours 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 aproximate some concept of student learning. However, as we saw 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 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 thay 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 sometime 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 acccountability 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, 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 electric 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 only select 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 advantaged 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 if 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 where 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, curriculum and even text books 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 is 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 kind of 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 1)when we are convinced that we know what works, and 2)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 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 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, re-orient, and justify action. One can inform by providing facts and figures. To re-orient 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 and 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 not for control purposes. They use instead, examinations based in part on essay type questions and problems. These examinations are closely aligned to the curriculum. Their 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 examation 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 in the examinations. They can also set targets for improvements 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

over-relying on negative clues. Most non-educational organizations and institutions who use rewards and sanction, tend to use negative sanctions only for a small portion of the populations they control. They usually use negative sanctions for the lower ten or twenty percent of the target population, and use diffentiated 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 (T.J. Peters and R.H. Waterman Jr. 1982 pp.223-234). 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 judgement:

"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 firms] philosophy and this extensive training to guide them. [A leading executive] summed up the problem when he said, "Substituting rules for judgment start a self-defeating cycle since judgement can only be developed by using it".[op cit p. 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 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 decision-making 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 teacher's perception 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, and Benson 1984). These authors recommend 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, 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. Moreoever, 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:

"All testing, auditing, information gathering, and incentive distributions should be organized around schools rather than school districts or individual classrooms." (Benson et al 1972, p. 47)

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

We need also 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 and teachers' 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 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 on 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

is 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 nursing, allied health, and 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 teacher's 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 contol 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 responibility and greater trust 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, 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 management 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 evaluations 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 school-wide rewards or sanctions. If we want to control students then it should only be done at important stages in the student's career, and it should be done with examinations 1)that are linked to the curriculum, 2)that establish a minimum that defines pass and fail, and 3)that provide rankings based on accomplishments. Similarly, teachers should never be evaluated on the basis of the performance of their pupils 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 used 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 have 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 if 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 require coordination. This means placing responsibility in a single

central body charged with setting policy for school accountability.

Producing schoolwide accontability 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 who it is intended for.

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 provides 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 acountability sytems that organize schools in different catagories involving some schools in helping others, giving to some schools inlarged 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 bejond tying incentives to schoolwide scores.

Bottom up accountability is particularly important where 1)implementation depends on local participation and support, 2)problems are diverse and peculiar to local conditions, 3)measures need to be interpreted inlight 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 school wide 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 defining 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 it's performance.

4. Research School. A school involved in a research oriented development program designed to increase it's 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, 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 state wide 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 fail 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, merged and when all remedies fail, 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 state wide 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 schools districts seem to spend on homework. Local parent groups, school boards and others might make better decisions if such information was 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 1) teacher preparation and achievement, 2) teacher use of time, 3) student learning time in selected subject areas, 4) order and consistency, 5) parent and community support, and 6) selected student outcomes.

1) 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; percent of teachers in each subject areas; percent 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 are 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 1) to upgrade their training and qualifications and 2) the distribution of skills in each school. Specialized teachers in mentor schools would be able to, 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.

2) Measures of teacher time. The purpose of these measures would be to provide incentives for expanding effective teaching time and for the de-bureaucratization of the schools. Teachers would self report approximate time spent on teaching and non-teaching tasks. Efforts to reduce paper work 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 supervison.

These process controls could provide bottom up accountability with school by school comparisons. Some of them (i.e. reduction of paper work) might be used in statewide competitions. In time, norms for acceptable levels of non-teaching 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.

3) Measures of student learning time. These might be approximated by course enrollment data, turnover rates, and 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 may 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.

4) 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 measure 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).

5) 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.

6) 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 that 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 state-wide 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 exams 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 administrating and grading the examinations.

We would want to also use essay type questions and problems in this examination. The examination would be alligned 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.

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, drop out rates, rate of admission in various levels of post-secondary 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 to 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 in 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 porvide most of the needed resources. We describe a style of accountability that is highly flexible, leaves 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 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 proceed with caution.

Proposals of this nature will require considerable discussion before any implementation can take place. Statewide examinations of teachers and pupils are radical departure with 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.

Bibliography

Bacon, William, Public Accountability and the Schooling System, Harper and Row, New York, 1978.

Barro, Stephen M. "An Approach to Developing Accountability Measures for the Public Schools," Phi Delta Kappan, 52, 196-205, December, 1970.

Benson, Charles et al., Final Report to the Senate Select Committee on School District Finance Vol. I. Scramento, June 12, 1972.

Boyer, Ernest L. High School: A Report on Secondary Education in America. Harper and Row, New York, 1983.

Browdy, Harry S., "The Demand for Accountability: Can Society Exercise Control Over Education," Education and Urban Society, 9, 235-250, February, 1977.

California Assessment Program, Examiner's Manual 1983, California State Department of Education, 1983.

Davies, Otto A. et al. "An Education Management Information System (EMIS)" draft manuscript School of Urban and Public Affairs, Carnegie Mellon University, Feb. 1984.

Duncan, Merlin G., "An Assessment of Accountability: The State of the Art," Educational Technology, 11, 27-30, January, 1971.

Eisner, Elliot W., "On the Uses of Educational Connoisseurship and Criticism for Evaluating Classroom Life," Teacher College Record, 78, 345-358, February, 1977.

Gifford, Bernard R. The Good School of Education: Linking Knowledge, Teaching, and Learning. University of California, Berkeley. School of Education. March 1984.

Glaser, Robert, "Education and Thinking," American Psychologist 39, 2, 93-104, 1984.

Guthrie, James W., "Educational Accountability," Proceedings of the Academy of Political Science, 33, 24-32, 1979.

Guthrie, James W. "Enhancing the Education Profession in

California: An Unfinished Agenda," Testimony before the California Assembly Education Committee, October 13, 1983.

Honig, William "Education Reform: Next Steps" Sacramento Department of Education, Mimeo February 1984.

Kirst, Michael W. "Loss of Support for Public Secondary Schools: Some Courses and Solutions," Daedalus. Spring 1983.

Lennon, Roger T., "To Perform and To Account," Journal of Research and Development in Education, 5, 3-14, Fall, 1971.

Lessinger, Leon M., Every Kid a Winner: Accountability in Education, SRA, Inc., Palo Alto, 1970.

Lessinger, Leon M., "Engineering Accountability for Results in Public Educaton," Phi Delta Kappan, 52, 217-225, December, 1970.

Lessinger, Leon M., "A Historical Note On Accountability in Education," Journal of Research and Development in Education, 5, 15-18, Fall, 1971.

Lessinger, Leon M., and Ralph W. Tyler, Accountability in Education, Charles A. Jones Publishing Company, Worthington, Ohio, 1971.

Lortie, Dan C. Schoolteacher. University of Chicago Press, Chicago 1975.

Marland, Sidney P. Jr.,"Accountability in Education," Teachers College Record, 73, 339-345, February, 1972.

McDonald, Frederick J. and Garlie A. Forehand, "A Design for Accountability in Education," New York University Educational Quarterly, 4, 7-16, Winter, 1973.

National Commission on Excellence in Education. A Nation At Risk U.S. Department of Education, Washington, D. C. 1983.

Olmsted, Richard, "Review of Every Kid a Winner," Harvard Educational Review, 42, 425-429, August, 1972.

Ornstein, Allan C. and Harriet Talmage, "The Rhetoric and the Realities of Accountability," Today's Education, 62, 70-80, September-October, 1973.

Oscarson, Janice M., "Community Involvement in Accountability," Journal of Research and Developement in Education, 5, 79-86, Fall, 1971.

Peters, Thomas J. and Robert H. Waterman, Jr. In Search of Excellence: Lessons from America's Best Run Companies. Harper and Row Publishers, New York, 1982.

Pincus, John "Incentives for Innovation in the Public Schools" Santa Monica. The Rand Corporation, Mimeo January 1973.

Peterson, Paul E. "Did the Education Commissions Say Anything?" The Brookings Review 2, no. 2, 3-12, Winter 1983

Ralph, John H. and James Fennessey, "Science or Reform: Some Questions About the Effectiveness School Model" Phi Delta Kappa 689-694 June, 1983

Spencer, Bruce D. and David E. Wiley, "The Sense and the Nonsense of School Effectiveness," Journal of Policy Analysis and Management, 1, 43-52, Fall, 1981.

Stoddart, Trish, David J. Losk and Charles S. Benson. Some Reflections on the Honorable Profession of Teaching. University of California, Berkeley, March 1984.

Wynne, Edward A. "School Award Programs: Evaluation as a Component in Incentive Systems" Educational Evaluation and Policy

NEW DIRECTIONS FOR STATE EDUCATION INFORMATION SYSTEMS

Michael W. Kirst

Stanford University

June 1984

This is a PACE Project sponsored paper. PACE, Policy Analysis for California Education, is a joint undertaking located at the University of California, Berkeley and Stanford University. Its Directors are James W. Guthrie and Michael W. Kirst. PACE is funded by The William and Flora Hewlett Foundation. However, the analyses and recommendations contained in this paper are not necessarily endorsed either by The Hewlett Foundation or the PACE directors.

NEW DIRECTIONS FOR STATE EDUCATION DATA SYSTEMS

Michael W. Kirst*

PACE Project (1984)

Stanford University

Overview of the Issues

California's education data system has made substantial improvements in the past 15 years. For example, school finance data have been dramatically improved in quantity and quality, especially with regard to equalization. Data on categorical programs were refined during the 1970s, and are increasingly useful for monitoring local compliance with state regulations. Indeed, the type and quality of state data tend to reflect past and present state policy priorities. The new state agenda of 1982, however, requires new types of data and has highlighted some of the gaps and shortcomings of prior data systems.

California is badly lacking in the collection of systematic information regarding areas such as curricular matters, instructional practices, educational personnel, school climate, and expectations for student attainment. This paper will provide a start on such tasks by suggesting specific policy recommendations for: closing data gaps, enhancing state and local use (particularly "bottom-up" demand for data for local decisions), coordinating data streams, matching policy data to new Senate Bill 813 needs, and providing top level data leadership. (A subsequent paper will address the detailed design of a state data system that builds on these recommendations.)

The problems, however, are not merely data gaps. At no point has there been a state "data czar" in the State Education Department. Various "data streams" are collected in an uncoordinated fashion and fed back to districts on a fragmented, piecemeal basis. Categorical programs, teacher preparation, and pupil assessment all have independent data collection and analytical efforts that tend to divide education policy by artificial "program barriers."

*The writer would like to acknowledge the critical comments of Bill Burson and Vin Madden of the California State Department of Education. Comments were also made by James W. Guthrie and David Stern of University of California, Berkeley.

California has several major data streams that follow traditional boundaries. For Example, California Assessment Program (CAP) data are distributed to one group of local users and bilingual categorical compliance information to another. The entire system has a top-down orientation with insufficient attention given to bottom-up user demands or access. CAP is a notable exception. Otherwise, local users find it difficult (assuming that they desired it) to combine state data streams for a comprehensive review of school district performance. Similarly, Commission for Teacher Credentialing (CTC) data are managed separately from the State Department of Education, preventing easy access or coordinated surveys for policy makers. State Department financial apportionment data are useful for local financial modeling and projections.

The new education agenda, featuring higher "academic standards," is for the most part simply being grafted on to state data priorities of the 1970s--equalization and categorical program compliance. This presents a clear danger of overloading surveys such as CAP by forcing them to carry more baggage than that to which locals can respond or will acquiesce in providing. For example, CAP added 10 items in 1983 to probe "school climate." This increased state demand for data is resented by local districts, in large part because LEAs do not find much of the state data easy to use or relevant to their particular problems. Moreover, local districts do not want to hear "bad news" from the state.

A share of the difficulty, however, resides with the LEAs who often do not use their own data (like NEEDS assessments) for policy making either. A forthcoming Stanford study finds much local information is collected with no connection to local policy or administrative decision making processes and structures. The lack of local district use of state collected data is a mutual problem of low motivation and barriers at each policymaking level.

Lack of local use of state data poses several problems. For example, classroom teachers must complete California Basic Education Data Systems (CBEDS), but they have little motivation to prepare accurate reports. Principals and teachers lack understanding of and commitment to CBEDS.[1] The state has difficulty coercing reluctant local educators to provide accurate numbers. The State Department of Education merely notes that

---

1. Data in this paper on CBEDS is derived from Barbara Jean Sims' paper, "Who Makes Up CBEDS," prepared for a UC Berkeley class on management information systems, 1984. Ms. Sims conducted 28 state and local interviews of educators.

failure to submit complete CBEDS data will result in "incomplete data in federal and state reports." This may not be a great concern to local teachers and principals. State administrators reported in interviews that they could find CBEDS users only as close to the classroom as the district level.

Although the state collects school site performance data, accreditation is handled privately and independently by Western Association of Schools and Colleges (WASC), apparently in isolation of potentially useful state data. Commission on Teacher Credentialing teacher status data is not on computers nor is it merged with either SDE or WASC. We do not know, for instance, how many are teaching in subject fields for which the teacher does not have proper preparation. University of California (UC) and California State University (CSU) student performance data on placement tests and college grades are yet another separate but relevant information stream. Feedback to local schools from UC and CSU is haphazard, piecemeal, and substantially less than maximally useful. These data are seldom used as a basis for rethinking college prep curriculum.

The State Education Department has made considerable progress in merging categorical compliance and CAP data--the Consolidated Programs Description Data Base is a good example (see Appendix I). California Assessment Program researchers have been able to construct profiles that enable one to identify particularly outstanding secondary schools on several dimensions. This promising start on combination and increased policy relevance of data streams needs to be intensified.

Recommendation One

California education policy data should be supervised and coordinated by a single statewide office.

We hesitate to use the label "Information Czar" but this does express the administrative essence. Some single office should be responsible for coordinating and combining data streams across state agencies (SDE, CTC, EDD, US, CSU, etc.) The "Information Czar" might have a Board of Directors composed of the State Superintendent, UC, CSU, CC Chancellor, CTC, State Finance Director, Legislative Analyst, and the Legislative Audit Committee. This same board should play a lead role in filling data gaps (see our detailed recommendations below), eliminating low use items, and designing new policy data formats. This office need not include financial apportionment data, but should focus initially on a plan for SB 813, CBEDS, CAP, and the categorical programs. This office should combine data sources in order to provide a holistic view of schools rather than

individual program reports.

The Information Czar should have some specific and widely publicized criteria for approving collection of school performance measures. Some desirable characteristics of proosed measures are:

1. They must be reflective of desired outcomes that are important.
2. The attribute to be measured must be susceptible to accurate measurement on some quantifiable basis.
3. People must be convinced of the validity and reliability of the measure as an indicator of the desired outcome.
4. Attainment of merit, based on the measure, must theoretically be equally possible for every school, regardless of how well it is currently doing in the educational process, or of the kind of student body it serves.

Recommendation Two

Major gaps in state education policy data should be filled as soon as possible.

1) There are few state data on middle schools. The California Assessment Program's 8th grade test will help, but it is only part of an overall picture of middle schools. We know little, for example, about how middle school tracks and course choices determine academic course work taken in the senior high school.

2) While there are useful UC data, there is no data integration between community colleges and elementary/secondary schools. We do not know how students from specific high schools perform in community colleges or what their GPAs are. We do not know how they score on academic placement tests. Since the community colleges are designing a new data system now, integration with secondary school needs is particularly appropriate. This may require a substantial amount of money for the community colleges.

3) Most of the state school site performance data focuses on the four-year college bound. The combinaiton of UC, CSU, College Board, and CAP data provides a good base. The state is gathering much more specific data on academic course-taking patterns, but not on "life in the general or middle tracks." While categorical data provide insight on the lowest achievers, these data are

program compliance rather than policy oriented. The state does have achievement data on compensatory education pupils.

4) While SB 813 requires some new types of data, this collective effort should not be undertaken in isolation from existing data streams. Moreover, new SB 813 data needs can be combined with reductions in other data requests. Two good candidates for elimination are: a) some financial data that are not used much; b) much of the categorical-oriented data that is no longer used by Department compliance officers.

5) The California Assessment Program cannot be a train used to carry vitally needed data to assess the impact of SB 813. CAP is already freighted with many new items such as school climate. The biggest reason for such current overuse is that CAP is the only vehicle which reaches students. The state needs to consider a CBEDS student information sheet. Any appreciable increase in CAP will probably lead to increased local education agency (LEA) resistance and lower data reliability.

6) Local education agencies that report large numbers of Limited English Proficient (LEP) students must offer costly programs and seek waivers from the State Board of Education for single language teachers. Consequently, while the Hispanic pupil population is growing dramatically, there continue to be problems with the reliability and validity of data regarding pupils with limited English speaking ability (LEP). There are incentives for LEAs to underestimate LEP students and thereby also understate the need for teachers with dual language capability.

7) There are serious shortcomings in existing data on the new policy dimensions regarding length of the school day and year. New additions to CBEDS will help this area somewhat, but it is still weak. (Pupil mobility causes severe data problems that will be analyzed in a subsequent paper by David Stern of the University of California.)

Recomendation Three

State education data systems should emphasize more local user needs and adaptions.

The current system has several unique and effective procedures for local use, including CAP formats and local users' handbooks. More needs to be done, however, to assist LEAs in assessing their academic standards and local situations. Selected data at the local level should be on microcomputers and its accessibility by the state should be through telecommunication networking. Such a network would permit the

state to extract its fair share of data and allow local districts to access state developed software to analyze their own data locally.

One promising technique for enhancing local use is an annual report of school site performance that incorporates standard statewide data with locally devised categories. (We have provided to several SDE officials examples of this approach. See Appendix 3 for an overview of the Florida school site report). The Florida approach was conceived before the current stress on curriculum, academic content, and college performance. Consequently, it needs to be supplemented with numerous educational content and process variables to be useful in California. However, the general concept has merit and might provide a bottom-up performance data constituency. Florida reports are disseminated widely to parents, newspapers, and citizens. The Florida system of financial rewards for the best school site reports has stimulated substantial interest among LEAs. The state could assist local use of data by making it easier for locals through computer networks to tap into state data bases (e.g., a distributed data base).

A bottom-up performance data system will not be easy to implement. Much of the current needs assessment data is not used by LEAs for policy decisions. Locally devised tests are used for individual pupil assessment, but less often for policymaking. Local school districts seldom collect systematic information on dropouts or post-secondary student outcomes. The SB 813 agenda, however, moves beyond a categorical compliance orientation to decisions that LEAs must be committed to in terms of academic time, content, processes, and standards. This new focus requires local commitment to data use. This will require an aggressive SDE technical assistance and field training effort. In May, 1984, the SDE began meetings all over the state to talk about a local school profile that might incorporate the recommendations in this section. (PACE will be reporting later this year on potential uses of statewide performance data with WASC accreditation.) Mississippi is now using output measures based on the school effectiveness literature in their accredidation.

State data systems will never be "complete" or finished, because as new policy areas arise a transition will take place. The new "excellence" issues of 1983, however, require a major increment of change. Data can help make education systems undertake self-correcting course changes and prevent over-reaction to fads and media bandwagons. A recent study revealed that the federal government spends $9 million on education statistics, but $36 million on health, $137 million on labor, and $54 million on agriculture statistics. Given these financial differences, it is not surprising that the nation knows more about its livestock than its students. Obviously, state

governments cannot rely on the federal government for either conceptual leadership or as a supplier of comparative state performance data.

The redesign of California education data outlines above will cost money. California's public education, however, is a $13 billion a year undertaking and more information expense is warranted. In the future, local schools can computerize much performance data. However, use of data in education policy at the local level is problematic, especially comparative performance. State leadership should include utilization as well as generation of improved education data.

What Will Information from the Database Look Like?

Database information is available in many forms. Your data team representative can use the terminal and give you information verbally or in handwritten form. Data can be printed out in lists or tables to suit individual requests. There are also standard formats available for printing school and district profiles.

What Kind of Review Will a Report Have Had?

If a fully certified data team member accesses data and gives you a verbal or handwritten report, no additional checks are required. If you receive a computer printout that you intend to use internally, it will have a small circular certification stamp to indicate that it has been reviewed by data source manager(s) for any of the sources used to prepare the report. If you request a printout to be distributed outside the Department, the report will also have been checked by the database coordinator.

What Kind of Training Have Data Team Members Had?

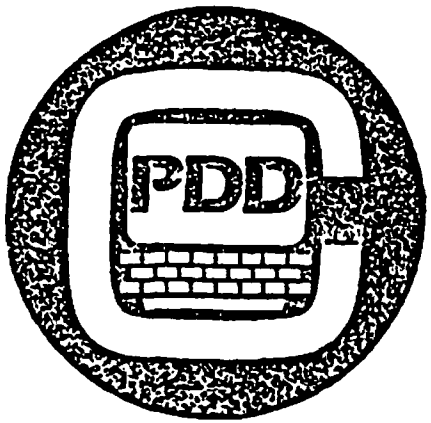
A six-month training program has been provided for data team members. It was two-fold: (1) A series of technical training sessions taught participants to access the database, retrieve data, and format reports. (2) A series of content training sessions covered all data sources, teaching what data are available and how to interpret and use the data. A person who has completed all the training and is fully certified to access the database will have spent at least 70 hours in formal sessions, follow-up exercises, and homework.

Data team meetings are held every other Thursday to update members. The meetings keep everyone informed about technical changes in the system, changes in the data available, and updates to the data. Members share experiences with using the database and plan for making it work effectively. Subcommittees working on specific projects report to the group and get further direction.

QUESTIONS AND ANSWERS

ABOUT HOW TO USE THE

CONSOLIDATED PROGRAMS DESCRIPTION DATABASE



APPENDIX I

What Is the Consolidated Programs Description Database?

The Consolidated Programs Description Database (CPDD) is a compilation of data that have been computerized. The data have been collected from schools and districts by various SDE offices and have been carefully selected to represent a comprehensive "picture" of each school and district. While there is a focus on data related to Consolidated Programs, the CPDD has data for all K-12 public schools and districts in California.

The original purpose in developing the database was to centralize data about students, schools, and LEAs and make it easily accessible to managers for planning, decision-making, and policy formation. The computer design allows flexibility to manipulate data: data can be "mixed and matched" according to management needs. The design also allows for training of SDE staff to access the system.

What Types of Data Are Included in the Database?

Data are available about:

- Funding (allocations and entitlements by program)
- Student participation
- Ethnicity
- Compliance status
- Languages
- CAP scores
- Program review ratings
- . . . and more

Where Are the Data From?

ALL CPDD data come from SDE forms or applications (e.g., PQRI, A-127D, CBEDS) or from other SDE computer systems (e.g., Compliance Tracking System, Bilingual Teacher Waiver System). There are currently 16 different sources from which we select and compile data. The data in the CPDD have been edited and certified by the responsible units.

Who Will Be Using the Database?

Authority for providing information from the CPDD does not come from the project itself, but from SDE management. It is each unit manager's responsibility to determine to whom the data team representative can provide information.

How Do I Use the Database?

First, you need to have a question that the database will help you answer. Secondly, you need to talk to your data team representative, your primary link to the database. This person has been trained to access the computer, to know and understand the data that are available--not only in the database, but elsewhere in the Department--and to help you figure out what data you want.

What Will Happen When I Submit an Information Request?

Your data team representative will need to sit down with you for a few minutes to discuss your request. The purpose of this is to clarify how the data will be used, identify criteria you want applied, and find out how you want the response formatted. Your representative can also help you figure out alternatives if exactly what you want is not available.

An information request log has been developed to keep track of CPDD use and to guide your data team representative in responding to your requests. Your data team representative will use this form to work with you in clarifying and refining your request before working at the terminal. This is a crucial part of the information request process.

Appendix E

The University of the State of New York
THE STATE EDUCATION DEPARTMENT
Albany, New York 12234

INFORMATION CENTER ON EDUCATION

The Information Center on Education was established in 1967, and is composed of two units--the Bureau of Statistical Services and the Bureau of Educational Data Systems. It is charged with the responsibility for the identification, implementation and operation of data systems in all areas of education as well as the coordination of all data collection procedures within the State Education Department. The Information Center:

1. develops, implements and operates educational data systems for the purpose of assembling information in all areas of the State's educational enterprise;

2. analyzes, interprets and disseminates data relating to public and nonpublic elementary and secondary schools as well as colleges and universities of the State;

3. advises and assists Department units and school authorities in procedures for the implementation and use of data systems;

4. coordinates all data collection procedures within the Department;

5. advises and assists Department units in the design and use of forms and survey instruments;

6. recommends policies and procedures for processing statistics;

7. plans, supervises and conducts special statistical surveys and studies;

8. advises and assists other Department units and school authorities on the conduct and design of experiments as well as the statistical analysis of experimental data;

9. prepares projections and estimates.

The following paragraphs provide a capsule review of these activities.

Note: ICE staff = 17 professional
14 support

Does not include data processing which is provided from a central DP unit within the Education Department.

It will be apparent that the developments listed were not, and could not be achieved by the Information Center alone. None would have been possible without the cooperation of local educational agencies and the many operating units of the Education Department who worked, and continue to work diligently with Information Center staff members to initiate and improve Department information systems. While the many Department units involved will not be enumerated here, special note must be made of the contributions of the Division of Electronic Data Processing which has shared equally with the Information Center the responsibility for and effort expended in bringing about the developments noted.

BASIC EDUCATIONAL DATA SYSTEM

In the fall of 1967, and after two years of development, a completely new and comprehensive information system on public elementary and secondary schools was implemented--the Basic Educational Data System (BEDS). Through this system, most of the information needed by the Education Department on public elementary and secondary schools is collected on a given day in the fall of each year and, by the use of machine-readable forms and automatic data processing, various outputs are produced which provide timely information for multiple purposes. They include:

I. Descriptive Reports--Reports are produced for the various general and subject supervisory units of the Department and are of two basic types:

- Comprehensive school reports are produced for the Bureau of Elementary and Secondary School Supervision. These reports (one for each public elementary and secondary school in the State) include: enrollment by grade, daily session data, instructional room inventory, distribution of last graduating class (if secondary school), number of dropouts, special programs or activities (including participation in regional programs, closed or open circuit television, programmed learning, prekindergarten program, flexible or modular scheduling and others), class size data, teacher load data, a complete faculty listing (with information about each individual, e.g. degree status, certification status, experience) and selected salary data.

- Subject faculty reports are produced for the subject supervisory bureaus. These include: enrollment by grade, daily session data, instructional room inventory and detailed information about faculty in a particular subject area including name, degree status, certification status, years of experience (4 categories), type of appointment, title of each course taught, number of pupils in each class, grade level, type of pupils (e.g. below average, average), number of periods the class meets a year.

- These reports are the major source of information the bureaus have about the schools and their programs.

II. Statistical Reports--A wide variety of statistical analyses of characteristics of public school professional staff are produced, including such factors as salaries, degree status, certification status, experience, sex, age and racial/ethnic characteristics. Summaries of various factors are available by school, school district, county, geographic region and for the total State.

In addition, several pupil statistics are also generated including enrollment by grade, racial/ethnic characteristics and course registration data as well as student staff ratios, class size and student load. These data are also available at various levels of aggregation.

Copies of these reports are regularly returned to school districts and are used by the districts and teacher groups in contract negotiations.

It should be noted at this point that the information contained in the personnel file of the Basic Educational Data System is covered by a strict policy of confidentiality.

The Basic Educational Data System has enabled more rapid summarization of a greater quantity of information about the public schools in New York State than was previously possible and, at the same time, has substantially reduced hand tabulation, duplication of effort and multiple requests for the same data in the Education Department. During the first year of the Department's consolidation of data collection activities over 20 annual reporting forms were eliminated from use. It is, of course, not possible to determine the number that have been eliminated (or prevented) since then due to the ability of the Information Center to respond to information requests which would normally have required a special survey or new form.

The number of requests for information going out to the schools is further reduced because the system is able to respond, at the State level, to a wide variety of recurring reports from outside agencies that otherwise would have been sent to local school districts. These agencies include: New York State United Teachers, the U.S. Department of Education, the National Education Association, the National Center for Education Statistics, the Bureau of Census and other professional associations.

III. Special Request Reports--Each year, a wide variety of information is supplied from the Basic Educational Data System to users both within and outside of the Education Department. In addition to data regularly distributed to Department units and that returned to school districts as standard output, the Information Center handles literally thousands of special requests each year. The requests vary from those which can be handled quickly, by phone or reference to a publication, to those of a very complex nature requiring special analysis and computer programming. The nature of these files are such that complex interrelationships among data elements can be obtained along with a longitudinal record of the progress of education in the State. Lacking such a broad based information system, the Department would be unable to answer these very legitimate requests for information. In addition, both the Department and the school districts would be called upon to file duplicative data requests which are now available quickly from a single source.

HIGHER EDUCATION DATA SYSTEM

The Information Center on Education also has responsibility for data collection and coordination in the area of higher education. As with elementary and Secondary schools, ICE has attempted to reduce the reporting burden placed on the 250 colleges and universities of the State. To that end, the Education Department has, since 1966, combined its information needs with those of the U.S. Department of Education. Basically, the Department uses the forms developed by ED--the Higher Education General Information System (HEGIS)--adding brief supplements where required. Information is received in the areas of enrollments and admissions, faculty, degrees awarded, institutional characteristics, student migration, libraries and finance. The data are entered into a computerized system and are available at various levels of aggregation.

The Information Center coordinates the data collection for the Federal government by mailing, receiving. editing and returning all HEGIS forms to Washington thus assuring the presence of identical data at both the State and Federal levels.

OCCUPATIONAL EDUCATION DATA SYSTEM

The Occupational Education Data System was implemented in the fall of 1970. The system provides enrollment and student follow-up data necessary for preparation of the Annual Report of Occupational Education and the New York State Plan for Occupational Education required by Federal regulations.

In 1978-79 some 400 private business and trade schools were added to the universe of over 800 secondary and postsecondary institutions providing data for this system.

NONPUBLIC SCHOOL REPORTING

The Information Center also collects data on nonpublic schools in a form compatible and interrelatable with that collected on public schools. The revised reporting procedure was used for the first time in the fall of 1969.

CONTINUING EDUCATION REPORT

This annual report was developed in conjunction with the Division of Continuing Education and includes data on enrollments, teachers and fees associated with continuing education programs operated by public school districts and BOCES.

HANDICAPPED CHILDREN INFORMATION SYSTEM

The Information Center has developed and implemented an automated system for processing the placement of over 9,000 handicapped children in approved nonpublic schools. Automation of an antiquated card filing system has reduced by 75 percent the time lag between school district requests for placement and Department action of approval or nonapproval. By 1980 the time lag will have been reduced by over 90 percent. Information from this file is used to keep track of institutional placements and to assure legitimate maintenance, tuition and transportation payments on behalf of these children.

The statistical portion of this system contains enrollment and service area data on all school age handicapped children in the State. Summary reports of handicapped children by type of handicap, age, and location of service are provided to units in the Department for use in planning and for Federal reporting.

EDUCATIONAL MANPOWER INFORMATION SYSTEM

Using existing Department Personnel and Certification files, thereby obviating the need for additional data collection, the Information Center has developed an elementary/secondary Educational Manpower Information System designed to serve as a base for the planning and development of professional education manpower training programs in the colleges and universities of New York State. Specifically, the System provides information in the following areas:

1. the existing educational manpower in public elementary and secondary schools;

2. the potential pool of educational manpower produced by teacher training institutions;

3. the potential pool of educational manpower not currently employed in a New York State school district;

4. the turnover rates of educational manpower in all sectors of the public education system.

PROGRAM COST SIMULATION MODEL

In a time when expenditures for education are growing and resources are diminishing, there is a growing concern with the issues of efficiency, control and accountability in the financing of public education. Of specific interest is information dealing with costs of various programs, e.g., science, mathematics, bilingual, handicapped.

While New York State school districts are required to submit an annual financial report to the Education Department, that report does not generally request expenditures associated with specific programs. In the absence of such information and to meet the increasing

demand for program cost data, the Information Center has developed an "Education Program Cost Simulation Model".

The model uses two Department files maintained by the Information Center, namely, the School Finance file and the Basic Educational Data System Personnel file. It is extremely flexible and can be used in a variety of ways to provide estimates of program costs in the public schools of New York State.

SURVEYS

As required, the Information Center designs and conducts surveys for various Department program offices. Such surveys, which can be conducted on either a sample or universe basis, are undertaken to provide in-depth information not available in the regular reporting stream.

DATA COORDINATION

The Information Center has been charged with the coordination of all data collection with the Department. As a part of that responsibility, the Information Center reviews and approves all data collection instruments to be sent by Education Department units (or those prepared by contractors to the Education Department) to any institution or person within the University of the State of New York. If the Information Center withholds clearance of any form, the decision may be appealed to the Commissioner.

DATA ANALYSIS AND DISSEMINATION

As a service unit, the Information Center on Education is called upon daily to analyze, interpret and disseminate information to offices and agencies both within and outside the State Education Department. The Information Center advises and assists other Department units in the design and conduct of experiments and in the statistical analysis of experimental data. It also has responsibility for the preparation of projections and estimates.

As a part of its dissemination activities, the Information Center produces a number of annual publications which are described on the attached list.

11. Directory of Nonpublic Schools and Administrators	11. Published in the early fall, this is the sole reference available showing names, addresses and telephone numbers of nonpublic school principals. It also shows the registration status of nonpublic high schools and is in exceptionally high demand.
12. Professional Personnel Report	12. A detailed presentation of the demographic characteristics of public school professional staff for the current school year, this publication is available in the spring of the year. It is used as both an historical time series and for research purposes.
13. College and University Opening Fall Enrollment	13. Published in the early winter of the current academic year, this publication presents opening fall enrollment data to the higher education community for planning purposes.
14. College and University Admissions and Enrollment	14. Published after the close of the academic year, this report presents a detailed analysis in summary and disaggregated (by college)form. It is used as both an historical time series and for comparative research purposes.
15. Colleges and University Degrees Conferred	15. Published after the close of the academic year, this report presents a detailed analysis in summary and disaggregated (by college and subject area) form. It is used as both an historical time series and for comparative research purposes.
16. College and University Employees	16. Published after the close of the academic year, this report presents data on the number and salary levels of college and university employees by type and level of position, type of institution and sex of the employee. It is used both as an historical time series and for comparative research purposes.

17. College and University Revenues and Expenditures	17. Published after the close of the academic year, this report presents information about the State's colleges and universities disaggregated by level and control. It is used as both an historical time series and for comparative research purposes.
18. College and University Racial/Ethnic Distribution of Enrollment	18. This biennial publication presents data in both summary and disaggregated (by control, level of student and subject area) form. It is used as both an historical time series and for comparative research purposes.
19. College and University Racial/Ethnic Distribution of Degrees Conferred	19. This biennial publication presents data by type of institution and degree level. It is used as both an historical time series and for comparative research purposes.
20. College and University Age Distribution of Students	20. This biennial publication presents data by type of institution and level of students. It is used as both an historical time series and for comparative research purposes.

Appendix 3

Annual Performance Reports

Whereas the statewide testing program provides the state with an early-warning system regarding minimum levels of student achievement, the Annual Performance Report is primarily for local client interests. This report would appear once each year, probably in early spring. The principal would be ultimately responsible for overseeing its production, but it should have sections reserved for exclusive use of the Parent Advisory Council, students (above the ninth grade), and staff. The report would be published in local newspapers, posted prominently in the school, and, most importantly, sent home to the parents or guardian of each student. The report would be the primary printed instrument by which clients could assess the effectiveness of their local school.

Proliferation of reporting forms and data collection efforts has long been a frustrating fact of life in both the private and public sectors. If Annual Performance Reports are well designed, they should help to reduce some of the burden. For the state, federal government, and local school district as well as for the individual school site, the Annual Performance Report should be the primary data compilation instrument. The school district could aggregate the information it needed from school reports, and then pass them forward to the state. Rather than imposing an additional informational burden on local school personnel, the Annual Performance Report might well consolidate all other such efforts.

The contents of an Annual Performance Report should include topical categories and items such as those illustrated below:

School Information

Name, location, enrollment, age of building, number of classrooms, number of specialized rooms, school site size, state of repair, amount spent on maintenance in the last year and last decade, library volumes, etc.

Staff Information

Number of staff by category, proportion in various license classifications, age, sex, ethnic background, experience, degree levels, etc.

Student Performance Information

Intellectual performance: all results of student performance in standardized tests should be reported in terms of state-established minimum standards. Relative performance of different schools in the district should also be provided. Other performance information might also be included: student turnover rate; absenteeism; library circulation; performance of past students at next level of schooling (junior high, high school, college); etc.

Areas of Strength

Here the school can describe what it considers its unique or noteworthy characteristics. The purpose is to encourage every school to have one or more areas of particular specialization and competence, or to espouse a particular educational philosophy, or to employ a distinct methodology or approach. This section would inform parents about the tone or style of the school.

Areas for Improvement

This section would identify five areas in which a school needed improvement and would outline its plans regarding them. These problem areas might in some schools change over the years, but in others remain the same as the schools mounted a long-term improvement project. This section should encourage schools to be self-critical, to establish specific goals and to report on subsequent progress.

Parent, Teacher and Student Assessment of School Performance

Responsible parents, teachers and students should be permitted an uncensored opportunity to assess school performance. This section would permit various school constituencies to express their opinions of school success or failure with respect to such matters as instruction, curriculum development, racial relations, student participation in decision making, drug abuse, etc.[16]

School Site Budgeting

In order to provide school sites with the flexibility they need to match programs with client tastes, they must be given budgetary discretion. This is best accomplished through a system of lump-sum allocations to individual schools based upon formal, districtwide rules. Presumably these rules would allocate an equal amount of money for every similarly-situated student in the

MERIT SCHOOLS FOR FLORIDA
A Concept Paper

Walter I. Garms
University of Rochester
April, 1984

THE CONCEPT OF MERIT SCHOOLS

This paper represents an attempt to develop the concept of a merit school, and to define operationally how that concept might operate in Florida.[1] The concept rests squarely on the notion that the best way to understand what happens in the education of children is not by viewing the individual classroom and the interaction between children and a particular teacher, but by viewing the effects of the school as a whole on children. Education is a particularly cooperative enterprise, where what is done with or to a child by other professionals reacts strongly in a positive or negative fashion on the effectiveness of a particular teacher's instruction. The intent of this paper is to examine how incentives at the school level can be used to improve the effectiveness of education in Florida.

The paper first notes some of the problems with merit pay for individual teachers. It then looks at the ways in which a merit schools system would answer these problems. Next, it develops criteria for a merit schools proposal. It discusses three alternative means for measuring merit, and recommends one method. Next, a series of measures is suggested, with the details of each (and the problems with it) discussed. The way in which the scores on individual measures are to be summed to provide a single measure of a merit is described. There is a discussion of how the merit awards should be distributed among schools and within schools. Finally, an important concomitant of a merit schools plan is discussed: freedom for the school to manipulate imputs to achieve desired outcomes.

WHAT'S WRONG WITH MERIT PAY FOR TEACHERS?

In recent years, there has been a good deal of discussion, and a fair amount of legislative activity in various states, regarding merit pay for teachers. The rationale is simple:

1. The schools are not doing as good a job as people would like them to do.

[1] The author gratefully acknowleges the contributions of the following individuals who have contributed ideas and suggestions: Robert Conrad Administrative Intern, Greece Central Schools, Clark J. Godshall, Teacher, Hilton Central Schools, Edward J. Maguire, Principal, North Rose Wolcott Schools, Joseph Occhino, Principal, Hannibal Central Schools, Seppo Pollanen, Special Education Teacher, BOCES II, Monroe County, and Thomas J. Strining, Principal, Webster Central Schools, (all from New York State).

2. The quality of instruction depends critically upon the quality of the teacher.

3. Teachers are not currently paid in a way that is closely related to the quality of instruction they provide.

4. If pay is based on the quality of the teacher, good teachers will stay in the profession (and be attracted to it), and poor teachers will leave the profession. In addition, teachers will act to improve the quality of their instructional efforts.

5. It is possible to design systems of merit pay based upon objective, measurable, criteria that are closely related to teacher quality.

6. The spirit of competition engendered by rewarding teachers on the basis of merit (or performance) will improve instruction).

If we examine this rationale, we can see that the first point is not arguable. Few people would disagree with the second point, although many would mention other factors that are also important. The third point, regarding the current methods of paying teachers, is also not usually in dispute. Almost uniformly in school districts across the United States, teachers are paid according to the number of graduate credit hours they have obtained and their number of years of service. Research does not indicate that there is a close correlation between graduate credit hours and teaching quality, and to the extent that there is a correlation between years of service and quality, it may actually be an inverse one.

The remaining points deserve closer scrutiny. Is it true that basing pay on "merit" (used henceforth as synonymous with teaching quality) will attract and hold good teachers, and encourage poor teachers to leave the profession? It certainly would provide some financial incentive. Whether that financial incentive is enough to achieve the goal depends upon the influence of other factors upon the teacher. Many teachers who teach in ghetto schools will find that additional money is not so important as better working conditions, free of the fear of physical violence and the necessity of spending large amounts of time on discipline. There may well be other factors that weigh heavily in a teacher's decision to stay in or leave the profession.

Will merit pay encourage teachers to do a better job? As with the question of entering or leaving the profession, it depends. In this case, it depends not just upon the strength of other factors. It also depends upon the extent to which the teacher perceives that it is possible to accomplish the things required in order to receive merit pay.

Point five is a difficult one. Teacher unions stoutly argue that the present state of the art is such that valid, objective, measurable criteria cannot be found. There is much to be said for this point of norm-referenced tests, designed to compare a student's performance with the average performance of others of his age or grade level. For purposes of measuring performance of teachers, criterion referenced tests, which measure the extent to which specific material has been learned, are preferable. However, even assuming that good criterion referenced tests are available, they cannot measure all of the things that a student should learn. In addition, good tests are hard to come by in the more subjective areas of the curriculum. Finally, there is little agreement on the relative weights to be put on the learning of reading skills, math skills, and skills in getting along with others.

There are other problems with the evaluation of teachers. A very important one is that teachers can reasonably argue that it is difficult to disentangle the influence of a particular teacher from the influences of other teachers with whom the student comes in contact, with influences of the student's peers, and of his home and neighborhood environment. How can a teacher be held accountable for the performance of a student when so many other influences are important?

Another objection has to do with the fact that students vary greatly in their abilities, interests, and learning styles. The teacher who teaches a class that is relatively homogeneous in these attributes, and whose teaching style is compatible with the student's learning style, will find it much easier to be successful than the teacher of a heterogeneous group.

Another problem with merit pay for individual teachers has to do with the fact that such pay puts teachers into competition with each other. Competition is not inevitably harmful; it is the basis of our free enterprise society. But competition within an enterprise which depends upon cooperation to achieve its goals can be destructive. It is easily possible to conceive of some teachers spending time doing things that will reflect discredit on other teachers, rather than improving their own performance.

Connected with this point is the fact that it is difficult to include principals and vice principals in a merit pay scheme, yet they may be extremely important to the success of the whole educational enterprise. There are similar difficulties in rewarding supervisors, teacher aides, and others who may be very influential in determining the amount a student learns.

WHY MERIT SCHOOLS?

The desire to improve the performance of the schools through offering incentives, together with the shortcomings of merit pay for individual teachers discussed above, leads to this proposal to consider instead the school as the unit for measuring merit. The proposal overcomes several of the most important drawbacks of

individual merit pay:

1. The individual teacher's contributions to the student's education cannot easily be disentangled from those of other teachers, and from other influences on the student. The individual school, on the other hand, constitutes a much larger part of the total educational influence on the student. Measurement of the achievement of students (or other appropriate measures) measures the total effect of the school experience upon the student. It is not necessary to try to disentangle the effect of one teacher from the effects of other teachers, of principal, supervisors, teacher aides, and the general school environment.

2. Merit pay for individual teachers fosters destructive competition. Merit pay for schools, on the other hand, enhances that spirit of cooperation upon which the educational process depends. Rather than trying to figure ways to achieve personal goals (while minimizing the achievement of other teachers with whom the teacher is in competition), in a school will find it to their advantage to cooperate in designing ways to improve the education of all of the students of the school. The skilled teacher will find it desirable to help less skilled teachers in the school, rather than finding such help counterproductive.

3. Additional measures can be used when one is measuring the performance of a school that are not appropriate when measuring an individual teacher. A simple example is a reduction in absenteeism, or in dropout rate, over which the individual teacher can be presumed to have little influence.

4. Parents could be expected to have a destructive effect on a merit pay system for individual teachers, for they will insist that their children be taught by a merit teacher rather than one who does not qualify. This would not be the case with a merit school, for with students geographically assigned to a school there is no alternative available to the parent. Parents could instead be expected to try to find ways to help the school achieve a merit rating (such as by helping to cut down on absenteeism).

5. Under a system of merit pay for individuals teachers, there is no incentive to manipulate the inputs of the school to improve educational achievement. A proposal to use some school money to improve instruction at the sixth grade level would be resisted by fifth grade teachers, for it would do nothing to help them achieve merit. Under a merit schools system, however, one could expect such a proposal to be approved.

It should be made clear at this point that there is a fundamental difference between the concept of a merit teacher and that of a merit school as described in this paper. The merit teacher is supposed to be one of the top teachers, as measured by some criteria that are hopefully valid and not dependent upon the students in the teacher's classes. If an award were made to

schools on the same basis of excellence, most of the awards would go to the suburban schools. What is being proposed here, however, is an award that would go to the schools that are meritorious because they are most improved over their previous condition. This would make it possible for all schools to compete on a basis of equality. Because there has been a good deal of misunderstanding about this (and concern about awards going to suburban schools), I suggest that a different name might be appropriate. One possibility would be to call it an Improvement Incentive Program.

CRITERIA FOR A MERIT SCHOOLS PROPOSAL

There are a number of important criteria to be satisfied if a good merit school proposal is to devised. These criteria will be mentioned briefly here, and then discussed at some length below.

1. The designation of a merit school must be based on measures. These measures must have certain characteristics:

a. They must be reflective of desired outcomes. There is no point in measuring something that is irrelevant. Measures of imputs are generally inappropriate; we are interested in outcomes.

b. The outcomes must not only be desired, they must be important. It is desirable to teach students to be neat, but that is not as important an outcome as many others that could be devised.

c. The attribute to be measured must be susceptible of accurate measurement on some quantifiable basis.

d. People must be convinced of the validity and reliability of the measure as an indicator of the desired outcome.

e. Attainment of merit, based on the measure, must theoretically be equally possible for every school, regardless of how well it is currently doing in the educational process, or of the kind of student body it serves.

f. Performance meriting an award in one year should not diminish the possibility of receiving one in the following year.

2. The awards for merit must be designed to provide an appropriate and adequate incentive for improvement on the designated measures.

3. Principals and teachers should have the maximum freedom to manipulate inputs in order to achieve the desired outputs.

4. The system should be designed so that it is not necessary to use identical measures for all schools.

AN APPROPRIATE MEASUREMENT SYSTEM

There is more than one possible approach to the measurement of merit in a school. Three different approaches will be explained here, and one will be chosen as the most preferable.

The first is an approach first suggested by Garms and Smith in 1970.[2] It is based on the assumption that the amount that children learn is a result of the combined effects of the efforts of the school, and of the out-of-school environment of the child. To look at only the former ignores much of what we have learned about home, neighborhood, and peer influences on learning. The method attempts to separate the effects of school influences from other influences through a statistical approach known as multiple regression. There is a single criterion that is measured: performance on a standardized test. The average achievement on this test for a school is predicted through knowledge of the socioeconomic status of the students in it. This is done by measuring a variety of factors for the children in the school. The data that were used in the 1970 study included the student's ethnic status, how many of his parents live with him, whether the family receives AFDC payments, the number of years of schooling of each of the child's parents, the number of rooms in the student's dwelling, and the number of schools the student has attended in the last three years.

Once the average expected achievement on the test for a school had been predicted based on the average socioeconomic status of the students, this prediction was compared with the actual achievement in the school. Better actual achievement than predicted would be the basis for some reward in mandatory help from a higher level.

The method had an important point in its favor. Excellent performance by the school in one year would not establish a new and higher base for performance in the following year. Instead, performance is always compared with predicted performance based on socioeconomic measures.

However, the method has several important flaws. A most important one is that prediction of achievement based on socioeconomic measures implies that there are systematic differences among socioeconomic groups in ability to learn. While much of the research that has been done supports the notion that present instructional methods produce differences in

---

[2] Walter I. Garms and Mark C. Smith, "Educational Need and Its Application to State School Finance," Journal of Human Resources V:3 (Summer, 1970), pp. 304-17

learning among socioeconomic groups, it does not necessarily follow that there are differences in innate ability among groups, and placing such an assumption into a system of state school finance is therefore inappropriate.

A second problem is the fact that it requires gathering data about students that constitutes, to many people, an invasion of privacy. While the research was done with no backlash from parents, it could be anticipated that a state system based upon it would result in much criticism.

A third problem is that there is only one criterion for excellence: performance on a single standardized test. Finally, the use of a sophisticated statistical technique is a problem, for it would be little understood, and therefore mistrusted, by most laymen.

A second possibility is the use of the "data envelopment analysis" method.[3] The method depends upon the use of linear programming methods, and according to the authors, "allows for production functions which may differ for each school, with multiple outputs and multiple inputs that may be related to each other in numerous ways (linear or nonlinear) that need not be specified." The method gives an overall evaluation of the "efficiency" of the school relative to the most efficient school, and indicates also how inputs could be changed in order to improve the efficiency of the school.

This method is an improvement over the simple multiple regression approach described above. It allows multiple outputs, whereas the regression method only allows one output. It also concentrates (although not exclusively on inputs that can be manipulated by the decision maker (the principal, for example). Its biggest drawback is that it is an even more complex mathematical procedure than multiple regression. Use of such techniques for management purposes, where the managers can be taught the method (and also know its shortcomings) is legitimate. Use of such a teachnique for rating schools, where the results are to be widely disseminated, could constitute a real problem. There are two simultaneous dangers. One is that some people would ascribe to the method a validity that is not warranted. The other is that people would refuse to believe in the validity of the method.

Another problem with the Data Envelopment Analysis method (as with the multiple regression method) is that the same variables must be used for all schools. This eliminates the possibility that schools could be allowed to choose some outputs

---

3. For an explanation, see Bessent, A., et al., "Productivity in the Houston Independent School District," Management Science 28:12 (December, 1982), pp.1355-67, and Authella Bessent, et al., "Evaluation of Educational Program Proposals by means of DEA," Educational Administration Quarterly 19:2 (Spring, 1983).

to be evaluated on, along with those standard ones on which all schools are evaluated. Yet another problem is the fact that the method relies on an analysis of stipulated inputs. While we know about some of the inputs that are important in education, we certainly do not know about all of them. And some of the ones we do know about are not susceptible to quantitative measurement. It is preferable for each school to be able to manipulate its inputs as it sees fit, with the goal of maximizing the outputs. The Data Envelopment Analysis method shows promise as a management tool, and deserves further evaluation for that purpose, but its complexity, and the use of fixed outputs and inputs, make it inappropritate for use in a publc merit schools, program, at least at present.

The third alternative is the prefered. It is a much simpler one conceptually, and allows for more flexibility. The idea is simply to choose some output measures that meet the criteria given earlier in this paper, and that are generally recognized as valid by professionals and laymen. Some of these might be mandatory, while others could be optional. A school would judge its performance using a number of these measures, with a weight assigned to each. The weighted sum of the measures would determine whether the school received pay as a merit school.

Let us take a single measure and see if it can be defined in such a way as to be a useful measure, for example, performance on a standardized test of reading. We will assume that this is a test at the fifth grade level, and that we are going to measure the average performance of all of the fifth graders in the school. The criteria, as noted earlier in this paper, are the following:

1. It must be reflective of desired outcomes. Measures of inputs are inappropriate.

2. The outcomes must not only be desired, they must be important.

3. The attribute to be measured must be susceptible of accurate measurement on some quantifiable basis.

4. People must be convinced of the validity and reliability of the measure as an indicator of the desired outcome.

5. Attainment of merit, based on the measure, must theoretically be equally possible for every school, regardless of how well it is currently doing in the educational process, or of the kind of student body it serves.

6. Performance meriting an award in one year should not diminish the possibility of receiving one in the following year.

The simplest possibility would be to choose the average score for the students in the school taking the exam. This would satisfy the first four criteria and the sixth, but would not satisfy the fifth. Those schools, typically suburban, where achievement is already high, would receive all of the awards.

A second possibility would be to use percentage gain on the test. A school in which the average score is 40 on a scale of 0 to 100 would have a ten percent increase if its average achievement increased to 44; the school with an average score of 80 would have to increase its score to 88 in order to have a 10% increase. This measure would make it possible for low-scoring schools to achieve a merit rating. In fact, it is possible that it would be easier for a low-scoring school to accomplish this than the high-scoring school. A serious problem comes when the school's accomplishment is unusually high. Suppose that its average score on the test this year is 91 out of 100. It is impossible for this school to achieve a 10% increase in its score.

This same notion of an upper bound on the possible score on the measure also makes it difficult to meet the sixth criterion listed above. If the school has done well the previous year, it may find that it is impossible to do as well again in the current year. A good deal of the problem here has to do with the assumption on the test that because there is a maximum score possible on the test that this represents the maximum that is possible for a student at that grade level to achieve. Part of this has to do with making the test simple enough that some students get all of the questions right. Another part has to do with the use of norm referenced scores. A score is given that is not a count of the questions correct, but an abstraction from that. A score of 95 means that 95% of pupils tested when the test was normed scored lower than that. Because of the normal bell-shaped curve of scores on any test, it may take only an additional two correct items on the test to move from the 50th percentile to the 51st, but it may take an additional 20 correct to move from the 95th percentile to the 96th.

We can get away from the last problem by using raw scores -- counts of items correct -- instead of the norm referenced scores. That does not eliminate the former problem that there is an upper limit to the number of items. It would appear to be necessary to use tests that would have enough items (becoming more difficult with each item) that almost no student could complete the entire test completely.

Another approach would be to stipulate that the measure is only to apply to the bottom quartile of students in the school. This would make it highly unlikely that the students measured would be near the top of the test, and would thus eliminate problems with both criteria 5 and 6. However, it does so by concentrating on only a portion of the students.

Yet another approach would be to specify that attainment of some high level on the measure, even if it does not represent an increase from the previous year, is evidence of merit. The trouble with this approach is that it assumes that there is an upper limit to what students can learn, and this may not be true. Even if it is true, it is certain that we are not near it at present. However, this sort of approach would be appropriate where there is a clear upper limit on what can be obtained. Dropout rate is a good example. It is obviously not possible to improve on the dropout rate if there are already no dropouts.

Another problem with the use of a test as a measure has to do with the question of who should take the test. Should we use the results of tests on students who are enrolled in the school this year but were not last year? If the test is given at the beginning of the year, this would mean that some of the students taking the test are ones with whom the school has as yet done nothing. It would appear inappropriate to do this. Some controls should be built in to see to it that students who contribute to the measure are ones the school has had a chance to work with.

Similarly, should those students who are in special education be required to take the same test that other students take? If not, is there an added incentive to classify students into such programs in order to raise the average achievement of the rest of the students? I believe that students mainstreamed into regular classes should take the tests. Students who are in special classes should either have special tests, or should not be included in the merit schools scheme until more experience is gained with it.

Also a problem is that if the test is given at the sixth grade level, the students tested this year are not the same students tested last year. Last year's fifth grade students will take the test this year, and they may be poorer or better on the average than last year's sixth graders. At the expense of a doubling of the amount of testing one could use gain scores for the same students, tested at the beginning of the year and at the end.

A LIST OF POSSIBLE MEASURES

It should be clear from what has been said that the definition of an appropriate measure is a very difficult one. Such definition is one that could occupy the time of the Florida Quality Instruction Incentives Council for quite a while. That seems an appropriate thing for it to do. However, giving that job to the Council to complete before the legislation goes into effect would delay things for at least a year, and probably longer. For this reason, I venture below to define some possible measures, and to comment on each. It is possible that some of them could appropriately be written into law this year, or be supplemented by others in the future as they were developed by the Council. I indicate for each measure the level or levels for

which the measure is most suitable.

1. (Elementary) Average percent gain in number of items done correctly on a standardized test of reading at the fifth grade level. The test is to be given at the beginning of the school year (pretest), and again at the end of it (posttest). The gain score is to be computed only for those students who took both tests. All students enrolled in regular fifth grade classes in the school are to take the test (including mainstreamed students). A student who is enrolled in the school during the period when the test is given must take the test, making it up if he is absent when it is administered. The test must be sufficiently difficult that no school shows an average score of greater than 80 percent of the items correct on the pretest (this eliminates the present state assessment tests).

This measure seems to meet all of the criteria. In particular, note that the use of gain scores for the same students eliminates the problem that performance meriting an award in one year might diminish the chances of receiving it in the following year. A school that, year after year, improved the attainment of its fifth grade students more than most schools, would merit the award every year.

This same sort of measure, based on standardized tests, could be used at other grade levels and with other subjects. This is, then, just one representative of a whole group of measures. It is strongly recommended that a test of cognitive skills development (i.e., critical thinking or problem solving ability) be included. This will decrease the tendency, in some other areas, to teach by means of rote memory. It is much more difficult to teach to the test when the test is one of problem solving. At the seventh grade level and above, there should also be tests in science and in social studies.

One thing to be considered with any measure is the incentive the measure itself provides for maximizing it illegitimately. For example, if the test is only given in reading and arithmetic, there will be a natural tendency to neglect other studies in order to concentrate on these two. This is not necessarily bad, if the studies covered by the tests are those that are the most important. One of the reasons students do not do better in school is that they do not devote sufficient time to the most important tasks. There are a number of subjects that must be (or are) taught in the schools that are clearly of less importance than reading, writing, and arithmetic. Some selective cutting back on the amount of time devoted to these less important studies would probably be good.

However, another incentive would be for the school to put its best teachers in the fifth grade, and its poorest teachers in the fourth grade, and possibly even to limit the amount taught in the fourth grade. This would provide the fifth grade teachers with students artificially behind where they should be, and therefore easy to achieve large gain scores. I call this a

perverse incentive. One fix for this would be testing at each grade level, but this implies an enormous increase in the amount of time, effort, and money devoted to testing. It might be possible to reduce the amount of testing by testing a random sample of students, if the number in a school is large enough. (Note, incidentally, that I have attempted to take care of another perverse incentive -- encouraging students who have done poorly during the year to be absent when the test is administered -- by insisting that students enrolled at the time the test is administered must take the test, even on a makeup basis.)

2. (All levels) The same as above (percent gain on a standardized test), except that the measure concentrates only on the lowest quartile of students taking the pretest. Here, the perverse incentive is to concentrate exclusively on the lowest quartile of students. However, if this is only one of a number of measures, the school will realize that an exclusive emphasis on the lower quartile of students will not be enough to gain a merit rating for the school. This illustrates a good reason for having a large number of measures: it counteracts the tendency to concentrate on only one thing to the exclusion of other important things.

3. (All levels) Percent reduction in failure rate on the state assessment tests. Like the previous measure, this concentrates on the poorest students. However, the measure is of reduction, rather than increase. As such, it has particular problems that deserve discussion.

Interestingly, a school with a small failure rate would have to decrease the absolute number of failures less to get a given percent improvement than a school with a large failure rate. A school that had a 30 percent failure rate the previous year and now has a 20 percent rate has made a 33% improvement. A school that had a 3% failure rate and has now changed that to a 2% rate has also made a 33% improvement. However, as the failure rate approaches zero, the task of decreasing the rate increases. In any case, there is an absolute limit beyond, which the failure rate cannot be decreased (namely, zero). The practical limit may be a number greater than zero, for it is impossible for the school to have complete control over failures.

It may therefore be desirable to define some small failure rate as reflecting excellent achievement even though it is no better than the previous year. The problem with this is that it does not assign a number to the percent improvement in failure rate. That violates criterion 3, which states that the attribute must be susceptible of accurate measurement on some quantifiable basis. Perhaps it would be best to assign the school a percent reduction that corresponds to the 75th percentile of percent reductions statewide.

Note that one could express this measure differently, as a percentage improvement in <u>passing rate</u>, rather than in failure rate. The main difference is in the difficultly encountered as

100% passing is approached. The school with a current passing rate of 60% only has to increase this to 69% to have a 15% increase; the school with a current passing rate of 90% cannot get a 15% increase. It seems desirable to have measures that take this difficulty into account, and this is why the recommendation is for an improvement in the failure rate rather than in the passing rate.

4. (High school) Percent increase in score of high school seniors on either or both subtests of the S.A.T. or A.C.T. Just as the previous measures concentrate on the lowest students, this one concentrates on the most able. One would have to define which students must take the test. It might be all those in the college preparatory track, if that is sufficiently well defined.

5. (High school) Percent increase in students winning specified statewide awards and scholarships. Note that, for the school that currently has no students winning such awards, winning the first one automatically puts them at the top in percent increase (the pecent increase would be infinite). I see no problem with this.

6. (High school) Percent increase in percent of students passing (not just enrolled in) courses that are evaluated with Advanced Placement tests or CLEP testing.

7. (High school) Percent increase in percent of students earning the Florida Academic Scholar Certificates.

8. (High school). Percent increase in percent of last year's seniors enrolled in an institution of higher education. This is a measure, like retention rate, in which there is an upper bound, making large percentage improvements particularly difficult for those schools that enroll large percentages of their students in higher educaton. However, reversing the measure (as was done by using dropout rate instead of retention rate) does not seem appropriate here. If this is only one of a large number of measures, most of which do not suffer from this problem, it is probably all right.

9. (High school) Percent increase in percent of last year's seniors enrolled in a university-level institution. This is, of course, a subset of the preceding measure, concentrating on those who are even more academically able. All six of these measures (beginning with S.A.T. scores) concentrate on the most academically able students.

10. (High school) Percent decrease in percent unemployed among last year's seniors. This measure is used instead of a measure of employment because of problems in defining employment (should it include college students? Military enlistments? Married females with children?).

11. (All levels) Percent increase in percent of ESL students who demonstrate competence on a test of English ability.

12. (Elementary, Middle) Percent increase in tests of physical growth and development. I am not familiar with the tests that may be available in this area, but if there are tests that are recognized as valid, it seems like a reasonable area to test.

13. (All levels) Percent increase in student participation in public performances or exhibitions in the arts areas, such as music and art. This is an area that is much less subject to accurate measurement, but if it is only one of a large number of measures, it can be included. Its inclusion indicates the desire to make the schools well-rounded.

14. (Elmentary) Percent reduction in percent of students who must repeat a grade. In a sense, this is a process variable, but it is of major concern to parents whose children are held back, and it has the valuable attibute of working against the perverse incentive in the standardized test measure to hold back students who are not apt to do well in the next year's standardized tests.

15. (High school) Percent reduction in dropout rate. Dropouts are defined as students who started as freshmen four years ago who are not currently enrolled in the school, have not graduated, and have not transferred to another school. To determine the dropout rate, dropouts are expressed as a percent of the size of the freshman class, excluding those who transferred to other schools. For example, suppose the freshman class consisted of 115 students. The senior class now consists of 80 students. Of the 35 students who are no longer here, 15 transferred to another school, and 3 already graduated. The remainder are unaccounted for, and presumably have dropped out. The number of dropouts is then 17 (115 - 80 - 15 - 3 = 17). The size of the freshman class, for purposes of computing dropout rate, is 100 (115 - 15 = 100). The dropout rate is 17%. If, in the following year, the dropout rate is 15%, the percent reduction in dropout rate is 2/17 x 100 = 12%.

The perverse incentive in this measure is an incentive to keep students on the rolls after they have dropped out. However, that is also a perverse incentive of the FEFP, since it rewards enrollment. Presumably the state regulations on counting enrollment are already sufficient to forestall this particular incentive. There would also be an incentive to water down courses, or to provide entertainment (including athletics) to encourage marginal students to stay in school. However, if this measure is only one of a number of measures, and achievement on standardized tests is another, this incentive should be counteracted.

16. (Middle, High school) Percent improvement in absenteeism rate. The first question faced in considering this measure is whether it is an outcome measure, or whether it is instead an input measure (the input is hours spent on education;

the output is educated students). I tend to think of it more as an input measure, but school absenteeism receives enough public attention that it us worth consideration as a measure for merit purposes. For the same reason that dropout rate is preferred to retention rate, absenteeism rate is preferable as a measure to attendence rate. The number of absentees is the number of students who, during the survey period, have unexcused absences. (I don't know whether a distinction is currently made between excused and unexcused absences. The idea is that illness is an excuse for absence.) The absenteeism rate is the number of unexcused absences expressed as a percent of enrollment. The percent improvement in absenteeism rate is calculated in the same way as percent improvement in dropout rate. A similar provision is made for declaring some arbitrarily low absenteeism rate to be automatically sufficient, and for assigning that rate some numerical value.

The perverse incentive here is to intensify efforts to get a note from parents for each student saying that the student was ill on a day when that student was absent, regardless of whether that is actually true. This is a difficult incentive to deal with. Presumably the FEFP has the same problem.

17. (All levels) Percent reduction in suspension rate. The reasons for this measure, which also is much like a process measure, are similar to those for the absenteeism measure described above. If schools see suspension as a way of getting rid of problem students who may drag them down on other measures, they will have an incentive to suspend. This measure, like the one on students held back, combats a perverse incentive of other measures.

There are important outcomes that have not been discussed here, primarily because there is no way to get an accurate quantifiable measure of them. Examples would be improvement of the artistic ability of students, or in general, performance in subjective areas of the curriculum.

I also have not included a number of potential measures because they are not outcome measures, but input or process measures. Examples would be percentage increase in the number of books in the library, or in number of in-service courses taken by teachers. Another is improvement in behavior of students. All of these, however, may be important preconditions for improving performance in these curriculum areas that are measured, and it could be expected that the school would take steps to improve them so it could improve academic performance.

IDENTIFYING MERIT SCHOOLS BASED ON THE MEASURES

Once a set of measures has been defined, the process for identifying the schools that deserve the term "merit school" must be established. Let us assume that there are 25 measures that are to be used in judging a school, of which 5 are local

measures. Some measures defined by the state might be used for all schools (a measure of absenteeism, for example), while others would only be used for some categories of schools. For example, performance on a standardized test at fifth grade level could not be used in a school that does not have a fifth grade. A specific local measure might only be used by one or a few schools. Some of the measures may be deemed by the state to be more important than others. All of this creates difficulties in creating a final ranking of schools for purposes of merit designation. I believe we should begin by separating schools into major categories that would be expected to have many of the same measures among schools in the group, but relatively few shared with schools in other groups. The most logical such grouping is into elementary and secondary schools, or perhaps elementary, middle school, and high school. There would be enough schools in each group to treat them separately. I would exclude, for the present, specialized schools such as those that are primarily vocational, or that primarily educate those in Special Education. If the intent were to award merit rating to 25% of the schools in the state, it could be agreed that 25% of the schools in each group would receive such awards. This would allow schools to compete with similar schools, and reduce contention that the measures are biased toward one type of school.

For each measure, the score on the measure for each school that uses that measure is listed, arrayed from lowest to highest. The school's score is replaced by its position on the list, with the school that scores highest having the highest number. Schools that score identically will all be given the same score, which is the average of their positions on the list. The schools are then given an adjusted score by dividing each schools's position in the list by the number of schools in the list, and mutliplying by 100. This insures that the school that does best on the measure will get a score of 100, and other schools will get less than 100.

An example may help. Suppose that on a certain measure there are 200 schools. The scores on that measure range from 36 to 80. Let us suppose that there are 3 schools that have a score or 55, and that this makes them numbers 119, 120, and 121 in the list when the scores are ranked form lowest to highest. Then each of the three is assigned to a position of 120 on the list. Each school's adjusted score on this measure will be 120/200 x 100 = 60. The top school in the state, number 200 on the list, would get an adjusted score of 200/200 x 100 = 100 (if there were no ties for top), and the bottom school, number 1, would get an adjusted score of 1/200 x 100 =0.5.

This, then, is a way of converting scores that have a variety of ranges into adjusted scores that all range from 0 (actually slightly above 0) to 100. There are other ways in which a conversion could be made. For example, one could calculate the range from lowest to highest, and calculate the difference between each score and the lowest score as a percentage of that range. This also would give adjusted scores

that range form 0 to 100. However, this puts too great a dependency on the extreme scores. A single school whose raw score is three times as high as any other school would bunch the adjusted scores for all of the rest of the schools together. My suggested method would not do that. Doing the calculations for each measure for a large number of schools is a laborious task by hand, but it is a cinch for a computer.

The use of local measures complicates things. If only one school in the state (or perhaps two or three, or those of only one county) use a measure, it is impossible to know how the school has done on that measure compared with other schools in the state. My suggestion is that the State Department of Education (or the FQIIC) work with representatives of the school districts and the unions to develop lists of acceptable measures that could be used at local option. The measures should meet the criteria I have listed earlier, and should be developed taking into account possible perverse incentives. A school is then free to use a measure from the optional list as a local measure, subject to the proviso that at least 20 schools in the state adopt it as a measure. If this is done, the local measures can then be incorporated into the scoring along with the statewide measures.

The adjusted scores for a school may now be summed into a final adjusted score. If the state decides that all of the measures should have equal weight in the determination of the final score, the process is simply that of summing all of the adjusted scores and dividing by the number of scores. This gives a final adjusted score. If, instead, the state wishes to put more weight on some measures than on others, each adjusted score is multiplied by the weight assigned to that measure, the weighted adjusted scores are then summed, and the total is divided by the sum of the weights.

It may be desirable to have site visits for those schools that meet some minimum criterion on the objective scoring. If merit is to be awarded to the top 25%, it might be desirable to have site visits to the schools scoring between 20% and 35% from the top. These site visits could perhaps measure additional, more subjective areas. They could do some auditing of measures where it seemed desirable. Perhaps more important, they could provide a shot in the arm to schools that might in the end not achieve merit, and provide help to the administration that needs support in getting the required autonomy to lead a school toward merit.

THE REWARD FOR DESIGNATION AS A MERIT SCHOOL

Up to this point I have described how a system can be developed for identifying merit schools. Ultimately, they will be the top 25% (or some other agreed-upon percentage) of schools in the state in each category (elementary, middle and secondary, for example) in terms of final adjusted score. Let us assume that the Legislature has designated a total amount that is to be

awarded to merit schools in a given year. There are two concerns: how the money is to be distributed among schools, and how it is to be distributed within a given school. The criterion to be applied here was expressed earlier: the awards must be designed to provide an appropriate and adequate incentive for improvement on the designated measures.

But us suppose that the amount provided by the state averages $3,000 per professional in each of the merit schools (or, say, $150 per student). Should the state give each merit school a flat grant of $3,000 per professional, or $150 per student? There are two questions here: should the distribution be based on the number of professionals, or on the number of students, and should the distribution be a flat grant, or be scaled depending upon the degree of merit? With regard to the first question, I believe the distribution should bwe based on the number of unweighted FTE in the school. Some schools may find that they can operate better by using fewer professionals and concentrating their resources on other things. This should not be discouraged. The individual child should be the unit of analysis.

The question of whether the grant should be scaled according to merit is somewhat more difficult. If there is just a flat grant of $150 per FTE to all merit schools, the highest scoring school may feel that it should get greater recognition. I rather favor breaking down the merit schools into two equal groups on the basis of final adjusted score, with the highest scoring schools getting, say, $175 per FTE and the lower half getting $125 per FTE. Breaking down any finer than this would, I believe, risk having the lowest scoring schools getting so little that they cannot use it as an effective incentive.

At the lower end, there will be great disappointment among the schools that did not quite qualify. I see no solution to this. To spread out the awards to more schools in ever-smaller amounts dilutes the effect. It is better to hope the schools that did not quite make it this year will try harder next year.

The distribution within the schools also deserves attention. One proposal would have the money distributed equally among teachers (and one would hope the principal and other professionals would also be included). I would not legislate this. I would leave the distribution up to the professionals at the individual school. Many schools might wish to distribute the money equally among the professionals, but some might wish to use part of the money for things that would help them to do a better job: a computer, better secretarial services, or whatever. (It would be delightful to see a school deciding to use part of the money to buy a better set of texts, or some laboratory equipment, to help them in getting a merit award the following year.) Or they might want to use it to reward school employees who otherwise would not qualify: attendance officer, teacher aides, etc. They might wish to distribute the money in other than equal amounts. I would put no restrictions on the way in which the

money is distributed, except to insure that the decision was reached in a manner that is reasonably democratic. One way might be to have the distribution decision made by a committee composed of the principal and either two or four teachers selected by their peers. Decisions would be made by majority vote, but this would insure that the principal's voice is heard.

SCHOOL CONTROL OF INPUTS

One of the criteria for the design of a merit schools proposal was that principals and teachers should have the maximum freedom to manipulate inputs and process variables in order to achieve the desired outcomes. I believe this is a very important point, and one that could make a merit schools proposal ineffective if ignored. The professionals in a school must believe that they have the means to make the school more effective. Otherwise, they are doomed in their efforts to improve education in the school. At present, teachers and principals often have very little control over these variables. The central office may hire teachers and assign them to the school, with little involvement of the principal. Textbooks are bought centrally, with little choice possible for a school. The pupil-teacher ratio is centrally determined, and may even be the subject of bargaining with the union. Teacher salaries are bargained, as are many working conditions. All of these restrictions make it very difficult for a school to do strikingly better by making major changes. At most, there will tend to be marginal improvements. What is necessary is a great deal more freedom for a local school to make decisions about how it will deploy its total resources, both manpower and money, to achieve its ends, within general district guidelines that should be focused on outcomes, not on inputs or process.

Of course, what I am proposing will not sit well with superintendents or unions (nor, for that matter, with the State Department of Education, which also has restrictive regulations). Politically, therefore, this part of the plan is difficult, but I believe it is essential if the concept of merit schools is to be really effective.

"Who Makes Up the CBEDS?"

by

Gene Dawson

CBEDS Background (1)

The California State Department of Education is developing a multipurpose data system on California education that contains basic information on staff, enrollment, finance, facilities, curriculum, and community demography related to public elementary and secondary schools. The California Basic Educational Data System, CBEDS, a part of the larger multipurpose data system, collects information on staff members and students at the county, school district, and classroom level. These data are collected once a year in October on "Information Day," then converted to file form. Subsequently, the data are used by the California State Education Department both for compilation of federal and state reports required by law and for response to state legislative requests for information, planning, and management. Certain CBEDS data are also made available to other state agencies, educators, and educational administrators for research and planning; to authorized professional organizations; and to universities and research organizations.

Available data are released in aggregate or partial form

only to authorized agencies or persons demonstrating a bona fide need for the information. Further, the California Information Practices Act of 1977 restricts disclosure of certain CBEDS data. The Act prohibits disclosure of personal information except for clearly defined official uses or for research when the individual to which it pertains is not identified.

Legal Authority (2)

The Education Code of California, beginning with Section 10600, provides for establishment of a basic education data system and requires schools, school districts, and offices of county superintendents of schools to cooperate with the Education Department in establishing and operating the system. Information collection through the CBEDS is mandatory, with the exception of the request for an individual's name and Social Security number on the Professional Assignment Information Form. Failure to submit information requested through the CBEDS results in using incomplete data for federal and state reports.

How CBEDS Works

As with all management information systems, (MIS), CBEDS has inherent problems. The problem of error control is the most pervasive. In a successful MIS, error is controlled by recognizing the interests of stakeholders, those involved with the MIS who have a stake in an efficient process and a use for a

reliable product.

Can CBEDS stakeholders and users control CBEDS error? First, are classroom teachers stakeholders in CBEDS? To be a stakeholder, one must have some positive interest in the system no matter how indirect or remote. In a typical MIS, the software salesman, the computer programmer, the CRT display clerk, the Chief Executive Officer, the data encoder, the stockholders in the company, all with varying degrees of direct contact with the MIS, have an interest in the successful operation of the MIS. They care what happens.

In contrast, classroom teachers are the "pieces" of data that other agents use. Ideally, these "pieces" should be dispassionate and would simply comply with the data requests. However, perhaps tacit recognition by the legislature that classroom teachers are not neutral bits of data, and may be actively negative, prompted the required compliance described in the CBEDS BACKGROUND section.

Second, are classroom teachers even users of CBEDS? The CBEDS 1981 Administrative Manual makes this statement:

> Features of CBEDS are:
>
> The collection of basic information on only three source documents. The County/District Information Form, School Information Form, and Professional Assignment Form collect basic information which

was formerly collected on over 40 different reports.(3)

For the classroom teacher who previously never filled out any of the forty eliminated forms, CBEDS adds to the paper work. As described in the CBEDS BACKGROUND comments taken from the 1982 CBEDS User's Guide, CBEDS is designed to serve a user far removed from the classroom. In the Sacramento office for CBEDS collections, even an enthusiastic supporter of CBEDS such as Vincent Madden, Manager of Data Acquisition and Forms Control, can find users only as close to the classroom as the district level, where, for example, CBEDS has successfully relieved the central office of reporting federally required ethnic counts. However, note that this relief, in addition to being outside the classroom teacher's concern, serves a part of the school system that is generally regarded by classroom teachers as the "other," the administration.

Vincent Madden suggests additional uses for CBEDS, such as providing data for union/district negotiations. Apart from the comments by district interviewees that local salary data were more accurate, this suggestion further illustrates the CBEDS designers' concept of who the potential users might be. No one sees the classroom teacher as a CBEDS user.

The solution to the California Basic Educational Data System problems will not be found with the stakeholders or users, because classroom teachers are neither. They are, however, the

agents that provide CBEDS data. In their self-reporting, they are the discrete pieces aggregated to form the CBEDS, but these pieces of individual data are not uniformly reliable. At the initial point of collection, error, here defined as incorrect data, enters the system for whatever reasons, rational-irrational, intentional-unintentional, that move human beings to do what they do.

Sources of Error

Error occurs even in the responses of those classroom teachers positively oriented to data collection and willing to comply with CBEDS. This class of error is illustrated by the classroom teachers who, misreading instructions, make mistakes in "bubbling-in" requested information on the electronically scanned response sheets. These are the teachers, for example, who honestly regard some of their activities as administrative and code themselves as part-time administrators.

Other errors also occur by mistake. For example, although the cover form for the school site asks for quantitative data (e.g., numbers of students per grade) the secretary makes one check per grade. Also, respondents, entering explanations of their anomalous assignments instead of bubbling-in the appropriate code, make "stray" pencil marks that confuse the optical scanner. In other cases, through inattention, the total number of students in the elementary school is entered in the

column of total graduated from high school.

Errors from faulty training are endemic. Following the hierarchical organization of the State Department of Education, CBEDS instruction flows down to counties that offer optional training to the district coordinator, who may or may not be the person actually coordinating data collection at the district level. From the county training session, the district coordinator arranges instruction for site principals in CBEDS terminology, interpretation, and changes from previous years. In this way, site principals become the first point of error control. The Administrative Manual outlines these responsibilities:

> The principal should check each completed Professional Assignment Form for completeness, accuracy, stray marks, and foreign objects. (4)

If only the principal would do this carefully, collection error would be eliminated from the CBEDS system.

The principal is indeed a pivotal agent; unfortunately, <u>all</u> educational agencies claim the principal as a pivotal agent. The principal is therefore often overloaded and responds to requests for accurate CBEDS correction by employing unsatisfactory coping mechanisms. Because the most functional response, hiring more help, is usually denied them, principals may choose dysfunctional

responses. Principals may (1) filter, by giving attention to the most demanding aspect of the collection, which may be to turn the forms in on time without close correction; (2) queue, by placing CBEDS in a waiting line where corrections may not be first in priority; (3) omit, by reducing effort on CBEDS; (4) approximate, by giving the CBEDS a gross rather than fine examination; or (5) trade errors, by accepting a higher error rate in exchange for a rapid CBEDS return.(5)

So far, the discussion has addressed unintentional error by the initial respondent, but intentional error, mentioned frankly in the interviews, raises the question: Why would classroom teachers refuse to cooperate in a data collection that, as described in the manual, seems benign? Writers in the field of implementation stress involvement of all actors in the process, reinventing the wheel if necessary, to build commitment to the program. However, CBEDS implementation is top down with consequent lack of understanding or commitment to CBEDS by classroom teachers, the initial data providers. During the first year of CBEDS collection, charges of "Big Brother," suspicions of the use of the data, and a general resistance to "more paper work," were expressed by deliberate falsification of names and reporting salaries as either absurdly high or low.

To ease teachers' concerns over privacy issues raised by the unions, name and Social Security number are not now required,

only the general descriptive information. By 1982, CBEDS was accepted by most teachers as just one more of the many required forms. However, indifference is not commitment, and interest in correct data collection is missing on the part of the classroom teacher. The form is considered too complicated. Teachers still will not bother to determine the correct code that describes their teaching assignment, which is a continuing source of CBEDS error.

Active opposition continues to be expressed by teachers who refuse to fill in names or Social Security numbers, erase names and numbers from their pre-printed forms, and neglect to return their CBEDS forms at school sites where principals do not require CBEDS collection. In removing the need for personal identification by name or Social Security number, incorrect responses cannot be traced to the respondent. Alienated teachers can make a decision to comply or not. They can exert disruptive power in their work life with little risk to themselves.

The several rationales offered here for the non-compliance action of teachers are descriptive of the real-life situations in which principals, as line supervisors, find themselves. Unable to reward or punish except in petty ways, subject to charges of harrassment and grievance by the union, unable to coerce, only request, the principals often believe themselves helpless as the classroom teachers have their way with CBEDS. Principals call

district coordinators, "What should I do?" The CBEDS coordinators, believing themselves equally impotent, reply, "Nothing."

Error Control

Relationships between classroom teachers, principals, district coordinators, and county offices, with the Data Acquisition and Forms Control (DAFC) office, are more clearly understood if an observer regards these agents as part of the collection process and NOT as users. Only one of the interviewed school districts used CBEDS information for local purposes, and that was in a trivial public relations demonstration of how the California teachers are aging as a cohort.

The CBEDS users, as described in the CBEDS BACKGROUND section, include large professional groups, the university system, and the legislature. However, the user of first importance is the Department of Education, itself, with the legislature as a close second. The challenge for DAFC is to provide timely, relevant, and accurate information to these two users. While the quality of timely or relevant data is a function of the state processing agencies, accuracy of the CBEDS is a function of the collection process.

Discussions in the previous sections indicate that CBEDS is not accurately collected. Because of the now possible anonymity,

no direct feedback to the individual is possible. Further, while compliance with CBEDS is mandatory, the "penalty" might even reinforce those respondents, concerned over "Big Brother," who might want to sabotage the system:

> Failure to submit the information requested through the CBEDS will result in the use of incomplete data in federal and state reports. (6)

The DAFC does try to improve CBEDS collection accuracy by writing careless districts with exhortations to do better in the future, but the plea to correct the salary reporting errors of 1981 had no punch. Since the DAFC has no control of CBEDS accuracy at the collection point with the individual respondent, control is established at the aggregate point in the system through predetermined knowledge of how the collected data should appear. Elaborate controls at the DAFC reduce error to an acceptable level. "Acceptable" seems to be defined as error that, in practice, can be confined within logical parameters. For example, the computer error control program "flags" salary amounts outside the state minimum/maximum. Clerks then assess the flagged notions. Salaries marked as

(2) (0) ( ) ( ) ( )

with the three blank bubbles, are assumed to be

(2) (0) (0) (0) (0) (0);

the clerk, filling in the bubbles, corrects a careless error. A salary reported as

(0) (0) (0) (0) (0)

is assumed to be an intentional error and is not included in salary averaging. Other flagged logical errors include free lunch counts greater than the total school enrollment, no graduating seniors in high schools with undergraduate classes in previous years, graduating seniors in elementry only school districts. Clerks call the school districts to clear ambiguous flagged responses. However, if the outputs are not flagged, error is undetected.

The Data Acquisition and Forms Control office is aware that as data become softer reliability weakens. The review board has rejected some requests for data collection as being impossible to determine by the classroom teacher, e.g., who drinks more of their lunch milk, boys or girls? Some items have been discontinued because they were too ill-defined, e.g., do you have inservice at your school site? By concentrating on the logical limits of CBEDS data error correction, and by not pressing for the impossible goal of, for example, totally accurate salary data, DAFC has chosen to accept a goal they <u>can</u> reach and a printout they can deliver to their users.

The DAFC has chosen the cybernetic model of system control,

the management by exception model. With this model, unanticipated consequences may develop. Suppliers of CBEDS data had an unpleasant surprise in March 1982, for example, when half of the California school districts received letters from the Local Assistance Bureau, which is responsible for monitoring legislated teacher-administrator ratios. These letters warned the school districts that their ratios were out of compliance and that the districts were subject to fine. Previously, this ratio report had been prepared by the local school district for the state to review. Since 1982, figures for this ratio computation have been taken from general CBEDS data sent the state by each school district. For the first time, mistaken coding of assignment by teachers, careless transmission of data by principals, lax supervision by district coordinators, and general tolerance for high error rate were revealed in the system.

CBEDS Collection Compared

Accurate CBEDS data <u>can</u> be collected at the local level, then forwarded to the Data Acquisition and Forms Control office, but the key persons among the actors are the district superintendents. These superintendents focus the interest of their staffs on accurate data collection. A comparison of two large California school districts illustrates the effect of superintendent commitment.

Increasing credibility with the state legislature and the

Department of Education through accurate reporting is the goal of the superintendent of the Los Angeles Unified School District. This emphasis is reflected by assigning a full-time staff member to ensure accurate CBEDS collection. In a system unique to the Los Angeles Unified School District, CBEDS is not collected separately for the state report, but is incorporated into the district data collection process in the fall. The district form gathers data of general CBEDS interest as well as data of specific interest to LAUSD such as more detailed ethnic breakdowns or curriculum offerings.

Because LAUSD has a sophisticated personnel information management system containing hard data such as the several job codes, salaries, ages, and years in the district, these data are easily pulled out for the CBEDS report and are accurate. CBEDS data, gathered from the LAUSD Information Form, are combined with the personnel data, then sent on tape to the Data Acquisition and Forms Control Office.

Los Angeles Unified School District site personnel are commited to data collection because the process is not seen as just another paper request from Sacramento but as the LAUSD information request clearly supported by the superintendent. Further, the collection feedback loop circles to someone who can make a difference at the initial point of error control, the site principal.

The first year the district information day was tried, errors were apparent. The central office held a "Correction Day" for principals with the forms from their sites. The LAUSD principals themselves, reviewing each form, made the corrections with wailing, gnashing of teeth, and much distress. However, since that first "Correction Day," the information forms have come from the sites filled in accurately, completely, without foreign objects or stray marks, and with 100% return.

In contrast, School District B assigned an early retiree as coordinator of the 1983 CBEDS collection. This coordinator had not been responsible for CBEDS before, did not have administrative clout, could not compel either 100% return or principal accuracy checks, and did not have either enough time himself or enough staff to check the forms even for completeness.

In an attempt to increase CBEDS accuracy and avoid the threat of fines from the Local Assistance Bureau (LAB) for another incorrect teacher-administrator ratio, District B did not use the preprinted forms from the state. Instead, respondents filled in all blanks anew. In addition, job codes were simplified. These attempts at an effective remedy failed - School District B received the LAB warning letter again in Spring 1984.

The LAB warning letter indicated illegal ratios, but these ratios were not a true reflection of the School District B

teacher-administrator ratios, which, in fact, were in compliance. The ratios, now computed from CBEDS data, have become an indicator of general careless CBEDS data collection. Teachers who mistakenly code themselves as administrators, teachers who feel they perform as administrators and therefore code themselves as administrators, and teachers who fail to return their CBEDS form, falsely weigh the ratios on the administrator side.

Because the district was liable for a fine, the ratios were corrected in Spring 1984 by the personnel officer who had the job of correcting the CBEDS based ratios each previous year. This task was not eased by the School District B personnel management information system, which lacks the flexibility to serve other than district basic needs. For example, in the computer only one prime job number can be listed for an employee even though the employee might be working in two different jobs.

But most importantly for School District B, the feedback loop to correct error does not return to the key collection agent at the first point of error control, the site principal. In contrast to Los Angeles Unified School District, where "Correction Day" made the principals sharply aware of error consequences, site principals in School District B never received feedback of any kind, good or bad. Seeing no reason to change their response, they did not.

Recommendations for an Accurate CBEDS Collection

1. CBEDS data gathering must be actively supported by the district superintendents who, in addition to their personal sponsorship, assign line administrators to coordinate the project. Because the site principals are not automatically either CBEDS users or stakeholders, the sponsorship of the district superintendent is necessary to increase the site principals' commitment to CBEDS and cause them to function in the role of collection point error control agents.

2. Modify personnel management information systems so that accurate data already available to the district can be easily gathered for the CBEDS report. CBEDS data collection will be eased by coordinating state and district requests for data, which will in turn encourage participant commitment by rendering CBEDS data locally useful.

3. Re-design the error feedback loop coupling so that those responsible for the careless transmission of error become those responsible for the error correction.

Footnotes

1. CBEDS Data User's Guide (1982 Data). CBEDS California Basic Educational Data System, California State Department of Education, Wilson Riles, Superintendent of Public Instruction, Sacramento. (p. 1.)

2. Ibid., p.1.

3. Administrative Manual for CBEDS Coordinators and School Principals. CBEDS California Basic Educational Data System, California State Department of Education, October, 1981. (p. 2.)

4. Administrative Manual for CBEDS Coordinators and School Principals. CBEDS California Basic Educational Data System, California State Department of Education, October, 1983. (p. 8.)

5. Perrow, Charles, "System Accidents: Complexity, Coupling and Catastrophy," manuscript, 1981.

6. CBEDS Data Users' Guide (1982 Data). CBEDS California Basic Educational Data System, California State Department of Education, Wilson Riles, Superintendent of Public Instruction, Sacramento. (p. 2.)

Problems in Measuring School Reform

Edward Haertel

School of Educaton

Stanford University

The only way to convince the skeptics that the schools are reforming will be to show them numbers. We must find quantifiable measures of schools and the process of schooling.

This evening, I would like to discuss three of the measures that are being considered as possible quality indicators for California's schools: First, SAT test scores; second, course enrollments; and third, hours of homework or number of writing assignments completed. I will also comment on the ways standards for these and other quality indicators might be established. Before turning to these matters, however, I need to say a few words about measurement itself.

Measurement is the process of matching numbers with objects in a way that reflects some quality of those objects. The objects may be students, classes, or schools, and the numbers may tell about achievement, attendance, or coursetaking. Whatever the specifics may be, if the measurements are to be fair and trustworthy, at least two basic principles must be observed.

The first principle is that different measurements may tell very different stories. Numbers reflecting different qualities are not interchangable, but once the complexities of the objects have been reduced to a tidy set of scores, everyone can see which is highest, next highest, or lowest, and it is all too easy to forget just what story those numbers tell.

The second principle is that when the measurements really

This is an expanded version of remarks presented at the Stanford University School of Education on June 14, 1984, in a panel discussion with State Superintendent Bill Honig and Professors Larry Cuban, Sanford Dornbush, Edward Haertel, and Michael Kirst of Stanford University.

matter, there had better be an intimate link between the numbers and the qualities they are intended to represent. If there is not, if the numbers can be changed without changing the qualities, then the measurements may soon lose their intended meanings.

I will briefly expand on these two principles, and then apply them in considering the three proposed quality indicators for California's schools.

Let me return to my first principle. Different numbers tell different stories, and we need to be clear on what any chosen set of numbers represents. If we want to measure quality X, then a measure of Y won't do. Coursetaking is not the same as learning, aptitude for college is not the same as content mastery, and the answers to simple questions about amount of homework or number of writing assignments are not the same as intellectual rigor. The problem of deciding what qualities of schools should serve as markers for reform is difficult; it cannot be finessed by seizing upon whatever numbers may be at hand, and letting the definition of reform default to whatever those particular numbers happen to signify.

Now consider the second principle, that if measurement is to drive reform, there must be an intimate link between the numbers and the qualities they are intended to represent. The technical term for this linkage is intrinsic validity. It implies that no harm should come from direct attempts to improve the scores themselves. This is important, because if schools or teachers are rewarded for better-looking numbers, they will do what they can to make the numbers improve. Note that measurements may be valid, and useful for some purposes, without possessing intrinsic validity. For example, number of books in the home may be a valid and useful measure of the learning environment, but that does not mean that putting more books in the child's home will automatically help her achievement.

Keeping these two principles in mind, I'd like to consider some of the measures proposed for the schools of California. Let's begin with the SAT test scores. You've all heard that the SAT is a poor choice as an indicator of educational health, because the students who take it are self-selected. This is a telling objection to the use of average SAT scores, because the average could be manipulated by encouraging only the best students to take the test. We can get around that problem by looking not at average score for a high school, but at the percent of ALL seniors who score--let's say--over 600. That way, there is an incentive to have as many seniors as possible take the SAT, and also to have as many as possible do well on it.

Questions of what the SAT numbers signify, and of how test

performance is linked to the processes we wish to reform, cannot be disposed of quite so easily. What is the SAT measuring? What do the numbers tell us? Quite simply, the SAT was designed to predict the success in college of individual students. It minimizes, as much as possible, the influence of particular advanced high school courses, and measures general learning skills developed over the entire twelve or so years of schooling. No specific content from advanced courses in English, literature, mathematics, or the sciences is required to earn a perfect score on the SAT. Of course, formal schooling is the major influence through which the aptitudes measured by the SAT are developed, but nonetheless, the test is resistant to the effects of specific courses, curricula, or school experiences, and this is by design! Most colleges use the SAT in conjunction with a student's high school transcript and grade point average, precisely because the SAT scores and the high school record are measures of DIFFERENT things. Now, aptitude for college is a schooling outcome that is worthwhile measuring, at least for students planning to go to college. We could question the importance of measuring college aptitude for ALL students, but that is a matter on which reasonable persons may differ. My first point is simply that we should get straight what it is we are choosing to measure when we select the numbers generated by the SAT.

Now consider the second point. What is the linkage between SAT test scores and the schooling processes we are attempting to reform? If the SAT had intrinsic validity, another question that would serve as well would be, "What reasonable steps could a high school take to increase its percent of 600-plus scorers on the SAT?" This is a question our policymakers may not have stopped to consider carefully. The answer isn't obvious. Or rather, the obvious answers don't bode well for the hoped-for reform. First of all, getting more students to take more advanced courses may not be a school's best strategy. None of the content of foreign language courses, chemistry, physics, or even advanced algebra is likely to help very much at all. Courses in English literature, by increasing vocabulary, might help some. Writing assignments are unlikely to raise scores, because no writing is required on the SAT. The fact is, we don't really know exactly how to develop the aptitudes the SAT measures, or what makes average SAT scores go up or down. A review in 1977 listed 89 different explanations that had been proposed for the SAT test score decline during the 1970's, from TV to the smorgasbord curriculum to changing perceptions among high school students of the test's importance. But research has provided some answers. Numerous studies of coaching for the SAT have shown modest gains when students are given direct practice in answering multiple-choice items like those on the test. The expected improvement in SAT scores is predicted quite well by the number of hours students spend practicing test taking. The more time is diverted from other

instructional activities to the task of answering multiple-choice questions, the more scores may be expected to improve. Even training in test-taking strategies is not much better than simple drill-and-practice in answering items. There are SAT preparation courses like this now in many high schools, because good SAT performance is already important to many college-bound students, to their parents, and to their teachers. But it is difficult to understand how creating an incentive to divert non-college prep students into such classes would encourage healthy reform.

There is another way for a school to increase its percent of 600-plus scorers which may be even less benign. A second way to maximize the proportion of high-scoring seviors is to encourage less able students to drop out. This incentive toward higher attrition may be further strengthened by other proposed quality indicators, for example the percents of students taking advanced courses. It needs to be given careful attention. The potential benefits of having weak students drop out might not be eliminated simply by adding low attrition to the list of quality indicators. Some very careful balancing of rewards and penalties across indicators would be called for to get just the mix of lower dropout rates and higher test scores that the policymakers had in mind.

I'd like to turn now from the SAT to another of the proposed quality indicators. Take course enrollments, or courses completed. These are numbers derived from students' transcripts. A school's score would reflect the percent of students taking advanced foreign language courses, physics, intermediate or advance algebra, and so forth, or the percent who had completed some particular set of courses, perhaps the four years of English, three years each of math, science and social studies, and half-year of computer science recommended by the National Commission on Excellence in Education. Again, I would ask first what these numbers really signify, and second, how they are linked to the instructional processes we wish to reform.

On the first count, the coursetaking numbers fair better than the SAT numbers. They are more direct indicators of some specific aspects of academic preparation. On the second count, there may still be cause for concern. The coursetaking numbers may be valid, but may lack INTRINSIC validity. In other words, the linkage between these numbers and academic quality might be seriously weakened once schools were given strong incentives to make the numbers go up. Course titles are like package labels, and we have no sure way of knowing what's inside. The easiest way to improve the statistics would be to change the labels on existing courses. State-adapted course content frameworks or lists of approved textbooks could curb this relabeling somewhat, but even listing the ingredients on the outside of a package may not tell us much about the QUALITY of the contents. Many of the

courses to be monitored are in areas where there are already critical shortages of trained teachers. Pressing the enrollments in these courses to new highs will result, in the short run, in either larger classes or instruction by less qualified personnel. The students now taking the advanced courses are probably among the best prepared in their schools. If students with lower developed aptitude are introduced into advanced classes, the level of instruction and the amount of content covered may decline, so that the larger quantity of advanced training is partially offset by its lower quality. The alternative of tracking, of creating less advanced 'advanced' courses, invites the subterfuge of relabeling courses taught already.

Let me turn finally to the numbers proposed to tell about writing assignments and homework. First, what do they signify, and second, will the things that could be done to change them bring schools closer to the goals of the reform? Homework, intelligently assigned, responsibly completed, graded, and returned, should lead to higher achievement. Also, researchers have concluded that students are not called upon to write nearly enough. But all writing assignments are not equal, and homework for homework's sake may do no good at all. As things stand now, homework and writing would be assessed by a few items on the California Assessment Program Student Information Question Sheet. All that these questions ask about is quantity. They cannot distinguish wise from unwise uses of homework, and the question of whether homework is graded and returned is not even asked. As to writing, a single question asks the number of writing assignments done for school during the last week. A paragraph counts the same as a theme. My point here is not simply that these questions on the CAP need to be improved. There is a deeper problem in substituting measures of quantity for measures of quality of instructional processes. The closer we move to assessment of actual learning activities, the more difficult it will become to obtain meaningful measurements of quality. The answer to my first question, then, what do these particular numbers represent, is that they probably don't represent very much. The answer to the second question, will steps taken to improve these measurements help achieve the hoped-for reform, is equally pessimistic. Making the sheer number of writing assignments or hours of homework an end in itself is an invitation to meaningless work. I have said further that I believe this is more than a simple problem of elaborating or rewording a couple of questions. It is far easier to count hours or papers than to inquire into their instuctional worth.

Thus far, I've described technical problems with several of the quality indicators that might be used to direct and document the planned reform. In the course of critiquing these particular indicators, I have tried to illustrate the kind of scrutiny that

any proposed index needs to be subjected to. There are some proposed indices that fare better than the three I've just discussed. Tests of achievement, rather than aptitude, should be valuable tools in improving learning. For the time being, the Advanced Placement tests by ETS or the College Entrance Examination Board's Achievement Tests are promising candidates. When the Golden State Examinations become available, they should be even better suited to California's needs. The CAP tests at grades 3 and 6 are of some value now, and problems with the grade 12 tests have been recognized and should be dealt with. Nothing that I said about the proposed use of the SAT should be construed as a criticism of testing per se, or of the SAT when it is used as it was designed to be used. But modern educational tests are precise tools, and they cannot be used interchangably. This is a mistake that has been made before. Perhaps the best-known example is the sorry history of the Project Headstart evaluation, where IQ tests, which measure general aptitudes, were used to assess program impacts. If specific achievement measures had been used to see if students learned what they were supposed to have been taught, Project Headstart might have been seen in a happier light.

In the time remaining, let me turn from the problem of measures to the problem of standards. Once we have found numbers that faithfully represent the qualities to be measured, and that have intrinsic validity, how are they to be used? There are three basic approaches, and all have their shortcomings. We can specify standards in terms of norms, in terms of changes, or in terms of some score. This is an invideous approach, because many schools near the bottom on any given indicator may in fact be doing a very good job, given the backgrounds of the students they serve. A simple ranking that is blind to these 'input' differences seems inherently unfair. Specifying standards in terms of changes gets around this problem by letting each school's previous performance serve as its own baseline. Under this scheme, schools would be rewarded for improvements in scores from year to year. The first objection that arises to this plan is that in some districts, there is very little room for improvement. If test scores, attendance, coursetaking patterns, homework, and so forth are consistently excellent, the district is actually penalized! Any scale will have some floor and some ceiling. Above a certain level of excellence, further efforts at improvement will bring diminishing or vanishing returns. But the change criterion may penalize not only those fortunate schools and districts near the tops of the scales, but also the districts working hardest with less advantaged students. Even in schools far from the tops of the scales, the reasonable steps that could be taken to improve performance may have been taken already. The schools already working most diligently to educate all their students may be the ones that lose out, unless they deliberately do poorly during the baseline year. The third method of

specifying standards, in terms of targets, would set goals for each school and reward those that met their goals. This would work fine if there were some way to assure the setting of reasonable goals. Individual schools or districts would have every reason to set targets that were as easy as possible to attain, if monetary or other rewards were attached to their attainment. If the state dictated the goals, they could not be uniform for all districts, because the educational needs of students in different districts vary so greatly. Setting differentiated goals for different districts could become enormously complex, burdensome, and politically charged. In addition, it would tend to give the appearance that different degrees of excellence were expected of different students, that all were not deemed capable of the same performances.

A method of standard setting that avoids some of these pitfalls, and that has in fact been proposed, is to divide schools somehow into more comparable groups, and to look at the percentile rank of each school compared to other schools with students having similar educational need. There is still the danger that people may construe students in the different school groups as inherently unequal, but if care is taken in the use and interpretation of norms constructed in this way, they may offer the best hope of an equitable solution to the standards problem. One refinement that might be considered would be some kind of appeal process for districts experiencing unprecedented demographical changes or disruptions, e.g., school closings or a large influx of immigrant students.

In summary, I have discussed some of the indicators proposed to measure the quality of California schools. I have urged more careful attention to the question of just what qualities the proposed indicators actually reflect, and also to the question of what might happen when schools are given an incentive to improve their standing on those indicators. I have tried to show that some plausible measures may not mean what they appear to, and that direct attempts to change even valid measures may have negative consequences. I believe that technically sound measures, appropriately chosen and wisely used, can aid significantly in the effort to improve the quality of education in the state of California, but these are matters that will require more careful attention if the efforts at reform are to succeed.

In closing, let me urge that however the reform is assessed, the greatest hope for constructive change lies in trust, cooperation, and commonality of purpose among all parties. Students, parents, teachers, principals, district and state administrators, scholars and the public must work together, not as adversaries, if California's schools are to reach once more toward excellence.

Toward a Statewide System for Public School Accountability: A Report from California

David Stern

University of California,
Berkeley
October 1984

This is a PACE Project sponsored paper. PACE, Policy Analysis for California Education, is a joint undertaking located at the University of California, Berkeley and Stanford University. Its Directors are James W. Guthrie and Michael W. Kirst. PACE is funded by The William and Flora Hewlett Foundation. However, the analyses and recommendations contained in this paper are not necessarily endorsed either by the Hewlett Foundation or the PACE directors.

The author wishes to thank James Guthrie, Richard Murnane, and Richard Pratt for helpful comments on an earlier draft.

Toward a Statewide System for Public School Accountability: A Report from California

David Stern

If people who work in schools were rewarded for achieving certain results, more of those results might be achieved. But this is much more easily said than done. Performance incentives require information that is clearly related to legitimate objectives. This paper discusses how to define and use such information. For concreteness, the paper focuses on a particular plan for performance incentives presently being implemented in California. The overall conclusion is that the plan as presently proposed represents a big step in the right direction, but most of the purported benefit will not occur unless the state delegates more control to local school authorities.

Background

A renewed popular concern for educational standards, and the Reagan administration's policy of reducing the federal role in education, have stimulated much recent reform by educational authorities at the state level. For example, in 1982 and 1983 most states increased course requirements for high school graduation, and many were trying new forms of differentiated compensation to reward good teachers. These two reforms and a multitude of others in the same vein were enacted in 1983 by a massive (214 page) reform bill in California. As in many other trends, California is moving faster and perhaps going farther than most other states.

California's superintendent of public instruction, Bill Honig, was elected on a tougher-standards platform, helped push through the 1983 reform package, and is advocating even more far-reaching measures. In an article on "Setting the Course for School Reform" (1984), Honig outlined a "merit schools" plan that would go over the heads of school districts and make individual school sites directly accountable to the state in certain ways. The first step in this "statewide accountability strategy" is to define a set of "specific performance indicators." The state will rate each school and make the results public. This step is

being implemented in fall, 1984. The next step is for the state to award money to high-scoring schools:

> "After assigning relative weights to the measures mentioned above and providing an overall school score we would reward the schools that are in the top 20 percent of their socioeconomic group--for actual performance-with a $50-per-student, three-year grant to be used at the schools' discretion. We would also provide a similar amount to the top 20 percent of schools in each group that grew the most in terms of overall academic scores based on a three-year average. Being in the top 10 percent in terms of either level or growth in specific individual categories--such as dropout rates or enrollment in advanced-placement courses--might qualify a school for a $5-per-pupil grant."

This second step will require legislation which has not yet been enacted.

Honig's proposal, and others like it, raise important and difficult questions: What should be measured? Does grouping schools by student characteristics make comparisons more fair? How can performance grants be made effective as incentives for improvement? How can contributions of individual schools be measured accurately when students move from one school to another? What additional data are needed?

Superintendent Honig's hope is to create a "powerful engine for school improvement." The point is not accountability for its own sake, but the continual improvement of educational programs. This requires both the commitment and resourcefulness of teachers and administrators. The Honig proposal, like other plans for payment by result, is designed to reinforce educators' commitment by offering cash grants <u>contingent</u> on high performance. This is different from the usual procedure. Even when, in recent decades, funds have been given to schools for certain designated purposes, payment has not <u>depended</u> on evidence that the purposes are actually being met. In theory, paying for results can enhance motivation, and one of the issues to be considered here is how that effect can be achieved in practice.

This paper argues that the proposed "merit schools" plan should be both refined and extended. Since educational improvement depends on both the commitment and the resourcefulness of teachers and administrators, enhancing motivation alone is not enough. To increase resourcefulness as well, data bases and analytical capability being assembled to

implement Honig's plan can also be made available to teachers and administrators in local districts and schools. Creating the statewide system can be a step toward building a decentralized "educational management information system." As described by Otto Davis and others (1984), such a system would enable local personnel to "investigate the educational process and. . . monitor student outcomes. . . to provide feedback and decision making support." It would "bring together administrators, teachers, and other practitioners in a common effort to understand and to improve the educational process and student outcomes" (pp. 4-5). The last section of this paper will discuss in concrete terms how the current momentum for reform could be carried this far.

Choice of Measures

In developing a statewide accountability system, the California department of education published an initial list of 37 performance indicators for individual schools (see Attachment I). These would be measured, where applicable, for all schools in the state. Each district would also add to the list, but these local measures would not be used in allocating merit grants. In keeping with the emphasis on final results, the state list contains more performance measures specifically for high schools than for elementary or middle schools. The list includes 8 measures of enrollment in advanced academic subjects, 19 indicators based on test scores, and 10 other measures. The indicators are not all independent of each other. For instance, proportions of students taking calculus, chemistry, physics, and foreign language are each counted seperately, as is the proportion of students meeting course requirements for admission to the University of California. However, these latter requirements include two years of math, a year of laboratory science, two years of a foreign language, in addition to a year of American history, four years of English, and one advanced elective. This double-counting obviously gives these indicators more weight.

The initial selection of measures reflects an emphasis on advanced academic study that is characteristic of the current reform era (see, for instance, the report of the National Commission on Excellence in Education, 1983). There are 11 separate measures based on examinations for college entrance. Only three measures, related to attendance and retention of students, reflect how well a school is doing with students who are not academically inclined.

State Superintendent Honig has an advisory committee of local district superintendents, a sub-committee of which reacted to this list of school performance measures. Their memorandum declared that:

> . . . the raising of standards does not mean lessening the commitment to provide high-quality education to a diverse student population. We need to continue our efforts to provide for those students who, because of language, emotional, physical and cultural handicaps, require a special type of program in order to help them reach their full potential." (Subcommittee on Quality Indicators, p. 2)

Among the local superintendents' specific recommendations were inclusion of occupational training, arts, social science, foreign language, and physical education on the list of courses in which enrollment is to be counted. They also suggested developing performance indicators based on crime rates of graduates compared to dropouts, and on graduates' success in the job market.

Similar concerns were expressed by school board members and local administrators in a series of workshops where the accountability plan was presented by officials of the state education department. A description of the workshops concluded:

> " 'Elitism' was a frequent response of the workshop participants to the accountability program. Many in the audience felt that the program was focused too heavily on high achieving, college bound students which make up 10 to 15 percent of the student population. The increased demand implied by the accountability program would have the effect of frustrating the other students and increasing dropout rates. The answer to this objection in that schools should no longer write off 85 percent of the student body as incapable of achievement in academic subjects, and that such academic achievement is very much needed by a growing proportion of students to comete effectively for jobs. Having a solid foundation in academic subjects and having the learning skills that come with such a foundation will be important for getting jobs and succeeding at work." (California Department of Education, 1984)

The ancient question of what to teach in school is not likely to find its final answer in California or any other statewide accountability system. However, if performance measures are defined in such a way that they seem elitist or irrelevant to large numbers of students and teachers, the accountability system will not achieve its intended motivational purpose.

Average versus distribution. In addition to emphasizing achievement in advanced academic subjects, the proposed California standards include no explicit incentive to improve low-achieving students' achievement in basic subjects. The only measures of achievement in basic subjects are from the California Assessment Program (CAP), and these are average test scores at selected grade levels for each school. Average scores measure only central tendency, not dispersion. Two schools may have the same average, but one may have both more high-scoring and more low-scoring students. There is evidence that most teachers care about the dispersion of scores, and try to help the slower students catch up (Brown and Saks, 1981). Many teachers believe it is important to prevent slow students from failing, as well as to stimulate quick ones. This value could easily be incorporated in state standards. For instance, schools could be recognized or rewarded for having fewer students in the lowest quartile of the statewide or national test score distribution. Again, not including such measures will reduce the accountability system's legitimacy, and probably its effectiveness, for teachers who value prevention of failure.

Level versus change. Another general issue in choosing performance measures is whether to reward schools that score at a high level or schools that show the most improvement (see Garms, 1984). Obviously, an elementary school whose students are performing at low levels when they enter kindergarten, or a high school whose freshmen enter with only fifth-grade reading skills, will be at a disadvantage if performance levels at grade three or grade 10 are used to measure merit. Teachers and administrators in such schools are likely to be more demoralized than stimulated by such an incentive program. However, if merit is defined in terms of change rather than absolute level, these schools can compete. This is sometimes called the "value-added" approach, by analogy with the production of commodities such as bread or cars. The value added by bakeries or automobile makers is the difference between the price of their finished products and the cost of flour, steel, or other intermediate goods the bakeries or car makers must buy.

Using change or value added has certain disadvantages, however. Some achievement tests have ceilings, and some other

performance measures like attandance and retention also have maximum values. Schools at or near the maximum have little or no room for improvement. Using only value added may therefore provide no incentive for high-performing schools. Besides, a "merit schools" program will not only create incentives for educators but will also hold up certain schools as examples for the public, and if these are predominantly low-achieving (but fast-improving) schools, the public may wonder whether the state really knows what a good school is. For these reasons, the Honig plan in California includes both level of performance and amount of improvement as criteria for merit awards. (Simple change, however, may not be the best measure of value added: see the section on comparing schools, below.)

Combining measures. The initial list of statewide indicators contains 37 items, all measuring performance levels. For some or all of these, change scores will also be computed in determining merit awards. In addition, the initial list may be augmented to reflect educators' concern for students who are not in the top part of the achievement distribution, and their concern for dimensions of achievement other than in advanced academic subjects. The final list of performance indicators could well exceed a hundred items.

Should there be a separate award, or set of awards, based on each measure? In a state with more than seven thousand public schools, providing awards to 30 or 40 percent of the schools, as proposed in the Honig plan, would mean at least two or three thousand awards, if a school can get more than one. It would certainly be feasible to allocate these awards on the basis of a hundred different measures of merit.

However, Honig proposes "assigning relative weights. . .and providing an overall school score." Similarly, in selecting merit schools in Florida, Garms suggests that "the state" may wish to "put more weight on some measures than on others" (1984, p. 18). To derive such weights, there are two requirements: a sample of judges and a procedure to elicit judgements about relative importance.

The sample should represent various groups of "stake-holders" in education. Parents of public school students should be included. Students, at least from high schools, should be represented. Since taxpayers support schools, representatives of the taxpaying public belong in the group. Teachers and administrators, though their role is to provide the service rather than to consume it or pay for it, would be hard to exclude for political reasons--and, again, the objectives have to seem legitimate to them in order to be effective as motivators. School board members sould be included, too.

Once the sample of judges is chosen, there exist several procedures for determining weights. A simple procedure would have each member of the sample assign each performance indicator a numerical weight, from one to three, one to a hundred, or whatever-- where a larger number signifies the performance indicator is judged to be more important. Each indicator's average weight would then be used to construct an index of merit, equal to the sum of performance indicators each multiplied by its average weight.

An example of a more complicated procedure for deriving weights is "conjoint analysis" (see Rao, Gerritz). This has been used in marketing research to discover what differences between competing brands are important to consumers. A sample of respondents is asked to make hypothetical choices--in this case, the choice would be among a set of hypothetical schools. Each school would be described in terms of performance indicators: high on some, low on others. Respondents' ratings of schools can be analyzed to provide numerical weights signifying each performance indicator's importance. Results would probably be more accurate, in the sense of predicting how people choose schools, than would results from the simpler weighting procedure. However, conjoint analysis is usually done with only three or four dimensions to be considered, not several dozen.

Whether simple or sophisticated, procedures for eliciting weights are inevitably artificial and somewhat arbitrary. Arbitrariness is also a problem in selecting the sample. In practical terms, the resulting weights are likely to be unstable. If new weights are generated in subsequent years, they are likely to cause changes in the rankings of schools on the index of excellence--changes that are not due to changes in actual performance. Arbitrary and unstable weights will not guide a steady effort toward educational improvement, but, instead, will promote fickle and superficial changes.

It is worth mentioning one other procedure, which can generate weights for performance indicators without asking arbitrary questions to an arbitrarily chosen group of people. This is a linear programming procedure, known as "Data Envelopment Analysis" (Charnes and others, 1978). It is designed to measure the efficiency of "decision-making units", such as schools, that produce multiple outputs but cannot use market prices to measure relative values of diverse outputs. The procedure has been applied to public schools (Bessent 1980, 1982). It could be used to infer average values of output weights that are consistent with the way schools presently operate (cf. Pratt, 1984). However, this procedure is still too new and experimental to adopt as the basis for allocating money. Further research would be useful.

One final difficulty with all these weighting procedures is that they are linear. That is, the index of excellence is simply the sum of the performance indicators, each multiplied by its weight. This implies, for example, that the value of a one-point improvement in reading does not depend on whether the school's reading scores are already extremely high or extremely low. (There is no "diminishing marginal utility.") The value of a one-point rise in reading also does not depend on whether the school currently is performing at a very high or low level in mathematics or other subjects. (There is no "diminishing marginal rate of substitution.") A measure of excellence that is simply a linear combination of performance objectives therefore lacks theoretically desirable properties.

The main advantage of combining performance measures into a single index of excellence is the convenience of a one-dimensional measure. However, if fame and fortune are tied to such an index, no one will know what is really being rewarded. Winning an award will be more like winning a lottery than earning a predictable payoff as a result of deliberate, concentrated effort. Payment by result, therefore, would make more sense if payments were tied to each performance indicator separately. The importance of various objectives is then clearly reflected in the choice of measures and the amount of money tied to each. Submerging these values in the computerized obscurity of a single index could defeat the whole purpose of an incentive plan.

Comparing Schools

Schools with high test scores or high levels of other performance indicators are likely to have students who were already performing at high levels when they entered. Absolute levels of performance are therefore inaccurate reflections of the school's own contribution or value added.

A school's own contribution has been defined as the difference between students' actual level of performance and their expected level of performance. Schools in which students do better than expected are considered "effective." One conventional way to estimate students' expected performance is based on their socioeconomic and other background characteristics, using statistical regression analysis (Garms and Smith, 1970; Klitgaard and Hall, 1975). The California Assessment Program has been reporting results of such analysis on individual schools for a decade. Another way to estimate students' expected level of performance is to employ their actual performance at an earlier point in time (McDonald and Forehand, 1972).

This second approach is similar to simply using change in performance, rather than level of performance, as the measure of a school's output. However, using change scores may not give the same results as using the difference between actual and predicted performance, where predicted performance is estimated from performance at a previous point in time (see Cronbach and Furby, 1970). In practical terms, using simple change scores is likely to give a smaller estimate of the contribution of schools whose students performed at high levels in the earlier period.[1] On the other hand, Rogosa and others (1982) have argued that simple change scores are more useful than differences between actual and predicted performance. Estimated differences between actual and predicted scores are biased and inefficient measures of true differences between actual and predicted scores if measured scores are different from true scores. Rogosa and others are also wary, on purely logical grounds, of counterfactual "predicted" scores. They argue that the difference between actual and predicted performance is "logically subordinate to" simple change in performance.

The proposed statewide accountability system in California deals with this problem by using both absolute performance levels and simple change scores to select merit schools, and dividing schools into quintiles based on students' socioeconomic level and the number of limited-English-speaking enrollees. Schools compete only against others in the same quintile.

As a means of controlling for differences in students' characteristics, grouping schools into quintiles is less precise than regression or other conventional techniques that can use predictors measured on a continuous scale. For example, suppose one predictor is the percentage of students in a school who receive aid to families with dependent children (AFDC). Suppose the boundaries between quintiles are 3.0, 5.0, 9.0, and 15.0 percent. Then a school where 8.9 percent of the students receive AFDC would compete against a school with 5.1 percent on AFDC, but not against a school with 9.1 percent on AFDC.

Another disadvantage of grouping schools is that, if several characteristics of students are to be considered, they have to be combined into some index before the schools can be categorized. Since the characteristics are not perfectly correlated with each other (if they were, there would be no point in using more than one), two schools in the same group might have very different characteristics. For instance, one might have a high proportion of limited English speakers and a low proportion on AFDC, while the other might have the oppposite. Such an index therefore conceals relevant differences. It is not at all apparent what such an index would mean, or how it should be computed.

Like a single index of school performance, a single index for grouping schools with similar students would be convenient. Its convenience, however, is outweighed by its arbitrariness and obscurity. If students' characteristics or earlier performance are to be used to predict expected levels of current performance, then conventional regression procedures are more precise. They may also be easier to understand, since the California Assessment Program has been publishing such analysis for years.

Logically prior to the question of <u>how</u> to adjust for differences among students is the question of <u>whether</u> to do it at all. Superintendent Honig's 1984 article claims that such adjustment

> "discourages excuse-making for schools with large numbers of students in lower socioeconomic groups because comparisons are made with similar schools. Comparisons would also unmask weak programs in schools with more advantaged students."

In schools at lower socioeconomic levels, the intent is to create expectations that seem realistic to teachers and administrators, so that they will strive for improvement rather than give up because their students cannot compete with those farther upscale. Parents, however, may see this as conceding defeat. They may not be pleased to hear that their school is at the top in its group if it is still well below the statewide average. Parents sometimes charge that comparing less advantaged schools only with each other actually encourages excuse-making by educators.

This is a serious dilemma. On the one hand, motivating teachers and administrators requires setting reasonable expectations. Since students' socioeconomic and linguistic background strongly influences their performance in school, the likelihood of achieving high performance is much lower in schools where students have lower socioeconomic status and less proficiency in English. In California two-thirds of the variance in school average reading scores in sixth grade, and half the variance in school average math scores, are attributable to measured difference in students' socioeconomic background and proficiency in English (Fetler and Carlson, 1984). These background factors probably would explain even more of the variance if they were measured more accurately.

On the other hand, these correlations show that schools are still transmitting inequality from one generation to the next (Bowles, 1972). By setting different expectations for different socioeconomic levels, the statewide accountability system would appear to condone, and even reinforce, that transmission of inequality. To avoid condoning this pattern, state grants to schools where students' performance exceeds expectations could be larger if students come from lower socioeconomic groups or have less proficiency in English.

Incentive Effects

In principle, awarding cash to effective schools has a lot to recommend it. Like merit pay for individual teachers, it would provide material incentives for performance--incentives that standard salary schedules for teachers do not offer. Unlike individual merit pay, however, merit grants to schools would promote teamwork among the school staff (Stern and Harter, 1981). Under present arrangements for evaluating and paying teachers, many teachers resist taking responsibility beyond their own classrooms. More collaboration among teachers on matters of curriculum, instructional technique, dealings with individual students from one year to the next, discipline, relations with parents, extracurricular activities and other matters would probably improve students' performance in many schools. If a whole school is rewarded for good performance, then teachers who are already inclined to take more schoolwide responsibility have an incentive to do so and an argument for their colleagues to do the same. (There is always the possibility of some "free-riding," however.)

One problem in implementing this idea is that the merit grants would actually be paid to local districts, not to individual schools. If merit grants are going to function as true contingent rewards, then districts must use the money in a way that the staff of the "merit school" considers beneficial to themselves. Adding to the school's budget for equipment, materials, services or supplementary personnel might seem sufficiently rewarding to motivate the staff. Outright cash bonuses could appeal to the staff even more, but would require changing collective bargaining contracts in most districts.

However, a district administration might choose not to pass the money through to the school that won it. Restoring recent cuts in district administrative staff might have higher priority. In districts with more than one school at the same level, there would also be strong pressure to use the money to

help the schools that did not win. For instance, in a district with several high schools, there may be one that has long stood above the others in academic performance. "Excellence High School," however, may not be big enough to accommodate all students in the district whose parents want them there. Now the state creates more publicity about the superiority of Excellence High compared to the other high schools in the district. This intensifies the problem for district administrators. To appease parents in the lower-scoring schools, the district may spend the prize money to benefit those schools. If this were to happen year after year, teachers at Excellence would clearly derive no additional incentive from this program. Instead, the program would have the same effect as any general grant-in-aid from the state to the district.

To achieve intended incentive effects, state legislation must direct the flow of money within districts. A simple but radical approach would be for the state to pay the awards directly to staff in winning schools. A more conventional tactic would be to include some language requiring districts to pass the money along to schools on whose account it was earned. Enforcing this requirement would involve financial audits to ensure that merit grants did not supplant regular outlays in those schools. The legislature may be loath to create a new regulatory apparatus, but it may be the only way to make the money effective.

The more the state succeeds in making the material incentive real and important, the greater the likelihood of perverse effects. There are many ways to raise test scores for a given group of students, and only some of them actually increase students' stock of knowledge. Other techniques include coaching during the test, editing students' answers, abetting the absence of slow students from testing sessions, and teaching the test. Principals can assign their best teachers to the grade levels or subjects tested (Garms, 1984). Enrollment in advanced academic subjects can be inflated by simply changing course titles. If performance objectives are defined in terms of proportions of high school students who take advanced courses or score high on college entrance tests, schools have an incentive to raise those proportions by reducing their denominator: that is, let non-college-bound students drop out.

The state conceivably can disqualify schools for outright cheating, but this will take a large number of test proctors, curriculum checkers, and other "inspectors." Moreover, the state cannot prevent schools from concentrating on measured performance objectives at the expense of unmeasured objectives (for instance, see Gramlich and Koshel, 1975). This is one reason why a state accountability system should be seen as only a step toward a more decentralized system for educational improvement.

Student Mobility

Holding individual schools accountable for the performance of their students makes most sense if each student attends only one elementary, middle, and high school. In fact, however, many students end up attending more than one school at each level because their families move, or because they transfer between public and private school. Reports by sophomores sampled in the national High School and Beyond survey imply that most students move at least once between kindergarten and graduation from high school. About one out of four students will have moved three or more times (Jones and others, 1983, pp. 8-15). This is in addition to the normal movement from elementary to middle to high school. In California, there are elementary schools where only 10 percent of the sixth graders attended the same school in first grade.

Evidence of an individual school's effects can easily wash away in this turbulent turnover of students. Measuring how an elementary school's sixth graders or a high school's twelfth graders perform, and comparing that with expectations based on performance of each school's students in earlier grades, does not give an accurate measure of value added by each school if a large proportion of students in the later grades actually attended different schools in earlier grades. Conversely, a school with excellent programs in the earlier grades would not receive its full due credit if many students then transferred to other schools. Trying to solve the problem by restricting the analysis to students who stayed in the same school would be valid only if the school's effect on stayers and leavers were the same. This is a strong assumption.

Student mobility is therefore another source of inaccuracy in measuring a school's effectiveness. Inaccuracy or arbitrariness damages incentive programs by weakening the link between what educators do and how their school is rated. That link must be strong in order for an incentive system to work. The only sure way to correct for student mobility is to construct a centralized file of individual student records, so that individual students' performance can be linked to schools they actually attend. This will require further evolution of the state's educational data base, and the eventual assignment of a number or code to each student much like a Social Security number.

The Evolving Statewide Data Base for Elementary and Secondary Education in California

The state's current data base has the following major components.

CBEDS, the California Basic Educational Data System. Every year on "information day" all teachers and administrators fill out Professional Assignment Information Forms. These provide data on classroom or administrative assignments, numbers of students enrolled in each classroom and certain characteristics of those students (but not students' names), and certain information about each teacher or administrator, including race/ethnicity, sex, type of credential, number of years as an educator, and salary. CBEDS also collects forms for each school and district, with information on classified staff, categorical programs, and number of graduates from each high school. Roughly half the districts in the state still send the information on paper forms, but the process is rapidly becoming automated, with large districts sending their data on magnetic tape and a growing number of small districts sending microcomputer diskettes. Automation may improve accuracy and also reduces redundancy, since some information does not change from year to year, so the computerized files are simply updated.

CBEDS data are assembled by the state education department, which sends printed summary reports to each district. The department also prepares special reports on request; it receives about two thousand such requests each year. Currently department staff are developing ways for local district and school site managers to use their own CBEDS data for planning and other administrative purposes.

CAP, the California Assessment Program, annually surveys students' performance on tests of basic academic skills. Currently tests are given each year in grades 3, 6, 8, and 12. The purpose is to estimate the distribution of performance at each school, not to produce a score for each student. Accordingly, students' identities are not recorded. An efficient matrix sampling procedure is used, in which several different forms of each year's test are given at each grade level. In addition to the test scores themselves, CAP collects simple measures of students' socioeconomic status, reported by the teachers for students in the earlier grades, by students themselves in the later grades. For instance, teachers are asked to rate the occupational status of each third grader's parents on

a four-point scale. While these socioeconomic measures may not be accurate for individual students, school averages have been found to produce a stable ranking of schools from one year to the next, and to predict a large proportion of the variance in schools' average test scores. Finally, CAP also collects information on instructional practices and school climate, reported by school staff at lower grades and by students at upper grades.

CAP compiles the data and sends detailed reports on each school back to the districts. Schools are told their percentile ranking within the state, and also whether they are above, below, or within a "comparison band." If a school's average score on a test is above its comparison band, the school is performing substantially better than predicted on the basis of students' characteristics.[2] The CAP report also gives each school a considerable amount of diagnostic information on the specific skill areas measured by each test.

AFDC survey. To determine eligibility for compensatory education funds, school districts are asked to report how many students in each school's attendance area are receiving Aid to Families with Dependent Children (AFDC). To do this, school officials must obtain the list of AFDC recipients from the county welfare office, then match each child with a school.

Language census. To comply with laws regarding education of students with limited proficiency in English, the state conducts an annual census in each school. This gives a count of students by primary language and grade level.

Ethnic census. Similarly, the state annually collects from each school a count of students by race or ethnicity at each grade level.

Consolidated Programs. To simplify paperwork for local districts participating in various categorical programs, the state created a "Consolidated Application" and a consolidated program review procedure. In a further effort to reduce the burden of data collection, the bureau responsible for consolidated programs created a data base that integrates information from all sources listed above. This is presently called the Consolidated Programs Description Database (CPDD). The unit of observation in CPDD is the individual school, the most convenient common denominator. CPDD uses interactive software for database management. Therefore, users can obtain specified information on specified groups of schools. CPDD has begun to be recognized as an important step in integrating educational data at the state level.

FUSE, the Follow-Up of Students and Employers. The federal

Vocational Education Act requires states that receive federal funds to collect data on students' success in the labor market after they leave vocational training. Districts representing about 25 percent of the state's vocational enrollment are sampled each year. Districts submit background data along with the name, address, and telephone number of each vocational student from the previous year. The state then sends a short questionnaire to each student. The response rate by students has been about 60 percent. The questionnaire asks about the former students' current educational status and employment. If the respondent is employed and gives permission, the state sends another short evaluative questionnaire to the former student's current employer.

FUSE is a useful source of information. It could yield additional performance indicators for high schools, especially if it were expanded to survey all districts each year, and perhaps also to include some students not enrolled in vocational classes. In addition, FUSE is important as a precedent for the maintenance of data on individual students at the state level.

Continuation high schools enroll about 10 percent of California's 16- and 17-year-old high school students. These alternative schools date back to 1919, originally providing a part-time schooling option for working students, and now serving also as a refuge for students who prefer them to the larger, regular comprehensive high schools. Despite their importance as an instrument for reducing high school dropout rates, continuation high schools were not included in the CAP until December, 1983. In addition to CAP and CBEDS data, each continuation school files a year-end report with information on enrollment, attendance, characteristics of students, graduation rates, and destinations of students who left without graduating.

Other. The state department of education has assembled data from the College Board on Scholastic Aptitude Test and achievement test performance by students from each high school. The University of California and California State University systems have also provided data on performance of students from each high school.

At the district level, the state collects financial data on revenues by source and expenditure by object. At the county level, the state collects annual data on enrollment and staff in various kinds of private schools.

Evolution of this "system" occurs through a continual process of proliferation and integration. New state policies create demands for new data. Local districts resist requests for data, and their resistance forces the state toward better integration. This interaction is evident as the new statewide accountability procedure begins to unfold.

Future Directions

It is uncertain whether the centralized accountability system as currently proposed in California will ever be fully implemented. In any event, the preceding discussion has given many reasons why it is difficult for a centralized system to create effective and appropriate incentives for school improvement. However, the state does pay most of the bill for schools in California, just as the 50 states together now pay most of the bill nationwide. To satisfy its desire for accountability, and create performance incentives that are not arbitrary or perverse, a state could implement a decentralized system.

One step toward a decentralized system of information for school improvement would be to make the state's data available to local districts and schools. With a relatively small investment in new hardware, every school in California could have a terminal providing access to CPDD, which itself could be more fully integrated with other state data sources. New commitment and reorganization within the state education department would be necessary to manage this kind of distributed data processing system. One such system already in operation is the Florida Information Resource Network, recently developed by the Florida Department of Education (1984).

Here is an example of how a comprehensive information system could assist in improving California education. Local educators could use the state's data to identify other schools that out-perform their own, but have similar characteristics. They could go beyond the "comparison band" analysis that CAP presently does for them by finding the actual names of schools that perform well on any particular measure, from specific skills in third grade reading to College Board advanced placement tests or high school retention rates -- and, among the high-performing schools, finding those with students and other characteristics similar to their own. Teachers and administrators could then make direct contact with their colleagues in the other schools. This would greatly speed diffusion of successful strategies among practitioners. Lack of mechanisms for such diffusion has greatly hindered the improvement of educational practice before now.

Another possible advance that could happen soon is the linking of local data bases for rapid transfer of data on individual students who move from one school or district to

another. Many districts and schools have already put their student test data and academic records on computers. Merging these into a regional or statewide data base would make it possible for information to follow students much more quickly when they change schools. This would facilitate communication and collaboration among the succession of teachers who become responsible for each student. It would also make it possible to monitor the subsequent performance of all students who pass through a given grade level at a given school, whether they stay in that school or not.

In these and other ways it is possible to improve the information available to teachers and local administrators in the near future. A decentralized information system can then be used to implement incentive plans negotiated locally. The state can encourage or direct local districts to institute plans. However, the state should not try to operate a centralized system. A centralized system will likely be too imprecise and arbitrary, and produce too many unwanted side effects. The momentum behind current proposals will probably produce more improvement in educational practice if it carries beyond the state capitol to the local level.

Summary of Recommendations

The proposed statewide accountability system would more likely improve school performance if it were modified in the following ways:

Include more measures of performance in basic subjects.

Include measures of performance by students in the lowest quartile.

Report (and reward) each performance indicator separately, not combined in an index of "quality."

Use conventional regression procedures to compare schools, instead of grouping them by an index of socioeconomic status and proficiency in English.

If actual performance exceeds predicted performance by the same amount in two schools, award more money to the one where students have lower socioeconomic status or less proficiency in English.

Beyond modifying the current proposal, the state should move toward decentralizing the whole accountability system.

Notes

1. Suppose the true relation between a school's earlier level of performance at time 1 and later performance at time 2 is

$$y_{2s} = a + by_{1s} + cz_s + e_s ,$$

where y_{1s} and y_{2s} are levels of performance at times 1 and 2 for school s; z_s is a set of other variables that predict performance at time 2; a, b, and c are numbers that are the same for all schools in the population; and e_s is the true difference between actual and predicted performance of school s. In the population of schools, e is independent of y_1 and z.

The true model of change is obtained by simply subtracting y_1 from both sides:

$$y_{2s} - y_{1s} = a + (b - 1)y_{1s} + cz_s + e_s.$$

But if previous performance, y_1, is not included as a predictor of change, the model would be

$$y_{2s} - y_{1s} = d + fz_s + m_s ,$$

where d and f are numbers that apply to all schools, and m_s is a new measure of the effectiveness of schools.

$$\text{Since } y_{2s} - y_{1s} = a + (b - 1)y_{1s} + cz_s + e_s$$
$$= d + fz_s + m_s ,$$

we can solve for m_s:

$$m_s = e_s + (b - 1)\, y_{1s} + (c - f)\, z_s + a - d.$$

If y_1 and y_2 are measured on the same scale, then regression toward the mean tends to give $b < 1$. If so, and if y_1 is not positively correlated with $(c - f)z$, then schools with high levels of y_1 will tend to have relatively low values of m_s compared to e_s.

This discussion ignores aggregation bias, which could be at least as serious a problem.

2. Fetler and Carlson (1984) describe the construction of comparison bands:

> "Weighted regressions of achievement scores on the three background variables were used to obtain predicted scores. Weights were the inverse of the school achievement score standard error. A linear function of the standard error was developed that defined a band of achievement for each school, symmetric around the predicted score, such that 25 percent of the schools in the state fell above the band, 25 percent fell below, and 50 percent fell within. This procedure was repeated for each content area every year." (p. 7)

Note that these comparison bands are not the same as prediction confidence intervals, since prediction confidence intervals are wider for schools where the values of predictor variables are farther from the sample means, while these comparison bands are of constant width around the regression plane.

REFERENCES

Bessent, Authella and E. Wailand Bessent. "Determining the Comparative Efficiency of Schools through Data Envelopment Analysis." Educational Administration Quarterly 16(2), Spring 1980.

Bessent, Authella and E. Wailand Bessent. "Productivity in the Houston Independent School District." Management Science 28(12), December 1982.

Brown, Byron W. and Daniel H. Saks. "The Microeconomics of Schooling." Chapter 5 in David C. Berliner (ed.): Review of Research in Education, Volume 9; American Educational Research Association, 1981.

California Department of Education. Accountability in California Public Schools: Local Reactions to a Statewide Program. Sacramento, California: Department of Education, Division of Planning, Evaluation and Research, August, 1984.

Charnes, Abraham, W. W. Cooper, and E. Rhodes. "Measuring the Efficiency of Decision Making Units." European Journal of Operational Research 2(6), November 1978.

Cronbach, Lee J. and Lita Furby. "How Should We Measure 'Change' -- or Should We?" Psychology Bulletin 74. 1970.

Davis, Otto A., Harry R. Faulk, and Holly H. Johnston. An Educational Management Information System (Draft). School of Urban and Public Affairs, Carnegie-Mellon University, February 23, 1984.

Fetler, Mark and Dale Carlson. Identification of Exemplary Schools (draft). California Department of Education, 1984.

Florida Department of Education. Florida Information Resource Network, Third Annual Report. Tallahassee, Florida, May 1984.

Garms, Walter I. Merit Schools for Florida (second draft). University of Rochester, April 1984.

Garms, Walter I. and Mark C. Smith. "Educational Need and Its Application to State School Finance." Journal of Human Resources 3, Summer 1970.

Gerritz, William. Parental Choice in Education: How Do Parents Choose Schools? Berkeley, CA: University of California, School of Education, 1984.

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

Honig, Bill. "Setting the Course for School Reform." Education Week, April 18, 1984.

Klitgaard, R. E. and G. R. Hall. "Are There Unusually Effective Schools?" Journal of Human Resources 10, Winter 1975.

Levin, Henry M. Cost-Effectiveness, A Primer. Beverly Hills, CA: Sage Publications, 1983.

McDonald, Frederick J. and Garlie A. Forehand. A Design for an Accountability System for the New York City School System. Princeton, NJ: Educational Testing Service, June 1972.

Pratt, Richard W. Estimation of Frontier Educational Production Functions with Data Envelopment Analysis. Dissertation proposal, School of Education, University of California, Berkeley; July 1984.

Rao, Vithala R. "Conjoint Measurement in Marketing Analysis." Chapter 15 in J. N. Sheth (ed.): Multivariate Methods for Market and Survey Research. Chicago: American Marketing Association.

Rogosa, David, David Brandt, and Michele Zimowski. "A Growth Curve Approach to the Measurement of Change." Psychological Bulletin 92(3): 726-748, November 1982.

Stern, David and John Harter. "Public Schools and Teachers' Unions in the Political Economy of the 1970s." Chapter 7 in Don Davies (ed.): Communities and Their Schools. New York: McGraw-Hill, 1981.

Sub-Committee on Quality Indicators. Recommendations of Criteria for Measuring Success. Ventura, CA: Ventura Unified School District; June 27, 1984.

Attachment 1

STATE STATUS AND TARGETS

I. ENROLLMENT IN SELECTED COURSES

A. Course Enrollments in Selected Courses 1982-83

Course	Number of Takers	Percent of Enrollment	Statewide Targets 1985-86	1987-88	1989-90
Calculus	12,206	4.5%	8.0%	13.0%	18.0%
Chemistry	66,600	24.6%	27.0%	32.0%	40.0%
Physics	28,067	10.4%	16.0%	21.0%	25.0%
Foreign Lang. (3rd year or more)	58,355	21.8%	25.0%	29.0%	32.0%

B. Percent of Students Meeting A-F requirements (UC Admissions Criteria) — Available Spring of 1985

C. Percent of Students Completing at Least

4 years of English	71%	
3 years of math	65%	To Be Announced
3 years of science	32%	

II. TEST DATA

A. CAP Scores

Grade level		Actual Scores 80-81	81-82	82-83	83-84	Statewide Targets 85-86	87-88	89-90
G-12	Reading	63.4	63.2	63.1	62.2	62.7	63.7	64.7
	Math	68.0	67.7	67.7	67.4	67.9	68.9	69.8
G-8	Reading	--	--	--	250	259	267	275
	Math	--	--	--	250	259	267	275
G-6	Reading	252	254	*(73.9) 258	--	(74.9) 263	(75.9) 268	(76.9) 273
	Math	253	253	(64.0) 260	--	(65.0) 264	(66.0) 269	(67.0) 273
G-3	Reading	254	258	(74.0) 263	--	(75.0) 267	(76.0) 272	(77.0) 276
	Math	254	261	(78.4) 267	--	(79.4) 272	(80.4) 278	(81.4) 283

*Percent correct equivalents are shown in parentheses.

B. SAT Scores

		Actual Scores 1982-83	Statewide Targets 1985-86	1987-88	1989-90
State Average	Verbal	421	428	436	444
	Math	474	481	489	496
Percent of High School Seniors Taking SAT		37%	41%	44%	45%
No. of Scores (per 100 seniors enrolled)					
≥450	Verbal	15.0	17.5	21.0	25.0
	Math	21.1	24.0	27.0	31.0
≥500	Verbal	9.1	9.5	10.5	15.0
	Math	15.6	16.0	16.5	20.0
≥600	Verbal	2.6	3.1	4.0	5.0
	Math	6.3	6.8	7.4	8.0

C. Advanced Placement Exam	1982-83	Statewide Targets 1985-86	1987-88	1989-90
Number of Advanced Placement examinations (per 100 seniors enrolled)	11%	13%	15%	17%
Percent of Students Passing Advanced Placement Exam	71%	72%	73%	75%

III. OTHER INDICATORS OF SUCCESS

			Statewide Targets		
A.	Writing Assignments		1985-86	1987-88	1989-90
	Percent of 12th grade students writing one or more assignments per week	22.0% (1978-79)	50%	75%	100%
	Percent of 8th grade students writing one or more assignments per week	Available Fall of 1984	50%	75%	100%
	Percent of 6th grade students writing one or more assignments per week	Available Fall of 1984	50%	75%	100%
B.	Homework		1985-86	1987-88	1989-90
	Percent of 12th grade students reporting: 1 hour of homework per day	Available Fall of 1984	66%	80%	90%
	2 hours of homework per day	"	15%	20%	25%
	Percent of 8th grade students reporting: 1 hour of homework per day	Available Fall of 1984	66%	80%	90%
	Percent of 6th grade students reporting: 1 hour of homework per day	Available Fall of 1984	66%	80%	90%
C.	Dropout Rate	1981-82	1985-86	1987-88	1989-90
	Percent of students leaving school and failing to re-enroll	Estimated 31%*	decrease to 28.5%	decrease to 26%	decrease to 23.5%
D.	Attendance	1983-84	1985-86	1987-88	1989-90
	Percent of enrolled students grades K-8 attending school on a given day	92%	93%	94%	95%
	Percent of enrolled students grades 9-12 attending school on a given day	84%	85%	87%	89%

Data-Based Accountability in Education

James W. Guthrie
Michael W. Kirst
Editors

A collection of six papers including:

The Design of School Accountability Systems
Guy Benveniste

New Directions for State Education Information Systems
Michael W. Kirst

Merit Schools for Florida: A Concept Paper
Walter I. Garms

Who Makes Up the CBEDS?
Gene Dawson

Problems in Measuring School Reform
Edward Haertel

Toward A Statewide System for Public School Accountability
David Stern