DOCUMENT RESUME

ED 336 414 TM 017 182

AUTHOR Cooley, William W.
TITLE Student Assessment in Pennsylvania. Pennsylvania Educational Policy Studies. Policy Paper Number 6.
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.; Pittsburgh Univ., Pa. School of Education.
PUB DATE 20 Dec 90
NOTE 32p.
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Accountability; Curriculum Development; Educational Assessment; Educational Change; *Educational Policy; Elementary Secondary Education; Minimum Competency Testing; Public Schools; School Districts; *School Responsibility; Standardized Tests; *State Programs; *Student Evaluation; Testing Problems; *Testing Programs
IDENTIFIERS *Pennsylvania; Testing for Essential Learning and Literacy Skills

ABSTRACT

The role of statewide testing programs and the direction Pennsylvania should take in statewide educational assessment are discussed. The major purposes proposed for statewide testing programs are: (1) informing state policy; (2) curriculum reform; and (3) accountability. The ultimate purpose of statewide testing programs is to improve student learning in the state's public schools. The state has the constitutional responsibility to provide a thorough and efficient system of public education. Results from Pennsylvania's Testing for Essential Learning and Literacy Skills (TELLS) program indicate that the present system is not adequate. It must be recognized that a test alone is not an accountability system. Student assessment should be designed so that the state and the districts are accountable for improving student educational outcomes. In designing a new state assessment system, Pennsylvania must: correct prior misuse of tests; establish a curriculum syllabus that tests must reflect; augment multiple-choice tests with other formats in order to assess a wide spectrum of desired student skills and knowledge. It is concluded that states should monitor outcomes at the district level, districts should monitor outcomes at the school level, and schools should monitor outcomes at the classroom level. Because districts differ in specific educational tasks, it is recommended that districts be held accountable for improving student performance, but not for the level of student performance. (SLD)

Pennsylvania Educational Policy Studies

PEPS is a joint effort of the U. of Pittsburgh's School of Education and the Learning Research and Development Center
This is policy paper number 6 in this series

Student Assessment in Pennsylvania

by

William W. Cooley

University of Pittsburgh

December 20, 1990

The purpose of this series of papers is to contribute to a more informed debate about critical policy issues facing Pennsylvania's public schools. This PEPS series draws upon a data base that has been established here at the University of Pittsburgh under the direction of William Cooley in cooperation with the Pennsylvania Department of Education.

Reactions can be shared:

by mail: LRDC, Pgh.,PA 15260
by PittVAX: COOLEY
by FAX: 412-624-7088
by PENN*LINK: PEPS
by phone: 412-624-7085
by BITNET:COOLEY@PITTVMS
by chat: room 743, LRDC

12/20/90

Student Assessment in Pennsylvania

William W. Cooley
Pennsylvania Educational Policy Studies
University of Pittsburgh

Background

One significant way in which Pennsylvania's laws and policies have impacted the commonwealth's public schools in the past decade has been the Testing for Essential Learning and Literacy Skills (TELLS). Begun in 1984, this program tests all third, fifth and eighth grade students in reading and mathematics. The initial purpose of TELLS was to identify students in need of remedial instruction. State funds were then distributed to districts based upon the number of students identified as needy by the TELLS tests. The extra funds were to be used to provide supplemental instruction to those students who fell below a minimum standard.

In 1984, when the State Board of Education established the Chapter 3 regulations that govern TELLS, it went back as far as 1963 for some of the necessary statutory authority. This was augmented by Act 93 of 1984 which set forth how districts would develop remedial programs and apply for state department approval so that they would be eligible for state funds for those programs. In the first testing, school year 1984-85, thirty-four percent of the public school students in

those three grades became eligible for at least one remedial program (reading and/or mathematics).

To understand how TELLS has affected what happens in classrooms it is necessary to take into account the fact that school building results began to be publicly reported. This changed the nature of the TELLS from a remedial program, designed to identify students who had fallen behind in reading and mathematics, to a school level accountability mechanism. The test then began to affect what aspects of the curriculum receives greater or less emphasis. That is not desirable if the test was not designed to reflect an ideal curriculum, and TELLS was clearly not, nor was it intended to be.

In considering the present situation, it is useful to distinguish between TELLS the testing program, and TELLS the state funded remedial program. Funds for the remedial program have been cut from the state budget for the current school year, so that even if the TELLS test is given, no extra funds would be allocated by the state for providing extra help to students who were identified as having inadequate essential skills. This seems to be the end of what was a rather poorly designed effort at compensatory education.

Meanwhile, Chapter 3, the State Board of Education Regulations which launched TELLS the testing program, is still part of the Pennsylvania Code. A subcommittee of

the Board is currently holding hearings as they consider revising these regulations. But, of course, whatever regulations they propose must be consistent with state law. One law, still extant, with explicit language regarding state testing is the paragraph on Educational Performance, Standards (24 P.S. 2-290.1, August 8, 1963).

> "...the State Board of Education...shall develop or cause to be developed an evaluation procedure designed to measure objectively the adequacy and efficiency of the education programs offered by the public schools. The evaluation procedure to be developed shall include tests measuring the achievements and performance of students pursuing all of the various subjects and courses comprising the curricula. The evaluation procedure shall be so constructed and developed as to provide each school district with relevant comparative data to enable directors and administrators to more readily appraise the educational performance and to effectuate without delay the strengthening of the district's educational program. Tests developed under the authority of this section ...shall be used for the purpose of providing a uniform evaluation of each school district."

Today, almost 30 years later, "directors and administrators" are still waiting for tests that measure "all of the various subjects and courses comprising the curricula" and "relevant comparative data" that will enable them to appraise and strengthen their educational program. What such comparative data would have to look like is one purpose of this report.

The Commonwealth's first effort at implementing this 1963 law was called Educational Quality Assessment (EQA). Launched in 1967 by the newly formed Bureau of Educational Quality Assessment, EQA was an honest effort at providing useful feedback information to districts. The 1963 law was passed in the context of a major school district reorganization act, and EQA was supposed to be a type of quality control effort. EQA became controversial because it tried to measure aspects of student attitudes and beliefs that many people felt were an invasion of privacy. Many valuable lessons were learned in that EQA experience.

This spring (1991), the Pennsylvania Department of Education plans to administer the TELLS test once again. This is because it is still required by the Pennsylvania Code. So even though there is no longer a state supported remedial program, and thus the purpose for which TELLS is a valid test has disappeared, the test will continue to be administered until the laws and

regulations are changed.

Beyond the spring of 1991 lies an important opportunity. Pennsylvania could reassume some national leadership by starting now to develop a statewide assessment system that could be a model for the rest of the country. Pennsylvania has had a long and varied experience in state assessment. It could build upon that experience, eliminating what has produced negative effects, and expanding what has been positive. This paper considers what that might look like.

Purposes of state-wide student assessment

There is considerable controversy today regarding the nature and purpose of state-wide testing programs. Unfortunately, the debates too often begin with "nature". People argue about who should be tested, or what should be tested, or how it should be tested, or when to test, or the manner in which test results should be reported. But it is "purpose" that should determine the nature of the test, so it is essential to agree on the direct purposes of state-wide assessment, or there will be endless and circular debates about its nature. The major purposes proposed here are: informing state policy, curriculum reform, and accountability.

Before turning to an examination of each of those more direct purposes, the ultimate purpose should be considered. The ultimate purpose which has guided this

draft is to improve student learning in the Commonwealth's public schools. This means that it must be possible to show how a state assessment system can accomplish that. Education takes place in schools. The people who live in those places--students, teachers, principals--are the ones who are going to carry the main burden of any state assessment system. They must be able to see how their cooperation with such an enterprise will contribute to goals which they share. This does not mean that the results of this assessment must be direct, useful feedback to students and teachers. But how the results can be an indirect benefit must be clear. This can and should be part of the design of a new state assessment system.

Informing State Policy: The most easily justified purpose of state testing programs is to inform state policy. Unfortunately, this seldom happens because state testing programs tend to be viewed as accountability mechanisms, rather than policy guiding mechanisms. One result of that accountability emphasis is that policy relevant variables are neither collected nor integrated with test results, so that educational practices and policies are not easily linked to outcomes. Thus what states tend to do is publish district or school results in the hopes that public display of low performance will "embarrass the inept into action." One problem of

course is that there are many reasons for a school's low performance, and "ineptness" of staff is only one such possibility. Also, it is not clear from the test results what a low performing school might do to improve performance, especially if ineptness happens to be the problem! The states tend to be monitoring in a way that neither the state nor the districts learn how to improve schools.

A report by Burstein, Baker, Aschbacher and Keesling (1986) has documented what is being done in existing state testing programs. Although their purpose was to explore how state test data could be used as national indicators of educational quality, it is clear from their documentation that little or no policy relevant variables are being collected along with student test scores. If anything else about the student or the student's educational program is collected, it is race, sex, and school building. Thus those are the only breakdowns of test scores that are possible. State testing programs can inform state policy deliberations if it is possible to link policy related and other explanatory factors to test scores. It is clear that responsibility for state educational policy rests with the state, and the state testing program can and should be designed in a way that can inform state policy.

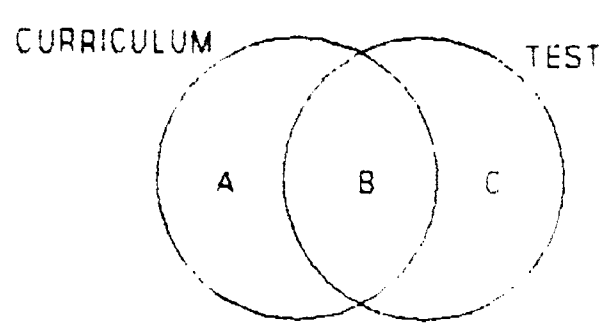
For example, a current policy debate surrounds the

relative merits of improving teacher salaries vs. having more teachers so that class size can be reduced. Statewide data, properly collected and analyzed, can inform that debate. Very little can be learned by looking at test results alone, but when integrated and analyzed with other data, much can be learned. Too often people want to move to a different testing program because the current one has not been informative, but most often, no one has done the analytic work necessary for deriving useful information from those test results. Useful, policy relevant information does not materialize by just giving tests.

Reform of the Curriculum: It is clear that state testing programs can directly influence what does or does not get emphasized in the curriculum. For example, Resnick (1987) has argued that state testing programs can have the effect of suppressing efforts to expand the teaching of higher-order thinking skills if such skills are not in the state assessment. It is critical that state policy boards and legislators understand how that works. The explanation begins with a simple model of what determines a student's performance on a test. A student's performance on a test will be a function of two major factors: (1) the student's abilities as measured by some prior test; and (2) the amount of relevant learning activity in which the student was engaged

between that prior test and the current test. Point (1) is illustrated by the fact that most of the variation in grade 5 TELLS performance, for example, is explained by how well a student did on the grade 3 TELLS. The relevance of the intervening learning activity depends upon whether the test sampled the particular skills or subject matter the given student was taught between the prior measure and the current measure, and whether the learning tasks were at an appropriate difficulty level for the student's current abilities.

Figure 1
Curriculum Overlap



Type A objectives are dropped. C are added to the curriculum

To increase test performance for a given set of students, the amount of relevant learning activity must be increased. There are at least two ways in which learning activities can be relevant to explaining performance on a test. The first way has to do with curriculum overlap, as illustrated in Figure 1. There is never a perfect fit between, for example, the mathematics objectives that are in a given districts curriculum and the mathematics objectives that are incorporated in a test like TELLS. Initially, the percent of the objectives in the test

that are type B will vary from district to district. Over time, as teachers become more and more familiar with what is tested in the TELLS test, and if the pressure to look good on TELLS increases, teachers will tend to shift their emphasis from Type A objectives to Type C objectives, resulting in higher TELLS scores, but not necessarily reflecting a general improvement in mathematics. Having the test determine the curriculum is not necessarily bad, except when the test was not designed to provide a logical, ideal, desired curriculum. Most standardized tests were designed to sample what is common across typical curricula for a particular grade. A mindless following of such tested objectives can produce a curriculum with a scope and sequence that are not optimal for facilitating student learning.

Manipulating curriculum overlap goes on all the time. (Madaus, 1988, provides an excellent summary of the research basis for that claim.) Teachers tend to want their students to look good on any externally imposed test, and they know this is one way to do it. It has people running around in circles. Some districts switching curricula rather than fighting the tests, others switching tests rather than fighting the curricula, others fighting the tests rather than switching curricula, and still others fighting the curricula to save their tests. It is essential that we

find a way out of this arbitrary, circular behavior.

Another strategy that principals and teachers use to increase the amount of learning activity relevant to a particular test is to allocate more time to those activities. This, of course, is one of the big side effects of testing programs that are used in accountability efforts. They encourage teachers and schools to emphasize what is measured in that testing program through manipulation of allocated time, without thinking about the relative value of what is being tested versus other school outcomes that are not being tested. Arbitrary allocation of more time to subject matter being tested, without a more general consideration of what is important to teach, is impossible to justify.

So if it is not sensible to arbitrarily manipulate the specifics within subject matter so as to increase the overlap between curriculum and test, or to arbitrarily shift the relative emphasis among curricula by manipulating the time allocated to different subjects, what is sensible? <u>It is essential to define a state curriculum if state tests are to be used to influence the curriculum.</u> What is needed is at least a curriculum outline or syllabus in sufficient detail to allow specification of both instruction and assessment. Also, this recommendation applies to the K-8 curriculum, where the need to improve student performance is critical, and

where there is a more common curriculum already in place.

Recognizing the importance of curriculum overlap in explaining student performance on an achievement test forces a consideration of the fundamental question of what is important to teach. To get out of the arbitrariness of changing tests to be more consistent with curriculum or curriculum to be more consistent with tests, it is essential to come to grips with what is important to teach in the first place. Curriculum theory and instructional science can contribute to this consideration, e.g., by making explicit the structure of what is to be taught, by studying how experts differ from novices, and by establishing the transfer value of what is to be taught (its utility in subsequent schooling or out in the "real world").

Recognizing the ways in which tests and curriculum interact must be sobered by another recognition. Putting something in a test does not automatically produce it. Some people talk as though all teachers are ready, willing and able to teach higher order thinking skills, but do not because it is not tested. Or that all teachers love to spend nights and weekends reading student essays, but they do not have students do a lot of writing because the tests are multiple choice. Tests can support and encourage new curriculum goals, but additional assistance to teachers may be needed to

realize improved student performance in areas where some teachers are unable to do what is necessary to achieve those goals.

Accountability: The word most frequently associated with state testing programs is accountability. Like equality, it is hard to be against accountability. What it means, or how it is achieved is another matter. To be accountable means to be responsible. For example, some people want the school principal to be accountable for the test results in his or her school. If I were a school principal, I would be willing to be accountable for the achievement of students in my school if the following were true:

1. Achievement was measured in terms of growth that occurred while the students were in my school.
2. I had adequate resources to monitor and improve the quality of teaching in my school.
3. I had adequate options for dealing with students who continually disrupt the learning of other students.
4. I had control over the instructional resources (e.g., budget, personnel selection, textbooks) available to my school.

It seems hard to expect principals to be accountable for student achievement if they are constrained by state

or district rules and regulations, or if they have no control over any resources that might be needed to do that job well. Because those four conditions tend not to be true, it is difficult to assign responsibility. Accountability systems do not work if it is too easy to blame someone else for not performing well.

One thing that is clear is that the state has the constitutional responsibility to provide a thorough and efficient system of public education. The TELLS results do indicate that the present system is not thorough, since about one in four students appear not to be mastering essential learning skills. It also appears not to be efficient, at least in the sense that student achievement results are unrelated to how much districts spend (Cooley, 1990). Thus a state testing program could be useful in holding the state accountable for its constitutional mandate. But who in the state is responsible: the Governor? the Secretary of Education? the Commissioner of Basic Education? What would happen if a state legislature passed a law holding the Governor accountable for improving student outcomes in Pennsylvania? For example, do you suppose that would change the Governor's behavior when the state's education budget gets established? States like to give tests that make district superintendents, or principals, or teachers accountable. How about one that makes the state officers

accountable for state wide improvement in educational outcomes?

The main point here is that a test alone is not an accountability system. A test, if designed within the context of a clear system of who is responsible for what, can be a useful ingredient in such a system. But there is currently no such clarity. If a state mandated test is to be part of an accountability system, who is responsible for the resulting outcomes must be defined, and the incentive systems that are inevitably imbedded in such systems must be carefully analyzed. It is recommended that student assessment be designed in a manner which makes is possible to hold the state and the districts accountable for improving the outcomes of their students. Toward that end it is necessary to define the outcomes, and thus the curriculum, for which they would be held accountable.

The fact that Amer ca has lost its competitive edge is now being blamed on the schools. For example, the Jeffersonian Compact coming out of President Bush's meeting with the Governors in Charlottesville, calls for the establishment of "clear, national performance goals, goals that will make us internationally competitive." Implementing this Compact will require "good information on the real performance of students, schools and states." The concluding paragraph is particularly noteworthy in

this discussion:

> "As elected chief executives, we expect to be held accountable for progress in meeting the new national goals and we expect to hold others accountable as well. When goals are set and strategies for achieving them are adopted, we must establish clear measures of performance and then issue annual Report Cards on the progress of students, schools, the states, and the Federal Government."

Who should be issuing whose report card is one of the big issues that has to be resolved as a state designs an accountability system. But at least we have the governors agreeing that they too can be held accountable.

Purposes to Avoid: There are some purposes for giving tests that have no place in state wide assessment. The classification of students for special programs is an example of one purpose that state testing programs should avoid. The TELLS experience illustrates why that is an undesirable practice, as explained in a previous PEPS report (Cooley, 1989). Testing for special programs is best done locally if it is to be done at all.

Providing information to parents on what their child is accomplishing in school is another purpose of assessment, but not state assessment. Student portfolios, for example, are an excellent way of

satisfying that very important need. Teachers showing parents the fruits of their child's school labors is not exactly a novel idea. What seems to be novel is the notion that such portfolios of student work can easily become part of a state wide assessment system. Much hard work needs to be done before portfolios can be part of a state assessment system that is intended to serve the purposes for which state assessment is usually done.

Different kinds of assessment

The big villain today in student assessment seems to be the "standardized" test. It has been charged with bias, irrelevance, triviality, unfairness, and all sorts of other evils. What is a standardized test? It is a test that is given in a standard manner, so that various kinds of comparisons are possible. It is not necessarily multiple choice, or limited to basic skills, or normed, or biased, or irrelevant, or unfair. The complaints of groups who want to abolish the standardized test, such as the National Center for Fair and Open Testing, tend to be concerned more about how the results are used. But it must be recognized that to achieve the purposes of state wide assessment being considered here, tests must be given so that comparisons are possible, and comparisons are only possible if the measurement procedure is defined in some standard way.

It is also important to recognize the different

types of comparisons which might be made. A norm-referenced test compares a given set of results to the results of some norming group, which is supposed to be a representative sample of the population to be compared against. A criterion-referenced test usually compares the results to some specified criterion associated with the subject matter being tested. If districts are to held accountable for improving student achievement, then comparisons over time are necessary. The recommendation here is to design a criterion-referenced assessment that allows comparisons in student performance over time. This, for example, would allow districts to determine if they are making progress toward their improvement goals.

An examination is still another type of test. An exam implies that the questions are clearly and directly related to the curriculum which the examinee has been studying. An exam is not a random sample of what students might study during fifth grade math, for example. An exam measures how well a student mastered what was studied. It is important to recognize that if a state assessment is based upon a state adopted curriculum framework (or syllabus), then it is an examination. Also important is the recognition that if examination exercises are to be used in comparisons, they must be performed and scored in standard ways.

Some skills cannot be measured in a multiple choice

format, but many can. Psychometricians have spent about 100 years understanding the properties of such tests, and to throw them out completely because they cannot test everything we want students to learn in school, or because some students do not do well on them, does not make sense. What needs to be done is correct the ways in which tests have been misused, agree on the curriculum syllabus that the tests must reflect, and augment the inexpensive but limited multiple choice formats with a variety of other types of constructed response and essay exercises, so that a wide spectrum of desirable student skills and knowledges are assessed, including writing and higher order thinking skills.

At the same time it is important to consider the notion of systemic validity that Frederikson and Collins (1989) have proposed. This calls for the development of test items, that if practiced, do not invalidate the test results. For example, drilling students on the words that happen to be used in the twelve vocabulary questions on the TELLS third grade reading test would have greatly enhanced a student's performance on the TELLS reading test, which had only 56 questions. The 12 vocabulary test items then no longer represent a random sample of the hundreds of words that a third grade student might know, but is a very biased sample of words they just happen to know. Such items are not systemically valid.

State assessment must be designed so that if students and teachers practice the types of exercises that are in the assessment, that does not invalidate the results of that assessment. Also, such practice must make sense pedagogically.

Socio-Economic Status and Achievement

This section deals with a fundamental question in assessment and accountability systems. Can and should the differences in the populations being served by a district be taken into account? In particular, should the socio-economic status (SES) differences among school districts be used in reporting assessment results? This section illustrates once again that procedures for reporting results depends upon the purpose of that assessment.

SES has clearly become a frequently used variable in education today. It became particularly prominent in the early days of Federal compensatory education funding, because funds were distributed to schools on the basis of economic need. As a result, school districts had to collect data that were descriptive of the families of their students. The most widely used SES indicators were the child's eligibility for free lunch, or whether or not the family received aid for dependent children (AFDC).

District researchers soon noticed that such SES indicators correlated very highly with standardized

achievement test results, particularly when such analyses were done at the school level. For example, in a district with many elementary schools, rank ordering the schools in terms of the proportion of children in each school that is eligible for free lunch produced about the same ordering of schools as using the proportion scoring below the bottom quartile on national achievement norms. Thus school districts suddenly had a very powerful predictor of the achievement level to be expected in each school. Because SES and its relationship to achievement causes so much confusion, it seems important to review some of the things that are known about this significant relationship.

First it is important to recognize how different levels of aggregation influence the strength of the relationship. For example, using nationally representative samples of students, family income correlates about 0.30 with achievement tests at the student level. This means that only about ten percent of the variation in individual student achievement is explained by home differences. Aggregating to the school level, the correlation is between 0.50 and 0.60 among school means nationally. If, however, the analysis is done within large urban school districts, the school level relationship is often greater than 0.90.

The high correlation between SES and achievement at

the school level is primarily due to something that statisticians call the grouping effect. This occurs when membership in the group (e.g., school) is related to either one or both of the variables being correlated. For example, the socioeconomic homogeneity of neighborhood schools produces a relationship between SES and school, and that relationship produces the larger correlation between SES and achievement at the school level than exists at the student level.

The way in which SES is measured also influences the strength of the relationship. As the indicators of SES move from measures that reflect family income to those that are more likely to directly influence the educational environment of the home (e.g., mother's education, number of books in the home, homework help), the relationship between SES and achievement increases. Unfortunately, in practice, the SES indicators used in accountability tend to be very crude measures of family income. (Cooley and Bickel, 1986, summarize these various SES-achievement relationships.)

The fact that the strength of the relationship increases as the SES measures more closely reflect those home processes that can influence student achievement is important in interpreting why the relationship exists. Students arrive at school with different school relevant abilities and motivations because of differences in what

happens in homes. Understanding the why of the relationship is important in considering the rationale for a particular application of that relationship.

One application of SES measures is in the search for explanations of why achievement in some schools is lower than in others. Having said that it must be quickly pointed out that a search for explanations is quite different than a search for excuses. If one admits that it is easier to produce higher achievement results in a school where there is strong support for high achievement in the home, then home differences must be taken into account when trying to estimate the possible influences of other ways in which the schools may differ.

Using SES in helping to sort out the relative effectiveness of different educational treatments is not the same as using SES as an excuse for not trying to raise the achievement level in a particular school. The following two sentences involve quite different uses of SES information. (1) K-8 elementary schools do not appear to be superior to K-5 schools after you take SES into account. (2) The students in this school did not do well on that achievement test, but what do you expect, given the low SES neighborhood that school serves. Low achievement is not inevitable in low SES schools. It is just that it is easier to produce higher achievement results in higher SES schools.

This latter point is important in considering another possible use of SES. Some states use SES measures in deciding where extra effort may be needed to raise achievement. School achievement levels are compared to those that would be expected (predicted), given the schools SES. If achievement is lower than expected, then special attention is given to that school to see what might be done to raise achievement. If a school's achievement is low, but SES is also low, the implication is "not too worry!" (i.e. What can you expect from such kids?). We do not believe that it is justifiable to use SES-based expectations in determining where educational opportunity may need to be improved. For example, in a targeted school improvement effort, where extra resources are provided to help improve student achievement in schools, the question of where to focus this effort would seem to be answered by where achievement is lowest, not where achievement is lower than would be expected, given SES. The justification for such extra effort derives from the need to equalize educational opportunity, and the most serious inequities are those that result from differences in home environment.

One recommendation with respect to the use of SES is to use it when seeking explanations of school factors that influence student achievement, otherwise you might

be attributing unusual success to school programmatic factors that happen to be related to home SES. The other recommendation is not to use SES as a statistical control variable when looking for low achieving situations that are to be improved through extra effort. Such extra effort is a scarce resource and should be distributed on the basis of reducing inequities in educational opportunity.

Another legitimate use of SES is to have low SES be the basis for distributing extra resources to schools. The AFDC component in the state ESBE formula is an example of that use of SES. It is quite different (and in terms of reducing inequalities in opportunity, quite justifiable) to use SES as a basis for extra resource allocation, than to use it as a way of adjusting achievement differences and assign extra resources where achievement is lower than expected. It is more difficult to produce high achievement in a low SES school, so the extra resources needed to do i .t job is justified. To illustrate, let's say that extra resources are given to a low SES school, and through that extra effort achievement is raised to the point that it is now comparable to that of higher SES schools. If achievement were the basis for distributing that extra effort, then it would be taken away from that school where it was needed (to offset differences in opportunity created by

home differences) and given to a school with higher SES but lower achievement. It seems safe to assume that when the extra effort is withdrawn, the low SES school would revert to lower achievement. It does not seem rational to establish an incentive system wherein raising the achievement level in a low SES school results in the removal of the extra support that helped to make that happen.

Recommendations

Planning must begin now if a new alternative assessment system is to be in place for spring 1992. The first thing that needs to be established is a clear purpose for this new assessment. The State Board has already begun this process in their effort to revise Chapter 3. The purposes recommended here are state and district level accountability, curriculum reform, and better informed state policy.

Planning can also begin now with a consideration of curriculum, since all three of the justifiable purposes for state wide assessment require a specification of what is important for students to learn. This specification should encompass as full a range of desired student outcomes as possible.

A state test will influence the curriculum if it is part of a district level accountability system. To assess the full breadth of student outcomes at the

district level it is neither necessary nor feasible to test every student. Experts in sampling could begin now to develop a plan that would not require the testing of every student but still allow district directors and administrators to assess their overall educational program.

The state does not have, and probably cannot have, a direct role in improving individual schools. Also, to achieve the desired breadth of outcomes in student assessment, it is not feasible to test all students in all schools. It is therefore recommended that no attempt be made to report results at the school level. Establishing a system of school level accountability can and must be the responsibility of each district. Districts must realize that if they do not establish effective school improvement procedures, they will tend not to do well in district level results.

Districts differ in the difficulty of their educational task, and this is a function of the socio-economic status of populations they serve. There is no satisfactory way of statistically adjusting for those SES differences in a state level accountability system. Therefore it is recommended that districts be held accountable for improving student performance, not level of performance. An assessment system can be designed that will achieve that objective.

Using state test results to inform state policy can be compatible with the purposes of curriculum reform and district level accountability. It is recommended that all three purposes be served by state wide assessment, and that must be taken into account when the state assessment system is designed. That is, it must be possible to link outcomes to programmatic information.

An assessment system that makes it possible to hold districts accountable for student outcomes would enable the Pennsylvania Department of Education to shift its emphasis from enforcing compliance with state rules and regulations, to an emphasis upon research and service. There are indeed some districts in the state that are in desperate need of help. A sound state assessment system could help to identify such districts, and a PDE staffed with people who know how to help, could be part of what the state does if state officials are to achieve the goals for which they are constitutionally accountable, maintaining a thorough and efficient system of public education.

These recommendations are also consistent with the notion that agencies should monitor at a level they can and should do something about. States should monitor outcomes at the district level, districts at the school level, and schools at the classroom level. Monitoring at different levels requires different kinds of information

and procedures. There is no reason why, and no clear procedures for, the state to monitor what is going on within particular public schools. That is clearly the districts' responsibility.

References

Burstein, L., Baker, E., Aschbacher, P. & Keesling, J.W. (1986). Using state test data for national indicators of education quality: a feasibility study. Los Angeles: Center for the Study of Evaluation.

Cooley, W. W. (1990). Important variations among Pennsylvania School Districts. Pittsburgh: Pennsylvania Educational Policy Studies, University of Pittsburgh.

Cooley, W. W. & Bickel, W. E. (1986). Decision Oriented Educational Research. Boston: Kluwer.

Cooley, W. W. & Bernauer, J. A. (1990). School comparisons in state wide testing programs. Pittsburgh: Pennsylvania Educational Policy Studies, University of Pittsburgh.

Frederiksen, J. R. & Collins, A. (1989). A systems approach to educational testing Educational Researcher 18 (9), 27-32.

Madaus, G. (1988). The influence of testing on the curriculum. Eighty-seventh Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press.

Resnick, L. B. (1987). Education and learning to think. Washington, D.C.: National Academy Press.