DOCUMENT RESUME

ED 342 798 TM 017 963

AUTHOR Moody, David
TITLE Strategies for Statewide Student Assessment. Policy Briefs, Number 17.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE 91
CONTRACT 400-86-0009
NOTE 5p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Basic Skills; *Educational Assessment; Educational Change; Elementary Secondary Education; Evaluation Methods; Holistic Evaluation; Psychometrics; Scoring; *Standardized Tests; *State Programs; *Student Evaluation; Testing Problems; Testing Programs; *Test Use
IDENTIFIERS Alternatives to Standardized Testing; *Authentic Assessment; *Performance Based Evaluation

ABSTRACT

Traditional standardized tests of basic skills are no longer considered meaningful by many leading authorities in educational measurement. Alternative approaches are not yet fully developed, although many efforts are being made. This paper explores the issues surrounding student assessment in the context of existing and evolving state practices, which frequently combine high-stakes evaluations with traditional multiple-choice norm-referenced examinations and negatively affect instructional quality. A new generation of alternative strategies for student evaluation is being designed to measure student performance in situations that bear an authentic relationship to real-world tasks. Authentic assessment is criterion-referenced and performance-based. It has intrinsic validity and a holistic approach. Because authentic assessments are typically much more difficult to score than traditional tests, they are expensive. The psychometric foundations of authentic assessments are still not fully developed. Vermont, Michigan, and Kentucky are leaders in the effort to use authentic assessment for statewide testing programs. The fact that many problems surround the changing nature of student assessment means that caution must be exercised in using assessment as a tool of educational reform. (SLD)

FAR WEST LABORATORY

1991

POLICY BRIEFS

NUMBER SEVENTEEN

Strategies for Statewide Student Assessment

David Moody

Introduction

Over the course of the last decade, statewide assessments of student achievement have assumed a position of prominence in the landscape of educational reform. Prior to 1980, nearly half of all states mandated no programs of this kind. Today, statewide assessments are all but ubiquitous, and the nature of the debate has shifted away from *whether* to conduct such programs to *what kind* of program to conduct.

As this *Policy Brief* explains, traditional, standardized, multiple-choice tests of basic skills are no longer considered meaningful by many leading authorities in educational measurement. Alternative approaches, on the other hand, are not yet fully developed, although innovative efforts are proliferating. The field of student assessment, in short, is undergoing a profound transformation at the very time that the demand for achievement data is growing ever more acute.

This *Brief* is designed to explore these issues in the context of existing and evolving statewide practices. First, the essential nature of the controversy associated with student assessment is explained. Then the variety of assessment strategies currently visible on the statewide level is reviewed. The *Brief* concludes with some issues policymakers need to consider as they mandate new strategies for statewide student assessment.

Far West Laboratory for Educational Research and Development serves the four-state region of Arizona, California, Nevada, and Utah, working with educators at all levels to plan and carry out school improvements. Part of our mission is to help state department staff, district superintendents, school principals, and classroom teachers keep abreast of the best current thinking and practice.

The Nature of the Controversy

The assessment of student achievement is a conflicted and sensitive area of educational policy and practice. The nature of the controversy can be traced to the multiple uses of testing data throughout the educational system.

Student testing data has in the past served two very different purposes: to guide instructional practice and to evaluate the practitioners. Teachers use test results for diagnostic purposes, to identify what students know and what they need to learn. State and local administrators use test results for accountability purposes, to identify strengths and weaknesses among programs and personnel.

Ordinarily, different kinds of tests are used for these two purposes. Teachers commonly devise their own tests, trade with other teachers, or use tests presented in classroom texts. Administrators, by contrast, tend to prefer commercially-produced instruments designed to compare the performance of large numbers of students (whole schools or school districts) with national norms.

Problems arise when the school performance measures carry high-stakes rewards and sanctions. Teachers then become acutely aware of these consequences and tend to teach to the test. This phenomenon is so well known that practitioners have coined an acronym for it: WYTIWYG — what you test is what you get. Teachers not only gear classroom activities to the evaluative assessments, but often also substitute those for their own diagnostic measures. At that point the distinction between instructional and evaluative assessment purposes begins to break down.

To tailor instruction to a standardized test may at first glance seem a desirable result; testing for essential skills, for example, should (by the logic of WYTIWYG) produce students with essential skills. Upon closer inspection, however, WYTIWYG contains a deeper, more sinister implication. It implies that the wrong kind of test may have far-reaching, negative effects on the quality of classroom instruction. In particular, a multiple-choice, "fill in the bubble" type of examination may lead to Trivial Pursuit-type instruction that produces students who can memorize well but are rarely challenged to exercise "higher-order" thinking skills: to think critically and deeply; to apply knowledge in novel situations; to integrate many discrete pieces of information; and to collaborate with others in the solution of complex problems.

Combine these two factors — high- stakes evaluative assessments that end up driving instruction and a testing instrument that reflects a narrow subset of legitimate learning objectives — and instructional quality is likely to suffer seriously. Unfortunately, statewide assessment strategies commonly meet both of these criteria.

State departments of education, for example, typically use student test scores to regulate flows of money to schools, to establish standards for graduation or for entrance into special programs, and to identify schools and districts in need of some form of external intervention. Most states make test results public, and the sheer attention focused on the comparative performance of schools and districts serves as a powerful pressure on school personnel. In addition to attaching high-stakes consequences, many statewide assessment strategies continue to rely on traditional, norm-referenced, multiple-choice exams.

Traditional and Authentic Assessment

The advent of large-scale testing programs coupled with rewards and sanctions has brought a spotlight to bear upon the deficiencies in traditional testing practices. A new generation of alternative strategies for student assessment is being designed specifically to redress these deficiencies. The cornerstone of these new strategies is *authenticity*: assessment is based on students' performance in situations that bear an authentic relationship to "real world" tasks. Writing essays, performing science experiments, participating in collaborative activities — these are now the substance of assessment practice itself, rather than the distant goal for which testing serves as a convenient substitute.

A comparison of traditional with authentic forms of assessment reveals their respective strengths and weaknesses. Strictly speaking, traditional testing practices embrace a variety of approaches to assessment. In the following discussion, however, the phrase *traditional testing* refers specifically to standardized, norm-referenced, multiple-choice exams of the kind exemplified by the Iowa Test of Basic Skills.

Authentic assessment differs from traditional testing in several respects. A first and fundamental difference pertains to the form in which the results are reported. Norm-referenced tests give only relative, comparative data: they compare the performance of a given student with the norm established by the performance of his or her peers. Such results are commonly expressed as percentiles — the percentage of students who scored better or worse than the student in question.

Authentic assessment, by contrast, expresses results with reference to actual, concrete skills that a student has or has not mastered. Such tests are said to be *criterion-referenced*. A criterion-referenced test will tell whether Jason or Jennifer can or cannot add or subtract fractions; a norm-referenced test will tell only whether they are doing better or worse than other students of their grade level or age.

In order to qualify as authentic assessment, however, criterion-referenced results represent only a first step. Another characteristic of authenticity pertains to the level and complexity of the skills under examination. A paper and pencil test of musical ability might reveal a certain, narrow subset of proficiencies — whether the student can correctly name the notes, for example. The actual performance of a composition on the piano, by contrast, calls into play technical, kinesthetic, and aesthetic abilities, in addition to the ability to decipher musical notation.

Authentic assessment, therefore, is both criterion-referenced and performance-based. In addition, it characteristically displays a quality of immediacy, vitality, and meaning to the test-taker: it has *intrinsic* validity. In most forms of traditional testing, validity is a laboriously achieved product of finding isolated, artificial tasks that correlate reasonably well with the desired skill. In authentic assessment, the test is the demonstration of the desired skill itself.

Another central feature of authentic assessment is that it is *holistic*. A traditional test of writing skills by necessity isolates a large number of discrete bits of knowledge. Direct writing assessment, however, requires the student actually to produce an essay, in which the countless bits of knowledge are blended into a single whole.

In part because they are holistic authentic assessments are typically far more difficult to score than are traditional tests. Indeed, the hallmark and principal virtue of the standardized, multiple-choice exam is ease of scoring — most are scored with great rapidity, little cost, and complete reliability by machine.

In order to score an authentic assessment, however, a team of human observers is generally required. Just as panels of experts score the performances of Olympic gymnasts, a collection of teachers is ideally available to evaluate a student's ability to conduct a debate, deliver a speech, or investigate a scientific question. Collective observation and evaluation of this kind require extensive preparation in order to coordinate scoring procedures and criteria.

While the art or science of authentic assessment is no longer in its infancy, neither is it a mature and stable discipline. Its basic weaknesses are its cost, and the incompleteness of its psychometric foundations. Yet authentic assessments evaluate a much fuller range of student abilities than is possible with a multiple-choice exam, and the testing process makes much more sense to students and teachers. Authentic assessment tasks are designed to be complex, integrated, and challenging; in this way, they mirror and support good instruction. Due to this combination of strengths and weaknesses, tradi-

tional and authentic forms of assessment will probably continue to coexist for some time to come.

The State of State Assessment

As of fall, 1991, 44 states and the District of Columbia had some form of mandated statewide testing program in place; of the remaining six states, all but Nebraska had plans in progress. Reading, mathematics, and language arts were each tested in 40 states or more. More than half of all states prescribed testing in writing, science, and social studies.

The most common usage of statewide assessments is to monitor the progress of individual students. For example, results may function as a criterion for grade promotion. Seventeen states require students to pass a statewide minimum competency exam in order to graduate; California, Delaware, and Virginia require school districts to select and implement such exams. Arkansas and Virginia employ tests as an *entrance* requirement for high school.

The next most common usage of statewide test results is for school-site accountability — both to state-level authorities and to the public at large. Results are often published, for example, in comprehensive reports of educational progress among schools and districts throughout the state.

Finally, in 20 states, funding decisions hinge upon results of student assessments. In some states, low-achieving schools receive added funds, while in others extra money is a reward to schools whose scores are high.

The majority of states continue to employ nationally normed, multiple-choice exams, such as the Iowa Test of Basic Skills, the Stanford Achievement Test, and the Comprehensive Test of Basic Skills. Because they are norm-referenced, these tests allow states to compare their students' performance with that of students throughout the nation.

Criterion-referenced tests — those whose outcomes are stated with reference to specific skills — are used in combination with norm-referenced exams in 25 states and are relied upon exclusively in 10 others. Such tests are most commonly employed in states that have articulated specific learning objectives for various ages or grade levels. Criterion-referenced results tell how many students have met those particular learning goals.

Among assessments that go beyond the multiple-choice format, writing samples are the most widespread, occurring now in some 28 states. A dozen states are moving toward performance-based assessment of mathematics achievement, and eight use such approaches for secondary-level science.

Pioneers: Vermont, Michigan and Kentucky

Several states, including Vermont, Michigan, and Kentucky, are in the process of reconstructing their strategies for student assessment. Vermont is noteworthy for the extent of its commitment to authentic assessment; Michigan for its leadership in devising low-cost, performance-based testing programs; and Kentucky for its radical revision of the educational system as a whole.

Vermont has completed the first year of a pilot program of statewide student portfolios in writing and math. The state allowed participating teachers to use their own judgment in selecting students' best and most characteristic pieces of classroom work for inclusion in portfolios. This process worked reasonably well in the assessment of student writing; but scorers of the math portfolios discovered that some forty percent of the submissions consisted of worksheets. Specification of more lifelike, performance-based math activities is high on the list of changes for this year's full-scale implementation of portfolio assessments in grades 4 and 11.

The challenges Vermont faced in shifting to authentic assessment were both greater than and different from those anticipated. Scoring, for example, proved to be surprisingly straightforward — although the math teachers preferred a quantitative scale, while the English teachers insisted on verbal descriptors. Nevertheless, all the participants continue to endorse the process. Vermont leads the nation in another respect as well: its assessment strategy is deliberately designed not to include high-stakes consequences.

Michigan is a pioneer in finding ways to undertake performance-based assessments with the minimum possible cost. The state has developed a battery of assessment programs on an incremental basis, adding one or two in different subject areas each year since 1986. To date, performance assessments have been developed in art, music, math, science, social studies, and physical education.

The key to carrying out such a program on a low-cost basis is to find people who strongly desire to participate, and to enlist their assistance at every stage of the process: development, administration, and interpretation. In Michigan, administration of the assessment procedures and interpretation of results are undertaken by graduate students in each region of the state, as well as by preservice and inservice personnel.

A courtroom challenge to Kentucky's formula for financing school districts had an unexpected outcome: the State Supreme Court ruled that the educational system as a whole, "in all its parts and parcels," was constitutionally invalid. The 1990 legislature faced the daunting task of mandating a new code for education from the ground up. In so doing, it wrote into law a dual system of statewide assessment strategies: a nationally normed, traditional test of academic skills was coupled with an extensive, to-be-developed, performance-based system.

What distinguishes Kentucky's strategy, however, is not this dual approach, but rather the extremely high stakes attached to site-level outcomes. The law mandates precise levels of improvement in student percentiles that trigger release of supplementary funds. Schools whose performance declines by set amounts incur an avalanche of penalties: parents are notified that their children are now free to transfer; all certified staff are automatically placed on probation; and a Kentucky "distinguished educator" visits the site to determine which employees, from teachers through district superintendents, shall continue to hold their jobs.

The bifurcated nature of the assessment field today is reflected in a pair of developments at the national level. On one hand, The National Assessment of Educational Progress has recently entered upon the state-by-state assessment of student achievement, using a traditional testing format. It is now possible to compare the math achievement of eighth-graders in Colorado with that in Maryland or Hawaii on a common scale.

On the other hand, in a parallel initiative from the federal level, the Office of Educational Research and Improvement has funded a new center (under the auspices of UCLA's Center for Research on Evaluation, Standards, and Student Testing) designed to help states exchange information about authentic, performance-based assessments, such as the portfolio approach under development in Vermont.

Issues to Consider

The diversity of mandated student assessment programs across the 50 states reflects competing viewpoints about appropriate ways to test students. As policymakers re-think student assessment systems, they need to consider a variety of issues:

- Testing programs are not an independent element in the educational system; on the contrary, they interact with actual instruction in significant and often unexpected ways. These must be anticipated and monitored with care.

- State demands for accountability, or "top-down" reform, could directly impede a major "bottom-up" reform strategy: the restoration of authority to teachers and principals at the school site. Top-down regulation of the system may inadvertently contribute to passivity and burnout among those charged with the actual delivery of educational services.

- The costs of state-mandated testing must be considered. What fraction of the total educational budget does student assessment warrant? In an era of scarce resources, this decision must be weighed against alternative educational needs such as funding for restructuring, professional development, or curricular innovations.

- State-level support for assessment research should be a corollary of the state-level demand for educational accountability. As the demand for assessment data continues to grow, policymakers must recognize that the field of student testing is undergoing fundamental change. In order to ensure an effective transition, more research into the statistical underpinnings and the instructional consequences of alternative assessment strategies is needed.

The assessment of student academic achievement is evidently not as straightforward a matter as the evaluation of athletic accomplishment, for example. Learning is not as susceptible to objective observation and measurement as is physical prowess, and high-stakes educational testing may interact with the very system it is attempting to measure. The prudent policymaker would do well to tread gently in this sensitive territory, and to exercise caution in seeking to use assessment as a primary tool of reform.

Briefs is published by the Far West Laboratory for Educational Research and Development. The publication is supported by federal funds from the U. S. Department of Education, Office of Educational Research and Improvement, contract no. 400-86-0009. The contents of this publication do not neccessarily reflect the views or policies of the Department of Education, nor does mention of trade names, commercial products or organizations, imply endorsement by the United States Government. Reprint rights are granted with proper credit.

Staff

Mary Amsler
Director, Policy Support Services

David Moody
Senior Research Associate

James N. Johnson
Communications Director

Fredrika Baer
Administrative Assistant

Briefs welcomes your comments and suggestions. For additional information, please contact:

James N. Johnson
Far West Laboratory
730 Harrison Street
San Francisco, CA 94107
(415) 565-3000

FAR WEST LABORATORY
FOR EDUCATIONAL RESEARCH
AND DEVELOPMENT
730 HARRISON STREET
SAN FRANCISCO, CA 94107
(415) 565-3000

OERI

OFFICE OF EDUCATIONAL
RESEARCH AND DEVELOPMENT
U. S. DEPARTMENT OF
EDUCATION