DOCUMENT RESUME

ED 116 263 95 EA 007 711

AUTHOR Mazzarella, Jo Ann
TITLE Accountability and Testing. NAESP School Leadership Digest Second Series, Number 5. ERIC/CEM Research Analysis Series, Number 20.
INSTITUTION National Association of Elementary School Principals, Washington, D.C.; Oregon Univ., Eugene. ERIC Clearinghouse on Educational Management.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE 75
CONTRACT OEC-0-8-0-080353-3514
NOTE 33p.
AVAILABLE FROM National Association of Elementary School Principals, 1801 North Moore Street, Arlington, Virginia 22209 ($2.50)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS Accountability; *Criterion Referenced Tests; *Educational Accountability; *Educational Assessment; Educational Objectives; Elementary Education; Norm Referenced Tests; Principals; Secondary Education; *Standardized Tests; State Programs

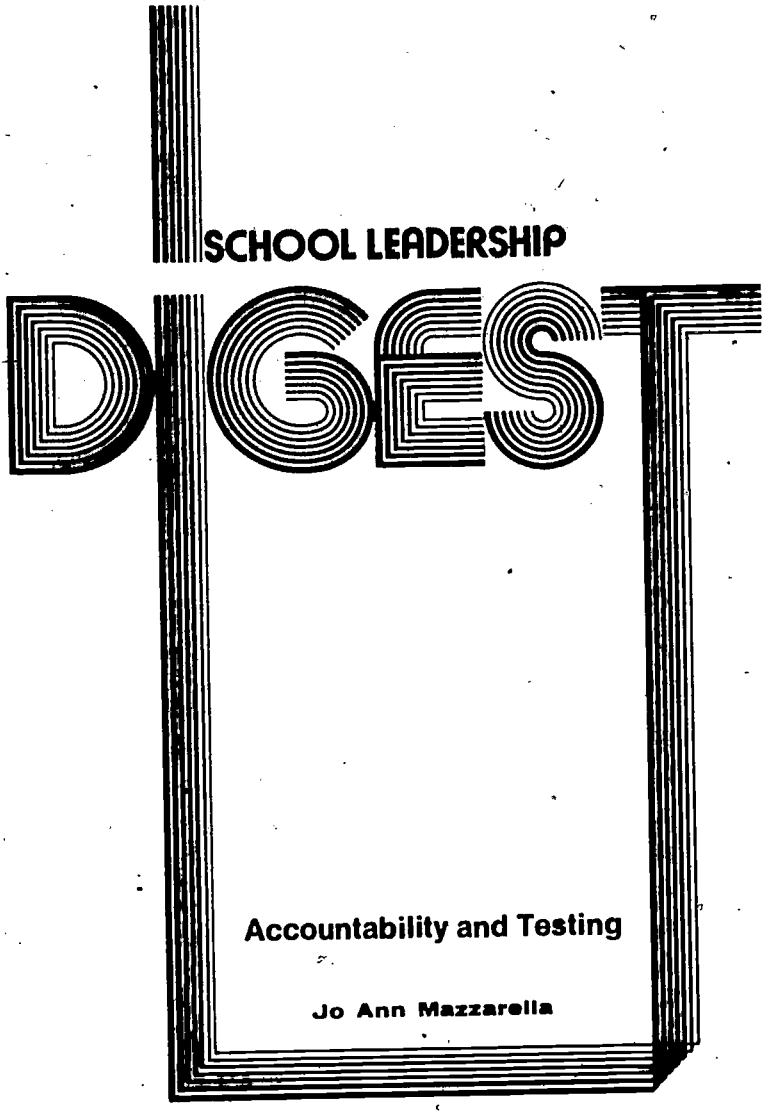
ABSTRACT

What is meant by "accountability" varies a great deal. It is not, however, the tools such as merit salary plans, voucher plans, and management techniques that are used to achieve accountability. Accountability has from its earliest days been tied to testing. In discussing testing, it is necessary to discuss the pros and cons of standardized, or norm-referenced, tests and of criterion-referenced tests; to consider the numerous against testing in general; and to examine the suggestions for alternatives to the usual methods of assessing student achievement. An administrator faced with the decision of what methods of evaluation to use for accountability will find that there are no easy answers. Most authorities on testing seem to agree that traditional standardized testing is not adequate. Yet there is still a great deal of disagreement about which other methods can do the job best. It seems clear that, for the time being at least, all the best methods of assessment and evaluation are going to involve a great deal of time and money. The method of evaluation chosen depends on one's definition of accountability, which in turn depends on one's idea of what good education is. (Author/IRT)

Second Series, Number 5

SCHOOL LEADERSHIP DIGEST

Accountability and Testing



Prepared by
ERIC Clearinghouse on
Educational Management

Published by
National Association of
Elementary School Principals

Library of Congress Catalog Number: 75-18249
NAESP School Leadership Digest Second Series, Number Five
ERIC/CEM Accession Number: EA 007 711
ERIC/CEM Research Analysis Series, Number Twenty

Printed in the United States of America, 1975
National Association of Elementary School Principals
1801 North Moore Street
Arlington, Virginia 22209

Additional copies are available from NAESP for $2.50.

CONTENTS

The Educational Resources Information Center (ERIC) is a national information system operated by the National Institute of Education. ERIC serves the educational community by disseminating educational research results and other resource information that can be used in developing more effective educational programs.

The ERIC Clearinghouse on Educational Management, one of several clearinghouses in the system, was established at the University of Oregon in 1966. The Clearinghouse and its companion units process research reports and journal articles for announcement in ERIC's index and abstract bulletins.

Research reports are announced in *Resources in Education (RIE)*, available in many libraries and by subscription for $42.70 a year from the United States Government Printing Office, Washington, D.C. 20402. Most of the documents listed in *RIE* can be purchased through the ERIC Document Reproduction Service, operated by Computer Microfilm International Corporation.

Journal articles are announced in *Current Index to Journals in Education. CIJE* is also available in many libraries and can be ordered for $50 a year from Macmillan Information, 216R Brown Street, Riverside, New Jersey 08075. Semiannual cumulations can be ordered separately.

Besides processing documents and journal articles, the Clearinghouse has another major function—information analysis and synthesis. The Clearinghouse prepares bibliographies, literature reviews, state-of-the-knowledge papers, and other interpretive research studies on topics in its educational area.

FOREWORD

Both the National Association of Elementary School Principals and the ERIC Clearinghouse on Educational Management are pleased to continue the *School Leadership Digest*, with a second series of reports designed to offer school leaders essential information on a wide range of critical concerns in education.

The *School Leadership Digest* is a series of monthly reports on top priority issues in education. At a time when decisions in education must be made on the basis of increasingly complex information, the *Digest* provides school administrators with concise, readable analyses of the most important trends in schools today, as well as points up the practical implications of major research findings.

By special cooperative arrangement, the series draws on the extensive research facilities and expertise of the ERIC Clearinghouse on Educational Management. The titles in the series were planned and developed cooperatively by both organizations. Utilizing the resources of the ERIC network, the Clearinghouse is responsible for researching the topics and preparing the copy for publication by NAESP.

The author of this report, Jo Ann Mazzarella, is employed by the Clearinghouse as a research analyst and writer.

Paul L. Houts
Director of Publications
NAESP

Stuart C. Smith
Assistant Director and Editor
ERIC/CEM

INTRODUCTION: THE PROBLEM OF DEFINITIONS

The task of implementing accountability programs can fill administrators with high enthusiasm or deep despair—enthusiasm when accountability seems to promise a truly effective way to improve the education in their schools; despair when accountability sounds like mere empty idealism, impossible to implement.

One step toward changing accountability from an ideal to a reality is choosing some method of determining whether educational goals have been reached. This usually means choosing methods to measure student performance. Which methods of assessment are best? The answer to this question depends to some extent on the meaning of *accountability*.

The term was first used in regard to education in 1969 when Leon Lessinger, as Associate Commissioner of Education, came up with an idea that seemed as reasonable as it was novel—that grant seekers should specify precisely the intended educational outcomes and costs of their projects. In addition, those receiving grants were to be audited to see whether they had indeed achieved these outcomes within the specified costs.

This rather limited concept expanded to become much broader in meaning, as is evidenced in this definition by Lessinger, Parnell, and Kaufman:

> Accountability in education means just what its dictionary definition says it means: responsibility. If you are held accountable for something, you are responsible for it, answerable to someone about it. In education, accountability means that educators of all kinds should be answerable to parents for how effectively their children are being taught and answerable to taxpayers for how usefully their money is being spent.

Accountability caught on immediately in America and has had enormous influence on American educational theory. Designs for programs can now be found in all subject areas from foreign language to vocational education, in kindergarten

through high school—and there are those who predict that accountability will someday be a part of all learning and teaching that goes on in America's schools.

Lessinger has estimated that since the appearance of his first article on accountability in 1969 at least 4,000 references dealing with accountability have been published. Since everybody is talking about accountability, it would appear that everybody is talking about the same thing, but this assumption couldn't be further from the truth. The term has a myriad of meanings, depending on who is using it.

The definition formulated by Lessinger, Parnell, and Kaufman is broad enough to include what most people mean by accountability, but many other more specific definitions have been formulated. The core of most of these definitions is how they answer the following questions: Who is accountable? Accountable to whom? Accountable for what? The table indicates some of the answers that have been offered.

Who is accountable?	*To whom?*	*For what?*
Teachers	Children and parents	Specifying costs (both past and future)
Principals	The teaching profession	Wise spending
Schools	The school board	Specifying educational goals
Superintendents	State departments of education	Achievement of goals
School boards	State or federal legislators	Students' acquisition of basic skills
Local school systems		Reporting to the public
State departments of education		Creating a suitable educational environment
Paid contractors		Behaving professionally
		Educational input or process
		Helping to create intelligent citizens

To have a complete view of the morass of meanings that surrounds accountability (and to be able to begin to extricate ourselves from that morass), we must examine yet another way of looking at the concept. Many seem to see accountability as synonymous with the methods employed to achieve it. For example, in the past when many educators spoke of accountability they meant performance contracting. Other writers and educators may actually be referring to things like merit salary programs, Jencks' voucher plan, or systems management techniques like PPBES. Our first step out of the morass is to remember that these systems are merely methods; they do not define accountability but are, as Lessinger and his associates pointed out in a 1973 volume, merely *tools* for the achievement of accountability.

The next step is to realize that the definitions reflect the differences in people's ideas about *what effective education is*; as long as educators continue to argue this issue (and let us hope they always will), they will continue to disagree about the definition of educational accountability. As Lessinger himself concluded in a published interview in April 1975, "Accountability is not defined yet." It will be up to teachers, administrators, and other educators to formulate the definition as we learn more and more about our educational responsibilities to children and how to achieve them.

STANDARDIZED TESTS: WHAT DO THEY MEASURE?

From accountability's earliest days, Lessinger and other proponents have been calling for the improvement of testing methods. In a 1970 volume he stated, "in place of relatively primitive tests now widely used, we must develop measures that are increasingly relevant and reliable." That was five years ago. The question is, Do we now have effective means of testing for accountability?

In the early days of accountability, during the heyday of performance contracting, contractors were paid almost exclusively according to students' gain scores on widely used standardized tests like the Metropolitan Achievement Test or the Iowa Test of Basic Skills. Standardized tests are used when the definition of accountability includes specifying and achieving educational goals concerning student performance in cognitive subject areas. When these tests are used, the goals have been stated in terms of comparisons; that is, students' performance is considered adequate if it compares favorably to the performance of most students in the United States.

Like the term accountability, the term *standardized testing* has many meanings. A report from the Association of California School Administrators explains, "For some, it merely means tests with norms. For others, it means the test is (1) published, (2) normed, (3) has explicit instructions for administration, and (4) was constructed to meet technical standards. Still others leave out requirement 1 or requirement 2 or both." In the pages that follow, a standardized test is a test that fulfills all four of these requirements.

Standardized tests are also called norm-referenced or psychometric tests. LeSage explains that a test becomes norm-referenced by giving it to a representative national sample of several thousand students. After these scores have been spread over a bell curve, the score of any student taking the

test can be expressed by how the raw score compares with those of the normative group. As Ebel puts it, "The aim of a norm-referenced test . . . is to indicate how the attainments of a particular pupil compare with those of his peers." This is done by percentile ranks or grade equivalents.

The content of standardized tests, according to LeSage, is determined by looking at popularly used textbooks and existing programs. As Schiller and Murdoch point out, standardized tests are designed to be a good measure of "what is generally taught."

Advantages of Standardized Tests

Probably one of the most important reasons standardized tests are used in accountability assessment is that of availability. Standardized tests have been widely used for years in schools, and it is an easy thing to apply their scores for accountability purposes. Another reason is their low cost; standardized tests require less time and money than it would cost to formulate and score a new test or new method of assessment. Schiller and Murdoch note too that standardized test scores such as grade equivalents are "easily understood by the public and by school personnel."

Another purported benefit of standardized tests is their quality. Although the validity and reliability of these tests for the purposes of accountability have come under a great deal of attack, proponents maintain that the tests are of higher validity and reliability than a "homemade" test that has not been perfected over the years by use on large numbers of students. This is probably the strongest reason that led Klitgaard, like many defenders of standardized tests, to conclude that in spite of the imperfections of standardized tests, "it is not clear what can take their place."

Averch and his colleagues note that standardized tests are most useful when the function they are to perform is that of comparing groups rather than individuals. Since they assess "what is generally taught," Ebel suggests they can help show if local programs are teaching what most people consider

important. Weber points out too that comparing scores in different curriculum areas over a state can tell a state if it has problems in one particular curriculum area. He further notes that standardized test scores can point out to a teacher or school system the existence of problems in one particular subject area. Also, the existence of national norms facilitates comparisons of schools and programs on a national level.

Disadvantages of Standardized Tests

> The use of most nationally-normed standardized tests to assess a given teacher's performance would be analogous to using a bathroom scale to determine how many stamps to put on a letter.
>
> *Alkin and Klein*

In the last several years, the use of standardized tests for accountability programs has been severely attacked. Why is this so? How can people reject tests that have been so carefully developed and normed?

One answer is that these tests have been developed primarily to *compare* students, not to assess their achievements. Those whose definition of accountability includes students' achievement of certain skills cannot measure their success with standardized tests. The tests tell nothing about what specific skills students have mastered; they merely tell how students compare to each other.

Still, it would seem that those who are interested in measuring success by how students compare to others across the nation might find standardized tests useful. However, there are other problems with the tests.

One problem is that standardized achievement tests may actually assess native abilities like reasoning ability rather than achievement. Some critics maintain that a test designed to separate good students from poor students must necessarily emphasize aptitude more than achievement. Others have pointed out that scores in almost all subject areas depend heavily on reading ability. It seems possible that we may be assessing a school's or a teacher's effectiveness by using tests

that assess things that schools and teachers are not able to teach.

Both Klein and Stake point out that standardized tests, by attempting to test "what is generally taught," may not be able to test what is being taught in a *particular* school or classroom. Porter and McDaniels have stated that "standardized tests are designed in such a way that they will not be sensitive to many unique instructional interventions." Teachers and administrators who are considering the adoption of certain standardized tests must look carefully at the amount of overlap between the test's and the school's learning objectives.

Other critics maintain that standardized tests aren't even good measures of "what is generally taught." The consensus of the articles in the July/August 1975 *National Elementary Principal (NEP)* is that current standardized achievement tests are very poor measures of student performance in all subject areas. Taylor, in that issue, indicts elementary science tests for being "incorrect, misleading, skewed in emphasis and irrelevant." To cite just one example from an issue filled with similar examples, one test asks if a damp towel placed in a warm dry room for one hour will then weigh more, less, or about the same as before. Taylor asks, "Does 'the towel' include the water it holds?" The implication is that the more deeply a student is able to analyze such questions, the more complex and difficult to answer they become.

Perhaps more importantly, Taylor, Schwartz, and other writers take issue with the values underlying standardized tests, for instance the assumption that memorization of the names of concepts is the best indication of mastery of a subject.

The critics in this issue of *NEP* maintain that reform must go beyond the development of better test items. Houts quotes Hoffmann who calls multiple choice, machine-gradable tests "insidious" because they "penalize the deep student, dampen creativity, foster intellectual dishonesty, and undermine the very foundations of education." While Hoffmann believes that these tests may successfully be used for limited

types of testing (such as a driver's exam), he maintains that they cannot successfully measure the most important products of education like creativity or profundity.

Thomas and McKinney have noted that because most standardized tests have been developed to correlate with future performance, they are not always correlated with present performance. This may seem confusing to those who thought standardized tests were meant to test current achievement. The contention is based on the fact that the validity of standardized tests is often determined by how well they "track" students; that is, how well they indicate which students will perform well in the future.

Another problem involved with using standardized tests in accountability is the inexactitude of their scoring. Krystal and Henrie note that for any one test score there is a 25 percent probability that the score is too high or too low. Lazarus points out that given the reliability range claimed by most tests, even the most reliable tests give only a very rough idea of student performance. As an example, he demonstrates that a score of 550 on a widely used test with a .90 reliability coefficient tells us, at best, only that the student probably falls somewhere *between* the fiftieth and eighty-fourth percentile.

Each score on a standardized test can be reported in three ways—as a raw score, as a percentile rank, or as a grade conversion. A raw score on a standardized test is merely a meaningless number to most people. Grade level scores are easy to understand, yet it would seem that they are too inexact to be useful. Cronbach, a longtime authority on all types of testing, states unequivocally: "grade conversions should never be used in reporting on a pupil, or a class or in research." One reason for this is that on some tests a student need answer correctly only two, three, or four more questions on the posttest than on the pretest to gain a full grade level. The same criticism can be applied to percentile ranks.

Many accountability programs make educators accountable for gains students make rather than for absolute levels of performance. Many authors have criticized the inexactitude of gain scores, which are usually obtained by subtracting the

pretest score from the posttest score. Stake has estimated that on one widely used test "on the average, a student's grade equivalent gain score will be in error 1.01 years." Weber has concluded that most standardized tests are sensitive enough only to measure gain scores over a period of at least three years. Implementation problems are obvious.

It is true, as Weber observes, that on standardized tests, group gain scores are more valid than individual scores. However, Olson points out that if the class mean on pretests and posttests is used to measure performance, high performance by a few can outweigh the poor performance of many. Thus, by directing efforts at a small number of high achievers, a teacher or program can produce impressive-looking gain scores.

Another criticism of standardized testing is the allegation that standardized tests are not valid for students who have severe learning deficiencies. Rosenshine and McGaw emphasize that achievement tests are designed for particular grade levels, and thus scores—especially gain scores—are not valid for students who begin above or below grade level.

The case against using standardized testing in accountability programs is massive—no matter which definition of accountability one chooses. Yet today these tests are still widely used methods of assessment for accountability.

By spring 1974, 30 states had enacted some form of accountability legislation. Of the 30 states that are now required by law to implement accountability programs, 18 have enacted state testing programs. Still others have enacted programs utilizing testing. Standardized testing is specified by law in at least nine of these programs.

There is very little information available about the details of state accountability programs utilizing standardized testing. Indeed, it is difficult to ascertain whether these programs are being implemented at all. It seems likely that although everybody is talking about accountability, very few people are doing anything about it, or at any rate, many who are doing something about it aren't talking.

CRITERION-REFERENCED TESTS: AN EXPENSIVE ALTERNATIVE

More and more educators, including Popham, Lessinger, and administrators of the National Assessment of Educational Progress, are turning to criterion-referenced tests for assessment purposes.

Ebel explains criterion-referenced tests this way: "The aim of [criterion-referenced tests] is to determine how many and which ones of a specified set of instructional objectives have been attained." A criterion-referenced test in mathematics, for instance, is divided into sections testing particular components of math such as adding two-digit numbers or multiplying fractions. Scores, given for each section, show whether the student has mastered each particular component. These tests are also called domain-referenced tests, mastery tests, or objectives-based tests.

Popham and Husek offer this definition: "Criterion-referenced measures are those which are used to ascertain an individual's status with respect to some criterion, i.e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced." The criterion used is often that of completing 80 percent of the items on a given section correctly.

Thus far the meaning seems fairly straightforward. However, there is a great deal of disagreement over the essential difference between standardized or norm-referenced and criterion-referenced tests. Shami and his colleagues point out that criterion-referenced tests can also be standardized (administered according to standard explicit instructions) and normed (their scores can be compared to a normative group).

Rather than delve into the intricacies of a definitional problem that may be mostly semantic, we will make the same distinction Averch and his colleagues have made: a

norm-referenced test compares a student's accomplishment with that of others; a *criterion-referenced* test indicates whether a student has accomplished certain skills.

Advantages of Criterion-Referenced Tests

It seems clear that using criterion-referenced tests for accountability avoids many of the problems encountered when using norm-referenced tests. They are endorsed most strongly by those whose definition of accountability includes the precise stipulation and measurement of educational goals.

In a 1969 article Popham states, "High quality instructional planning requires the explication of instructional intents in terms of measurable learner behaviors." Criterion-referenced tests are better than standardized tests at measuring such behavior.

Advocates of increased community involvement in education often prefer criterion-referenced tests because, unlike standardized tests, they can be locally developed to reflect local educational goals or objectives. If teachers, schools, or state boards of education develop their own tests, they will also be assured that the tests measure their unique textbooks or programs. "The criterion-referenced test," Knipe and Krahmer note, "is the only type of test that a school district can use to determine if it is working toward its curriculum goals."

Because criterion-referenced tests measure specific skills or knowledge, they are designed to do more than just measure correlates of learning or compare students. Criterion-referenced tests also have much more exact methods of scoring than the percentile or grade conversion techniques used on norm-referenced tests. Since students' scores merely indicate what learning objectives have been mastered, it is easy to calculate progress over time.

Another advantage of criterion-referenced tests is that they can be used for individualized instruction of students at many different levels. A teacher can decide which sections or items of a test he or she wants a student to complete according to the student's own level.

Disadvantages of Criterion-Referenced Tests

Although there are several ways that criterion-referenced tests fulfill the needs of accountability programs better than standardized tests, they present problems of their own. One is cost—both in time and money. Since criterion-referenced tests are useless if they do not reflect the particular educational goals being set, many educators are finding it necessary to develop their own tests. Some seem to be having success, but for most the task seems gargantuan.

Morrissett puts it: "The production of valid well-structured hierarchies of objectives and test items is not a task that can be undertaken by a teacher meeting five classes a day, nor by a Thursday afternoon curriculum committee." He points out that the National Assessment of Educational Progress spends $5 million per year developing items for use in just two or three subject areas. Ebel maintains that there are few people who have backgrounds that qualify them to develop valid, reliable criterion-referenced tests.

A possible solution to this problem is for schools to choose instructional objectives and tests that have been developed by private firms or state departments of education. In this case, a school may choose what it feels is a good test or selection of test items and then design curricula to fit the items. Popham in 1973 recommended the Los Angeles-based, nonprofit Instructional Objectives Exchange for such items. Although several other firms and state departments of education are moving in this direction, it is not clear if good "item banks" exist yet. In a paper published in 1974 on reading tests, Hogan stated unequivocally: "well-developed criterion-referenced tests are simply not available today."

Another problem arises from developing tests locally or even on a statewide basis. Those developing tests may shy away from setting high goals that seem "unrealistic" compared to past achievement. A related problem is raised by Krystal and Henrie: What happens if local special interest groups gain too much control over formulating learning objectives? Gubser gives an account of a new teacher

recertification program in Arizona that depends heavily on students' answers on tests that many feel are based on particular political beliefs and ideologies. A question that must be dealt with when developing criterion-referenced tests is, Are local goals always better than more widely held goals?

An additional problem sometimes found with criterion-referenced tests is that they usually cannot be used to compare students. Unless the tests are normed, the scores of these tests, like the raw score on a standardized test, do not tell anything about where a student stands nationally. For this reason, some suggest the development of criterion-referenced tests or, at least, criterion-referenced items that have also been nationally normed. Grady recommends merely using *both* types of tests.

Both Ebel and Haggart have noted another problem with criterion-referenced tests. As Ebel states, "Emphasis on discrete specifics may lead to neglect of the integration of ideas that gives unity and solidarity to a subject." Perhaps criterion-referenced tests will encourage students to collect differentiated skills or small bits of knowledge at the sacrifice of understanding underlying concepts or ideas.

In fact, many theorists, including Averch and his colleagues, have voiced the fear that the most important goals of education may be too broad and complex to test with criterion-referenced tests. Combs maintains that one of our most important educational goals is to create intelligent citizens who are "creative, flexible, open to experience, responsible to themselves and others and guided by positive goals and purposes." He notes, however, that these types of goals are "at odds with the specificity and precision demanded by most persons operating in the behavioral-objective performance-based criteria persuasion."

Combs further criticizes the behavioral objectives approach on which criterion-referenced tests are based for being a "closed system of thinking" because it allows only for planned outcomes. It would be tragic indeed if schools restricted themselves to teaching only those things that can be measured by a criterion-referenced test.

State Testing Programs

At least 13 states now use criterion-referenced tests in their statewide assessment programs, and there are indications that more may soon follow suit. Accountability programs using this type of testing are somewhat better reported than programs using standardized tests. Three of the most widely publicized programs are in Florida, California, and Michigan.

The Florida program, utilizing both criterion-referenced and norm-referenced tests, is based on Florida's 1971 Educational Accountability Act. The criterion-referenced component of the testing has thus far been devised by Florida reading specialists and teachers who chose performance objectives from a catalog provided by the Center for the Study of Evaluation at the University of California at Los Angeles. The program, projected through 1978, includes plans to measure student performance in such diverse areas as mental health and aesthetic appreciation as well as communication and learning skills.

The California program is based on the 1972 Stull Act, which requires each teacher to develop pupil performance objectives and criterion-referenced tests as a basis for evaluation of his or her work. In 1972-73 the San Diego Unified School District responded to the act with a plan prepared by teachers and principals for teacher evaluation based on student performance on certain learning objectives. Although a few other similar kinds of programs have been instituted in California, it is unclear what kinds of programs most schools in the state are instituting, or indeed if they *are* instituting serious programs. A paper from the Institute for the Development of Educational Activities notes, regarding California, that "teachers and administrators consider that state's accountability program 'a paper tiger'."

The Michigan program, begun in 1970, is one of the pioneer state accountability programs. It originally utilized norm-referenced tests, but after two years replaced them with criterion-referenced tests developed by the state board of education, teachers, and administrators. At present the

program measures performance only in reading and math, but plans are being made for testing in other areas. In the future, the state plans to avoid spending the millions of dollars necessary to test all students by testing only a representative sample of students on most objectives. A 1974 National Education Association-sponsored evaluation of this program severely criticized it for using performance objectives that purportedly were not field-tested or validated and that penalized minority students. The NEA committee recommended the use of local rather than statewide objectives.

WHY TEST AT ALL?

Although some form of testing student performance is a component of almost all accountability programs, many writers suggest that all forms of testing—whether standardized *or* criterion-referenced—present more problems than they solve.

Poor Measures of Good Education

One argument against using test scores as major criteria in accountability programs is raised by Soar. He maintains that it makes little sense to make teachers responsible for students' test scores when there is no research to indicate that there is *any* correlation between teaching and test scores.

Another often-cited argument is that tests now currently available are culturally biased against minority students who frequently have a different language, different experiences, and different ways of looking at the world than do the majority of students. Similarly, others contend that students' scores indicate less about the effectiveness of teaching and programs than about the socioeconomic backgrounds of the students.

At bottom, this is not only an argument against the validity of testing techniques but also an argument against making teachers accountable for the academic performance of deprived students. The truth is that we know very little about how to raise the achievement rates of these students. How can we hold teachers accountable for doing what no one yet knows how to do?

Another problem involved with using any type of gain scores as the main method of measuring whether educational goals have been met is the regression effect. The regression effect means that no matter what kind of test one uses, students who score *high* on a pretest will tend to score *lower* on the posttest, and students who score *low* on a pretest will

tend to score *higher* on the posttest. The existence of this effect is not an argument against testing per se. It does mean that gain scores may be invalid unless they can be compared to those of a control group, and such comparison is often difficult and costly.

Adverse Effects of Testing

Critics warn that we must be very careful that achievement testing programs don't put so much pressure on students that there is a sacrifice of academic honesty. If educational excellence is measured *only* by students' scores on tests, both students and teachers may be tempted to cheat. Such an outcome may be especially likely if teachers and students are asked to produce more than they really can. Teachers may coach, encourage, or hurry students during a test or even go so far as to improperly score tests if under pressure.

Another way that teachers may react to extreme pressure is by "teaching to the test." This means having students memorize the correct responses to the specific items on the test. Many have been critical of criterion-referenced tests for being easy to "teach to," partly because their items test mastery of specific performance objectives rather than broad general concepts and partly because teachers themselves often have a hand in making up test items. Of course, it is also possible to teach to a norm-referenced test if the teacher is able to obtain copies of the test before it is given.

Glass and Wildavsky suggest that an answer to some of these problems is for an outside independent auditing agency to administer tests and see that they are fairly conducted. This policing of tests, however, does not lessen the extreme pressures that make teachers and students desperate enough to attempt cheating in the first place.

Noncognitive Subject Areas

> The outcomes-oriented educator cleaves exclusively to objectives amenable to measurement.
>
> *Popham* 1969

> Measuring what we know how to measure is no substitute for measuring what we *need* to measure.
>
> *Combs*

Many advocates of current testing procedures are nevertheless quick to admit that there are important educational goals that, as yet, cannot be measured by *any* tests. Ebel mentions that we are not currently able to test "interests, values, aspirations, attitudes or self-concepts." Lessinger, in April 1975, noted that our tests cannot assess things like "insightful appreciation, understandings and flashes of insight." Soar notes that student characteristics like complex problem-solving ability or responsible citizenship behavior have growth rates that are so slow as not to be measurable.

Some accountability programs demand the specification of educational outcomes that can be measured very precisely; yet, some critics maintain that by concentrating on measurable outcomes we may be slighting the outcomes that are most important. Combs holds that our educational efforts ought to be toward creating people who exhibit "intelligent behavior, intelligent problem-solving, and good judgment." He also holds that it is important for students to discover the "personal meaning" of the knowledge they are learning. These things are not measured by either criterion-referenced or standardized tests. What this point of view suggests is that basing accountability programs solely on outcomes that are "testable" may cause educators to lose sight of the most important educational goals.

But if we eliminate tests, how can we determine if educational goals are achieved? Must we then discard the whole concept of accountability as impossible to implement?

Alternatives to Testing

Some educators have pointed out that traditional forms of testing are not the only way to evaluate school or teacher effectiveness. In spite of his affinity for criterion-referenced tests, Lessinger, in a 1970 issue of *Educational Technology*, recommends that accountability can make use of "a variety of modes of attaining evidence. One thinks immediately of

hearings of juries or expert witnesses, of certified auditors, of petitions or the like. Education can make use of all these modes and can use such means of acquiring evidence as videotape and pupil performance in simulated real-life situations, to mention a few."

Perrone reports that since 1972 the North Dakota Study Group on Evaluation has been examining current methods of assessment. Members of this group are moving toward the use of observation and daily record-keeping by teachers and students as good ways of measuring the important goals that tests cannot measure. A group at Educational Testing Service in Princeton is studying similar measures.

In 1974 Hawes quoted a superintendent using such forms of evaluation, "We don't feel there's a testing program out today that measures what we believe is important to evaluate. As examples, we want to appraise students' attitudes toward learning and using what they've learned. And we want to assess the more creative aspects of the student's ability."

Hoffmann, in an interview with Houts, calls for a return to the more individualized subjective forms of evaluation used before machine-graded tests. He suggests the development of testing in which concern is not just with the correct answer but with "the reasoning process used to arrive at the answer."

Some authors have suggested making educators accountable for the process that occurs in the classroom rather than the product. In a 1974 article Aldrich recommends that schools be held "responsible for the environments which they create and foster for children." Instead of testing students, the schools might evaluate things like materials and activities available, arrangement of time and space, and teaching skills.

Others have suggested making teachers accountable only for "behaving professionally." Stocker notes the National Education Association recommendation that teachers be evaluated on responsibilities like "adequate academic preparation" and "knowledge of and concern for students." It is unclear what methods would be used for such evaluation, but they would not be concerned with student performance.

The problem of developing alternatives to standardized

testing is being tackled by representatives to a conference on standardized testing convened by the National Association of Elementary School Principals and the North Dakota Study Group on Evaluation in November 1975. Twenty-five leading national education associations, government agencies, and education groups called for investigation into the uses and impact of standardized tests in the schools and for the development of more fair and effective means of assessment. The group plans to meet in the spring of 1976 to discuss findings and further recommendations.

Most alternative forms of evaluation have had so little application that it is hard to weigh their strengths and weaknesses. At the moment they seem to hold a great deal of promise, but it remains to be seen how much time and money they will cost.

There are, however, a few schools that are using alternative forms of evaluation. These programs seem to stress the definition of accountability that includes reporting to the public on broad educational goals. The programs are not restricted to narrowly prescribed educational outcomes but instead concentrate on accurately reporting all kinds of student achievement.

In a 1974 article Hawes tells the story of Devil's Lake, North Dakota, a 2,000-student system that evaluates student performance through daily teacher observation and note-taking, samples of students' work, and teacher inventories of children's skills and attitudes. This evaluation is reported to parents by means of personal interviews with the teacher.

According to Aldrich's 1975 report, the Marcy School in Minneapolis utilizes an "internal evaluator" who evaluates the total learning environment by observation of children in the classroom. This technique is useful for those whose definition of accountability includes making educators accountable for "process," that is, what goes on in the classroom.

The Prospect School in Vermont includes student journals of their daily activities as part of their assessment program. Carini, a staff member of the school, writes that this technique makes it possible "to report precisely to parents and others on growth of individual students."

CONCLUSION

An administrator faced with the decision of what method of evaluation to use for accountability will find that there are no easy answers. Most authorities on testing seem to agree that traditional standardized testing is not adequate. Yet there is still a great deal of disagreement about which other methods can do the job the best. It seems clear that, for the time being at least, all the best methods of assessment and evaluation are going to involve a great deal of time and money.

Administrators whose definition of accountability includes the stipulation and achievement of precise learning objectives will no doubt choose to assess student performance with criterion-referenced tests. Those concerned chiefly with assessing achievement of the broadest educational goals or with reporting the educational processes that occur in the classroom will experiment with the alternative forms of evaluation now being developed.

The method of evaluation chosen depends on one's definition of accountability, and this, as we have said, depends on how one answers the question, What is good education? Each educator must ultimately find his or her own answer to this question.

BIBLIOGRAPHY

Many of the items listed in this bibliography are indexed in ERIC's monthly catalogs *Resources in Education (RIE)* and *Current Index to Journals in Education (CIJE)*. Reports in *RIE* are indicated by an "ED" number; journal articles in *CIJE* are indicated by an "EJ" number.

Copies of most ERIC reports, but not journal articles, can be ordered from the ERIC Document Reproduction Service. If a report is available from EDRS, its order number and prices are given. When ordering, please specify the "ED" number. Unless otherwise noted, reports are available in both microfiche (MF) and paper copy (HC). Please include check or money order payable to EDRS.

Postage must be added to the cost of all orders. Rates are as follows. Microfiche: $0.21 for up to 60 fiche and $0.09 for each additional 60. Paper copy: $0.21 for first 60 pages and $0.09 for each additional 60.

Address requests to EDRS, P.O. Box 190, Arlington, Virginia 22210.

Aldrich, Ruth Anne. "Innovative Evaluation of Education." *Theory into Practice*, 13, 1 (February 1974), pp. 1-4. EJ 095 538.

Aldrich, Ruth Anne. "Marcy Open School: Feeding Back to Decision-Makers." In *Testing and Evaluation: New Views*, pp. 49-52. Washington, D.C.: Association for Childhood Education International, 1975.

Alkin, Marvin C., and Klein, Stephen P. "Evaluating Teachers for Outcome Accountability." Los Angeles: University of California at Los Angeles, 1972. *UCLA Evaluation Comment*, 3, 3 (May 1972), pp. 5-11. 7 pages. (Complete document, 11 pages, available as ED 068 495 MF $0.76 HC $1.58.)

Association of California School Administrators. "The Nature and Uses of Criterion-Referenced and Norm-Referenced Achievement Tests." *Special Report*, 4, 3 [1975]. 8 pages.

Averch, Harvey A.; Carrol, Stephen J.; Donaldson, Theodore S.; Kiesling, Herbert J.; and Pincus, John. *How Effective Is Schooling? A Critical Review of Research.* Englewood Cliffs, New Jersey: Educational Technology Publications, 1974. 258 pages.

Carini, Patricia F. "The Prospect School: Taking Account of Process." In *Testing and Evaluation: New Views*, pp. 43-48. Washington, D.C.: Association for Childhood Education International, 1975.

Combs, Arthur W. *Educational Accountability: Beyond Behavioral Objectives.* Washington, D.C.: Association for Supervision and Curriculum Development, 1972. 40 pages.

Cronbach, L. J. *Essentials of Psychological Testing.* New York: Harper and Row, 1970.

Ebel, Robert L. *State Testing Programs: Status, Problems, and Prospects. TM Report 40.* Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1974. 6 pages. ED 099 429 MF $0.76 HC $1.58.

Educational Testing Service. *State Educational Assessment Programs. 1973 Revision.* Princeton, New Jersey: 1973. 104 pages. ED 080 582 MF $0.76 HC $5.70.

Flanigan, George E. "The Stull Bill—Two Years Later." *UCLA Educator,* 16, 2 (Spring 1974), pp. 8-9. EJ 105 279.

Glass, Gene V. "The Many Faces of 'Educational Accountability'." *Phi Delta Kappan,* 53, 10 (June 1972), pp. 636-639. EJ 059 809.

Grady, Michael J., Jr. *Using Educationdl Indicators for Program Accountability. Cooperative Accountabilities Project Bulletin No. 5139.* Denver: Cooperative Accountability Project, Colorado State Department of Education, 1974. 60 pages. ED 096 740 MF $0.76 HC $3.32. (Also available for single copy Colorado requestors from Cooperative Accountability Project, Colorado Department of Education, 1362 Lincoln Street, Denver, Colorado 80203. All others: SEAR, Wisconsin Department of Public Instruction, 126 Langdon Street, Madison, Wisconsin 53702.)

Gubser, M. M. " 'Accountability' as a Smoke Screen for Political Indoctrination in Arizona." *Phi Delta Kappan,* 55, 1 (September 1973), pp. 64-65. EJ 084 434.

Haggart, Sue A., and others. *A Guide to Educational Performance Contracting. Technical Appendix.* Santa Monica, California: Rand Corporation, 1972. 148 pages. ED 060 392 MF $0.76 HC $6.97. (Also available from Publications Department, Rand Corporation, 1700 Main Street, Santa Monica, California 90406. Pub. No. R955/2-HEW. $7.00.)

Hawes, Gene R. "Managing Open Education: Testing, Evaluation and Accountability. Special Report." *Nation's Schools,* 93, 6 (June 1974), pp. 33-47. EJ 097 879.

Hawthorne, Phyllis. *Legislation by the States: Accountability and Assessment in Education. Revised. Report No. 2. Bulletin No. 3100.* Denver; and Madison: Cooperative Accountability Project, Colorado State Department of Education; and Division for Management and Planning Services, Wisconsin State Department of

Public Instruction, 1974. 109 pages. ED 098 681 MF $0.76 HC $5.70. (Also available from SEAR, Wisconsin Department of Public Instruction, 126 Langdon Street, Madison, Wisconsin 53702.)

Hogan, Thomas P. "Reading Tests and Performance Contracting." In *Measuring. Reading Performance*, edited by William E. Blanton and others, pp. 51-65. Newark, Delaware: International Reading Association, 1974. Complete document, 76 pages, available as ED 094 358 MF $0.76 HC $4.43. (Also available from International Reading Association, 800 Barksdale Road, Newark, Delaware 19711. Stock No. 718, $3.50 nonmember, $2.50 member.)

House, Ernest R.; Rivers, Wendell; and Stufflebeam, Daniel L. "An Assessment of the Michigan Accountability System." *Phi Delta Kappan*, 55, 10 (June 1974), pp. 663-669. EJ 099 423.

Houts, Paul L. "A Conversation with Banesh Hoffmann." *The National Elementary Principal*, 54, 6 (July/August 1975), pp. 30-39.

Impara, James C. "A System of Educational Assessment in the State of Florida." Paper presented at AERA annual meeting, Chicago, April 1972. 8 pages. ED 063 335 MF $0.76 HC $1.58.

Institute for the Development of Educational Activities. *Assessment and Accountability in Education: Threat or Promise? An Occasional Paper.* Dayton, Ohio: 1974. 24 pages. ED 098 670 MF $0.76 HC $1.58. (Also available from I/D/E/A, P.O. Box 628, Dayton, Ohio 45419. $2.00, payment must accompany order.)

Klein, Stephen P. *The Uses and Limitations of Standardized Tests in Meeting the Demands for Accountability.* Los Angeles: Center for the Study of Evaluation, University of California at Los Angeles, 1971. *UCLA Evaluation Comment*, 2, 4 (January 1971). 20 pages. ED 053 175 MF $0.76 HC $1.58.

Klitgaard, Robert E. *Achievement Scores and Educational Objectives.* Santa Monica, California: Rand Corporation, 1974. 73 pages. ED 093 989 MF $0.76 HC $3.32.

Knipe, Walter H., and Krahmer, Edward F. "An Application of Criterion Referenced Testing." Paper presented at AERA annual meeting, New Orleans, February 1973. 19 pages. ED 074 154 MF $0.76 HC $1.58.

Krystal, Sheila, and Henrie, Samuel. *Educational Accountability and Evaluation. PREP-35.* Washington, D.C.: National Center for Educational Communication (DHEW/OE), 1972. 56 pages. ED 067 514 MF $0.76 HC $3.32.

Lazarus, Mitchell. "Coming to Terms with Testing." *The National Elementary Principal*, 54, 6 (July/August 1975), pp. 24-29.

[LeSage, William]. "Standardized Tests: What Are They? How Are They Made? How Are They Scored?" *Instructor*, 82, 7 (March 1973), pp. 45-52. EJ 073 953.

Lessinger, Leon M. "Accountability and Curricular Reform." *Educational Technology*, 10, 5 (May 1970), pp. 56-57.

Lessinger, Leon M. *Every Kid a Winner*. New York: Simon and Schuster, 1970. 231 pages.

Lessinger, Leon M., and Savage, William. "Accountability in Education. Part I." An interview in *The University of South Carolina Education Report*, 17, 5 (March 1975), pp. 1, 4.

Lessinger, Leon M., and Savage, William. "Accountability in Education. Part II." An interview in *The University of South Carolina Education Report*, 17, 6 (April 1975), pp. 1-2.

Lessinger, Leon M.; Parnell, Dale; and Kaufman, Roger. "Learning." Volume I of *Accountability: Policies and Procedures*. (A Series of 4 volumes.) [New London, Connecticut]: Croft Educational Services, 1971.

Lessinger, Leon M., and others. *Accountability: Systems Planning in Education*. Homewood, Illinois: E.T.C. Publications, 1973.

Morrissett, Irving. "Accountability, Needs Assessment, and Social Studies." *Social Education*, 37, 4 (April 1973), pp. 271-279. EJ 075 419.

National Association of Elementary School Principals. "Major Educational Groups Join Forces to Probe Standardized Testing." News Release. Arlington, Virginia: November 14, 1975.

Olson, Arthur V., and Richardson, Joe A. *Accountability: Curricular Applications*. Scranton, Pennsylvania: Intext Educational Publishers, 1972. 224 pages.

Perrone, Vito. *Introduction to Testing and Evaluation: New Views*. Washington, D.C.: Association for Childhood Education International, 1975.

Popham, W. James. "Focus on Outcomes: A Guiding Theme of ES '70 Schools." *Phi Delta Kappan*, 51, 4 (December 1969), pp. 208-210. EJ 011 967.

Popham, W. James. "Instructional Objectives 1960-1970." *Improving Human Performance*, 2, 3 (Fall 1973), pp. 191-198. EJ 089 804.

Popham, W. James, and Husek, T. R. "Implications of Criterion-Referenced Measurement." *Journal of Educational Measurement*, 6, 1 (Spring 1969), pp. 1-9. EJ 006 705.

Porter, Andrew C., and McDaniels, Garry L. "A Reassessment of the Problems in Estimating School Effects." Paper presented at

American Association for the Advancement of Science meeting, Washington, D.C., March 1974. 38 pages. ED 091 292 MF $0.76 HC $1.95.

Rosenshine, Barak, and McGaw, Barry. "Issues in Assessing Teacher Accountability in Public Education." *Phi Delta Kappan*, 53, 10 (June 1972), pp. 640-643. EJ 059 810.

Schiller, Jeffry, and Murdoch, Ellen Press. "Implications of Using Standardized Tests in Performance Contracting." In *An Experiment in Performance Contracting*, pp. 51-91. Washington, D.C.: Office of Planning, Research, and Evaluation; Office of Economic Opportunity, 1972. Complete document, 236 pages, available as ED 064 782 MF $0.76 HC $12.05.

Schwartz, Judah L. "Math Tests." *The National Elementary Principal*, 54, 6 (July/August 1975), pp. 67-70.

Shami, Mohammed A. A.; Herskowitz, Martin; and Shami, Kalida K. "Dimensions of Accountability." *NASSP Bulletin*, 58, 383 (September 1974), pp. 1-12. EJ 101 986.

Soar, Robert S. "Accountability: Problems and Possibilities. Problems in Accountability and the Measurement of Pupils." Paper presented at AERA annual meeting, New Orleans, February 1973. 7 pages. ED 077 949 MF $0.76 HC $1.58.

Stake, Robert E. *Measuring What Learners Learn (with a Special Look at Performance Contracting)*. Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, University of Illinois, 1971. 41 pages. ED 052 234 MF $0.76 HC $1.95.

Stocker, Joseph. "Accountability and the Classroom Teacher." *Today's Education*, 60, 3 (March 1971), pp. 41-56. EJ 032 780.

Taylor, Edwin F. "Science Tests." *The National Elementary Principal*, 54, 6 (July/August 1975), pp. 72-80.

Thomas, Thomas C., and McKinney, Dorothy. *Accountability in Education. A Research Memorandum*. Menlo Park, California: Educational Policy Research Center, Stanford Research Institute, 1972. 80 pages. ED 061 620 MF $0.76 HC $4.43.

Weber, George. *Uses and Abuses of Standardized Testing in the Schools. Occasional Papers, No. 22*. Washington, D.C.: Council for Basic Education, 1974. 42 pages. ED 094 098 MF $0.76 HC $1.95. (Also available from Council for Basic Education, 725 15th Street, N.W., Washington, D.C. 20005. $0.50.)

Wildavsky, Aaron. "A Program of Accountability for Elementary Schools." *Phi Delta Kappan*, 52, 4 (December 1970), pp. 212-216. EJ 029 300.