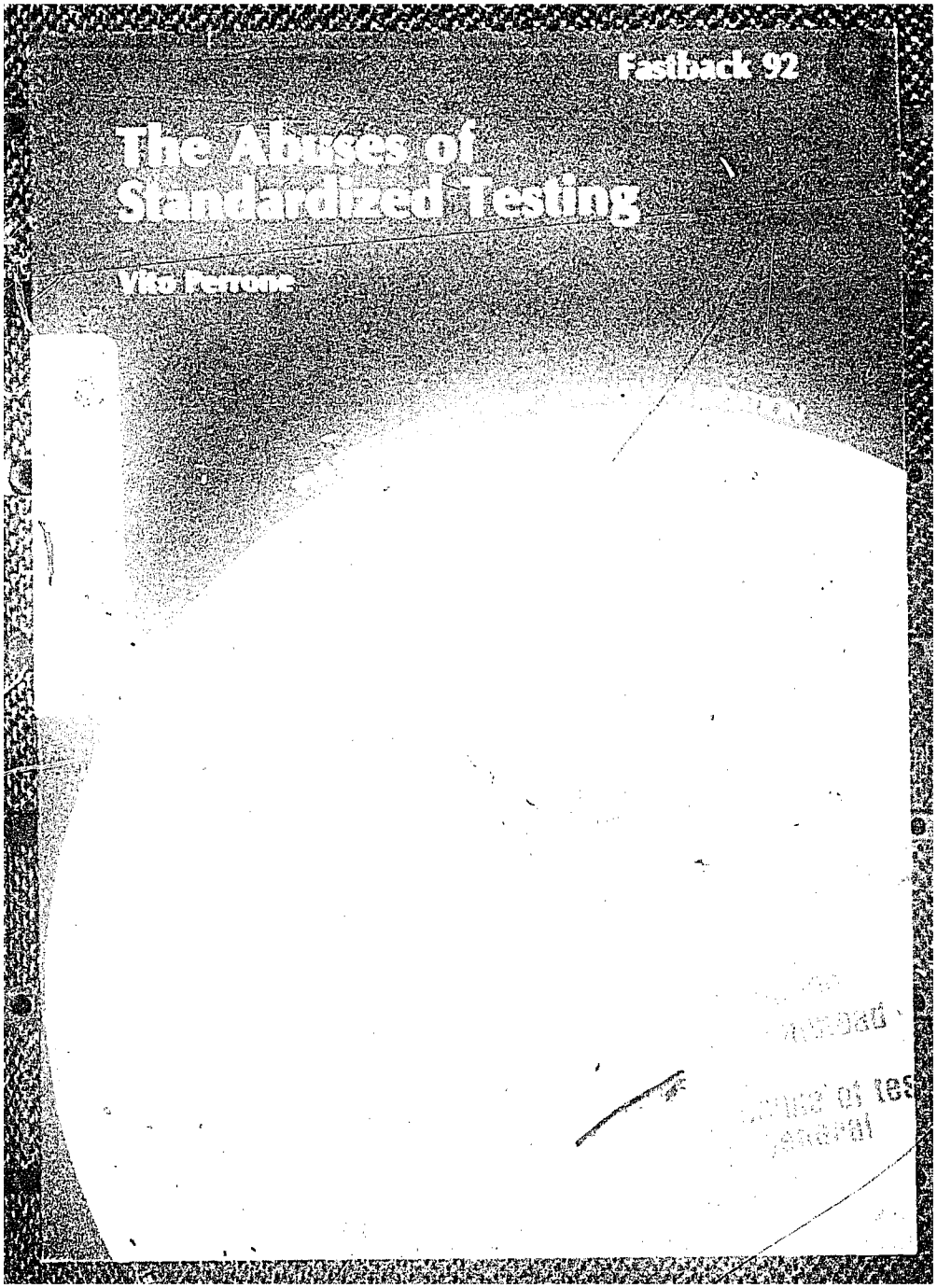
DOCUMENT RESUME

ED 141 399 TM 006 358

AUTHOR	Perrone, Vito
TITLE	The Abuses of Standardized Testing. Fastback Series No. 92.
INSTITUTION	Phi Delta Kappa Educational Foundation, Bloomington, Ind.
PUB DATE	77
NOTE	45p.; Some parts may be marginally legible due to small print of the original document
AVAILABLE FROM	Phi Delta Kappa, Eighth and Union, Box 789, Bloomington, Indiana 47401 (1-9 copies: for members, $0.60 ea., for nonmembers, $0.75 ea., discounts on larger quantities)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Achievement Tests; Educational Testing; Elementary Education; Evaluation Methods; History; *Intelligence Tests; Norm-Referenced Tests; Objective Tests; Political Issues; Scores; Socioeconomic Influences; *Standardized Tests; Test Bias; *Testing Problems; Test Interpretation; Test Results; *Test Validity
IDENTIFIERS	*Alternatives to Standardized Testing

ABSTRACT

The author takes the position that standardized tests, as presently developed and marketed, do have potentially positive uses. However, these advantages are outweighed by the tests' deleterious effects on children and programs. Standardized tests refer to published, norm referenced, achievement and intelligence tests which contain specific instructions for administration. This discussion includes a historical background of standardized testing, an explanation of the tests themselves, a proposed moratorium on testing and suggested alternatives to standardized testing. A bibliography is appended. (MV)


Fastback 92
The Abuses of Standardized Testing
Vito Perrone

VITO PERRONE

Vito Perrone is dean of the Center for Teaching and Learning and professor of history and education at the University of North Dakota.

He completed his M.A. in 1958 and Ph.D. in 1963 at Michigan State University.

Perrone has published extensively in professional and educational journals. His principal interests are alternatives in education, community organization, staff development, and educational evaluation.

Actively involved with Phi Delta Kappa, the Association for Childhood Education, the National Association of Elementary School Principals, and the North Dakota Study Group on Evaluation, Perrone also serves as a program consultant to the National Endowment for the Humanities. He is the author of Phi Delta Kappa's best-selling fastback, *Open Education: Promise and Problems* (fastback #3).

Series Editor, Donald W. Robinson

The Abuses of Standardized Testing

By Vito Perrone

Library of Congress Catalog Card Number: 77-72589
ISBN 0-87367-092-2

Bloomington, Indiana

TABLE OF CONTENTS

A Prefatory Comment

It needs to be said at the outset that standardized testing is at the center of an enormous educational controversy. It has become an issue as a result, in part, of accountability pressures; but another strong factor is the growing number of teachers who want increased control over curriculum, materials, and classroom practices. Another significant contributor to the conflict is the tension growing out of the increasing trend toward decentralization in school districts. Standardized tests, for example, were thought to make some sense when most students in a given school district were involved in a common curriculum. But as school districts have begun to decentralize, to foster alternative methods of education, common curricular patterns are being threatened. In addition, we are in an era when "equal educational opportunity" is being affirmed as never before. That affirmation has brought with it increased understanding of the ways in which children of the poor, who include a large percentage of America's minorities, racial and ethnic, have been deprived of equal educational opportunities. The role that standardized tests have played in this process fills the literature and helps fuel the debate.

The foregoing background is not meant to be all-inclusive. Many of us could provide other significant reasons why standardized testing is in the "eye of the storm." If my orientation were different, I could, I suspect, suggest—though I do not believe the argument has much substance—that educators have made standardized testing an issue to "cover up" their failures, to rationalize the "decline in test scores." But that *is not my orientation!* And that is a point I wish to make very clear to readers of this fastback.

I do not believe that standardized tests, as presently developed

and marketed, have a great deal to contribute to children, teachers, parents, or schools. Whatever merit they may have, and I believe they do have some potentially positive uses, is outweighed significantly by their negative qualities, by their deleterious effects on children and programs.

Now that I have established my personal orientation —one that will pervade this fastback—I need to provide some additional contextual information. Most of my examples will be related to the elementary grades, principally because this is where the majority of standardized tests are used. When I speak of standardized tests, I am referring to tests which are published, norm-referenced, and administered according to explicit instructions. Popular examples of standardized achievement tests are the *Metropolitan Achievement Test*, the *Stanford Achievement Test*, the *Stanford Test of Academic Skills*, and the *Iowa Test of Basic Skills*, to name only a few. Examples of standardized measures of mental ability are the *Otis-Lennon Mental Ability Test*, the *Stanford-Binet*, and the *Wechsler Intelligence Scale for Children*. My purpose in listing the foregoing is to assure as little misunderstanding as possible about what is meant when I refer to standardized tests and standardized testing.

A Basic Position on Standardized Testing

Standardized testing, essentially a post-World War I phenomenon, has become commonplace in America. Tests exist for almost every social human trait imaginable, including intelligence, alienation, self-concept, maturity, moral development, and creativity. They are used to select people for admission into and for exclusion from a wide range of educational programs, private and public projects, and jobs. Standardized tests affect Americans of all ages, in all fields; however, they come down most heavily on the young, those between the ages of 3 and 21. David McClelland suggests that standardized tests have been so thoroughly ingrained into American schools that "it is a sign of backwardness not to have test scores in the school records of children." While I am concerned about the effects of testing on individuals of all ages, I am especially troubled about its effects on young children at the primary school level. These are years when children's growth is most uneven. Not only are there great differences among individuals, but also within the individual over time. During this period a large number of the skills needed for success in school are in rather fluid acquisitional stages.

The most widely used standardized tests attempt to assess intelligence, language skills, reading readiness, achievement in various subjects, and, in recent years, one's self-concept. As I have said, the tests are commonly norm-referenced, so that users can rank order individuals or groups in relation to a particular, or norm, population. (Most test publishers claim that the populations used to establish norms are representative of the general population for whom the tests have been devised. This is, however, a questionable claim in many cases.) We have become so accustomed to their use that we often fail to ask ourselves whether the tests do in fact assess what they purport to assess or whether the assumptions that undergird the claims of the test makers are acceptable.

Some Hard Questions About Standardized Tests

Sheldon White suggests that in using standardized tests we are involved with "an affair in which magic, science, and myth are intermixed." He may well be offering an understatement. How many of us, for example, actually believe that the intelligence and competence of any individual can be adequately represented by his* score on any group-administered test? Or, that there is one "normal curve" that can provide a distribution capable of classifying all children? Such assumptions defy almost everything that we have come to understand about children's growth.

Even if one fails to take note of the implicit assumptions of the tests, an examination of the test items ought to cause enormous concern.** Are they clear? Are they fair? Do they address the *particular* educational concerns of teachers of young children? Do the tests provide *useful* information about individual children, about a class as a whole? Do they help young children in their learning? Do they support children's intentions as learners? Do they provide parents with essential information about their children? I have encountered few teachers able to provide an affirmative response to any of the foregoing questions. They do respond in the affirmative, however, to the following questions. Do teachers feel any pressure

*For clarity and economy, we use the masculine form of pronouns throughout this publication when no specific gender is implied. While we recognize the trend away from this practice, we see no graceful alternative. We hope the reader will impute no sexist motives; certainly none are intended. —*The Editors*

**A large number of publications have provided thoughtful critiques of sample test items from a variety of popularly used standardized tests. See: Deborah Meier, *Reading Failure and the Tests* (New York: Workshop Center for Open Education, 1973); Deborah Meier, Herb Mack, and Ann Cook, *Reading Tests: Do They Hurt Your Child?* (New York: Community Resources Institute, 1973); Banesh Hoffman, *The Tyranny of Testing* (New York: Collier Books, 1964); *National Elementary Principal* (March/April, 1975, and August, 1975).

to teach to the tests? If the tests were not given, would there be fewer skill sheets and workbooks, a broader range of materials, more attention to integrated learning? Would teachers prefer to use the time devoted to standardized testing for other educational activities? Do teachers feel that they can assess children's learning in more appropriate ways than through the use of standardized achievement tests?

Labels That Cripple

I have many other concerns about standardized tests. They have been used increasingly to make judgments about children. Children judged to be "below average" are not likely to receive, in most schools, the kinds of educational opportunities available to children judged "above average." Placement in remedial and other special education programs and in lower-level tracks is usually related closely to test results. Children placed in such settings are often viewed as failures; expectations tend not to be high for them. And children in such settings quickly learn to view themselves as failures, producing little. Children who are labeled in a manner that suggests limited ability find that their education takes on a narrow focus, one-dimensional tasks such as skill sheets, workbooks, and drills of one kind or another being most prominent.

Who are the children who tend most often to be labeled "below average"? The high proportion of children from lower socio-economic populations, which include large numbers of minorities, represented in special education and lower-level tracks ought to give us serious pause. Jane Mercer provides rather stark data: namely, that from 50 to 300% more black and Mexican Americans are identified as mentally retarded than could be reasonably expected from their proportion of the population. Our commitments to democratic practice and equality of educational opportunity should force us to speak out strongly against a process that consistently produces such results.

Multiple Patterns, Diversified Programs

Teachers of young children have long believed that children come to learning in many different ways, demonstrating in the process that they have multiple patterns of growth and achievement. This belief has given direction to programs which are diver-

sified as to aims and goals. In these programs, children are respected, regardless of racial background or socioeconomic class. Their interests have become basic starting points for learning. Such developmental programs have tended to support more formal instruction in reading; for example, only when children are ready and not because they are 6 years old.

Because teachers in such settings have been committed to increasing children's opportunities for successful experience and high levels of self-esteem, many learning options are made available. The clock then tends not to determine to such a large degree when children begin and end learning activities. Peer interaction—i.e., communication—is encouraged. Integral, rather than peripheral, to a child's life in the classroom are the creative and expressive forms of communication that have the capacity for developing feeling—the most personal of human possessions. (Too often, a teacher does little with the creative and expressive arts because they don't relate particularly well to the normative testing programs. They are not *basic* enough!) Static expectations for children, rooted in an array of basal materials and common curricula, do not reflect the diversity that actually exists and is supported in responsive primary schools. Yet the standardized tests are anchored in standard curriculum materials (basal textbooks, syllabi, and state guidelines) that have predetermined expectations toward which every one is expected to work. To actually develop a responsive, developmental, classroom environment is to risk lower scores on many of the standardized tests. Teachers and children do not need these kinds of external pressures.

Evaluation Consonant with Purpose

Does the foregoing suggest that evaluation is not important? Most definitely. I do not oppose evaluation; I consider it basic to the growth of programs, teachers, and children. But evaluation needs to be embedded in the classrooms. It needs to be consonant with *purpose*. Assessing children's growth, for example, is an intense activity, and it should occur daily, continuously. It is integral to everything that goes on in a classroom.

Historical Roots

Having now established a basic position, I will proceed to a discussion of some of the historical roots of standardized testing in this country. While I would certainly acknowledge that one can understand standardized testing practice and psychometrics without any special knowledge of the historical development of standardized testing, I believe that the history is important if one wishes to examine the implications of standardized testing and gain increased understanding of its assumptions.

Historical perspective is especially important to an understanding of the current critique. In suggesting a historical perspective, I do not wish to imply that the context today is the same as that which existed in the early years of standardized testing. The differences are profound, as will become clear in the narrative which follows.

The beginnings of standardized testing were taking form at the turn of the last century. It was a time of rapid change in many aspects of American life. Immigration was reaching new heights, especially with a heavy influx of southern and eastern Europeans who were considered less assimilable than earlier immigrants; industrialization, aided by a growing faith in science and technology, was firmly rooted; what seemed an uncontrollable urban expansion paralleled the increased levels of immigration and industrialization; and schools were under intense pressures to enroll larger percentages of the school-age population, especially at the secondary school level.

Psychology and education, as areas of academic inquiry, had long been stepchildren to the more traditional academic fields such as philosophy and history. Seeking status, they turned increasingly to science and technology as a base for their inquiry. That practitioners within these related fields adopted statistical procedures and scientific methods is not surprising. The standardized tests which they developed met many of the conditions of science as they were

understood at that time; in addition, and possibly more importantly, they represented an activity which could support many of the cultural assumptions of the day.

In a society clinging to egalitarian views, how could the great differences among people with regard to education, status, and power be explained? The social Darwinists suggested that such differences existed because intellectual attributes among individuals (and groups) varied significantly. The tests quickly bore this out. Individuals (and groups) could be classified on the basis of "scientific measures," selection processes could be established "without any blemish on democratic philosophy." While I would hesitate to suggest that social Darwinism was the predominant philosophical orientation at the beginning of the century, it clearly had widespread support from the middle and upper classes (who were at that time principally white, Anglo-Saxon, and Protestant) and was called upon to buttress the development of standardized tests.

Alfred Binet, an early champion of experimental psychology, is generally acknowledged as being responsible for legitimating standardized tests. In the years that followed the passage of compulsory education legislation in France (1881), Binet raised questions about the degree to which *all* children could benefit from regular school activities. By the turn of the century, Binet advocated special classes for those possessing "limited ability."

Along with Theodore Simon, Binet was invited in 1904 by the Ministry of Education to develop an identification process which might be used to select children for special classes. Simon and Binet capitalized upon the opportunity to develop a series of tests. The procedures they employed for norming and validation were not very dissimilar from those used today. (Another similarity exists in the results. Then, as now, scores reflected the social/economic structures.) While Binet was not immune to the use of such phrases as *mental ability* in relation to his test results, he had reservations about interpreting an individual's test results narrowly: he did not believe, for example, as many of his followers did, that the tests measured fixed intellectual qualities that were not amenable to further training.

After Binet's death, Lewis Termin of Stanford University began revision of the Binet testing process. The Stanford-Binet Test, published in 1916, was the result. Termin, in order to bring more

specificity to the results, attached a score which he called an *Intelligence Quotient*. In *The Measurement of Intelligence*, Termin described the "potential" for the test in clear fashion:

> . . . [I]ntelligence tests will bring tens of thousands of these high-grade defectives under the surveillance and protection of society. This will ultimately result in curtailing the reproduction of feeble-mindedness and in the elimination of an enormous amount of crime, pauperism, and industrial inefficiency.

Those with IQ scores in the 70 to 80 range were of particular concern to Termin. He wrote further in *The Measurement of Intelligence*:

> [Such intellectual deficiencies are] very common among Spanish-Indian and Mexican families in the Southwest and also among Negroes. Their dullness seems to be racial; or at least inherent in the family stocks from which they come. . . . Children of this group should be segregated in special classes. . . . They cannot master abstraction, but they can often be made into efficient workers . . . from a eugenic point of view they constitute a grave problem because of their unusually prolific breeding.

(Such arguments, though less direct, are not uncommon in the literature produced today by Arthur Jensen, Richard Herrnstein, and William Shockley, to name a few.)

Termin, along with others such as Henry Goddard (Vineland Training School in New Jersey) and Robert Yerkes (Harvard) who had similar interests in "tests of intelligence" and later "tests of achievement," eventually saw southern and eastern Europeans as also demonstrating mental deficiencies. It is not surprising that all three were active in the eugenics movement—the tests were for them excellent detection devices—and in political efforts to stem the flow of the "inferior" southern and eastern Europeans who were entering the United States as immigrants.

Goddard went to Ellis Island in 1912 to administer the Binet test, as well as others which he devised, to new immigrants. The results were hardly surprising, though to read Goddard's account in Leon Kamin's *The Science and Politics of I.Q.* it was shocking: He judged 83% of Jews, 80% of Hungarians, 79% of Italians, and 87% of Russians as "feeble-minded." In 1912 it was Goddard who also provided the classic tracing of the Martin Kallikak family, a case history still used in psychology texts as late as 1955.

When the United States entered World War I, Yerkes, then

president of the American Psychological Association, proposed on behalf of several of his colleagues (including Termin and Goddard) that psychologists could perform a service by administering tests to draftees as an aid in their military placement. Tests were given to 1,700,000 men. While they were not used by the Army for purposes of placement, they did serve to solidify the legitimacy of standardized testing and "improve" the technology of psychometrics.

The data from the testing were reported in 1921 by the National Academy of Sciences as *Psychological Examining in the U.S. Army* (edited by R. M. Yerkes). The data were analyzed further in *A Study of American Intelligence* by C. C. Brigham (Princeton University Press, 1923). There were no surprises. Whites scored considerably higher than blacks; individuals from Scandinavian and English-speaking countries scored significantly higher than those from Latin and Slavic countries. Following correlations between the test scores and the length of time an individual being tested had lived in the United States were essentially dismissed.

The results of all of this testing were influential in the passage of the 1924 Immigration Act, which placed discriminatory restrictions on the immigration of non-Anglo-Saxon populations. It was a major victory for psychologists such as Termin, Goddard, Yerkes, and Brigham. In fairness, however, it should also be said that the victory was not theirs exclusively. There were millions of Americans who were convinced that immigration should be restricted and did not need the psychological test scores for confirmation. (Provisions of the 1924 act relating to national origin quotas were maintained in the McCarran Act of 1952. By the 1960s, however, pressures were building to bring an end to all discriminatory legislation. In 1965 Congress passed the Immigration and Nationality Act, which eliminated the quota system based on national origins.)

Walter Lippmann was one of the few individuals of the time who raised a voice of protest; his outlet for criticism from 1922 to 1924 was the *New Republic*. In an early commentary, he wrote, "The real promise and value of the investigation which Binet started is in danger of gross perversion by muddleheaded and prejudiced men."

In the issue of November 15, 1922, Lippmann wrote:

> . . . [I]ntelligence is not an abstraction like length and weight; it is an exceedingly complicated notion which nobody has yet succeeded in defining. . . . If the impression takes root that these tests really measure

intelligence, that they contribute a sort of last judgment on the child's capacity, that they reveal "scientifically" his predetermined ability, then it would be a thousand times better if all the intelligence testers and their questionnaires were sunk without warning in the Sargasso Sea.

Lippmann closed his original series of six articles on the intelligence tests by suggesting that psychologists back away from their "pretentious" directions and "save themselves from the humiliation of having furnished doped evidence to the exponents of the new snobbery."

Termin's responses to Lippmann deserve mention. They are similar in content to what is produced by many contemporary apologists: namely, appeals to scientific authority, ridicule, and nonsequiturs. In one resonse (November 29, 1922), Termin suggested, along with considerable ridicule, that Lippmann had "some kind of emotional complex." Lippmann's response in the January 3, 1923 *New Republic*:

> Well, I have [an emotional complex about this business]. I admit it. I hate the impudence of a claim that in 50 minutes you can judge and classify a human being's predestined fitness in life. I hate the pretentiousness of that claim. I hate the abuse of scientific method which it involves. I hate the sense of superiority which it creates and the sense of inferiority which it imposes.

But Lippmann's charges, though powerful in tone, were not particularly influential. Tests, those producing IQ scores and those producing achievement scores, proliferated rapidly in the 1920s and 1930s. They fit many school needs of the day by providing external procedures to justify promotions in the schools—now more committed to age-grade patterns than ever before. (That they served to justify the continued pre-eminence of the privileged in American society seemed not to be a problem for the majority of educators.) And, as was noted earlier, they fit the scientific ethos of the period.* Even the progressives, especially in the twenties, gave passive

*Robert Thorndike and Elizabeth Hagan write: "The testing movement seemed especially suited to the temper of this country and took hold here with a vigor and enthusiasm unequaled elsewhere." *Measurement and Evaluation in Psychology and Education* (New York: John Wiley, 1962), p. 5. Readers should understand that standardized testing has become in large measure an American phenomenon. It has never had very much consistent support in the rest of the world.

approval to the testing activities. To attack "science" was not consistent with their basic approaches to education. In the thirties there were some shifts in progressive thought. (For example, attempts were made to have standardized testing examined within the Progressive Education Association. But the organization was in the beginning of its intellectual twilight by this time.)

I recognize that I have focused on several individuals whose sociopolitical views influenced their work decisively and, by today's standards, negatively. This might lead some readers to suggest that my treatment, at least to this point, has not been balanced. I could have highlighted, as most texts on measurement and testing do, the pioneering efforts of Termin and Yerkes, the contribution of Termin to our understanding of the gifted (he did make a number of significant contributions), and the efforts made by conscientious measurement scholars such as E. L. Thorndike (and I will comment further on Thorndike) to improve testing practice. I could have alluded (as several of the texts have) to the arguments about the influence of heredity on intelligence and achievement test scores and to the concerns about the tests really testing social class. I might not have mentioned eugenics at all (as most texts do). The point is that the contemporary literature produced by measurement scholars does not provide any balance. My purpose has been to raise aspects of the history of testing in America that have not been fully in view; in other words, to bring some balance to the discussion.

Were there other forces besides eugenics and the scientific ethos that helped make testing such a popular enterprise? There was a belief often expressed in the early literature that the tests represented a democratic, objective process for selecting students for admission into colleges and into particular professions, for passage to new grade levels, and for receipt of academic honors. Some egalitarians argued that a test score removed the possibility of social status or faulty, prejudiced, teacher/school judgment determining one's entry, for example, into the elite schools. (The elite schools received the same students after standardized testing as before, but the illusion of democratic practice survived the twenties and persisted well into the 1950s and early 1960s.) How many times have we heard about the child who came from a lower socioeconomic and minority background who was singled out because of high test scores and

given academic opportunities that might not otherwise have been available? There have been many such cases, but when we consider all of the individuals of lower socioeconomic and minority backgrounds who owe increased levels of academic opportunity and altered social or economic status to test scores, the numbers were surely small in the decades between 1920 and 1950 and there is little to suggest that conditions have changed.

Ralph Tyler, in his critique on testing, comments that standardized testing began "as a means for selecting and sorting people, and the principles and practices of testing that have been worked out since 1918 are largely the refining of means to serve these functions rather than other educational purposes." Sorting and selecting, Tyler would suggest, were viewed in the early period as natural and necessary functions of the schools; hence, why would tests designed for such purposes be questioned seriously?

A stable force in the early testing movement was E. L. Thorndike who, in the long run, may well have had a greater influence than the eugenicists. His *Introduction to the Theory of Mental and Social Measurements*, published in 1904, was an important contribution to the measurement field. The *Thorndike Handwriting Scale*, produced in 1909, was the first popularly used standardized test in the public schools.

"Whatever exists at all exists in some amount" was a classic Thorndike phrase and one which says a great deal about Thorndike's basic approach to testing. In "Nature, Purposes and General Methods of Measurement of Educational Products," *(The Measurement of Educational Products,* Seventeenth Yearbook, Part II, National Society for the Study of Education, 1918) Thorndike commented about some of the problems that were beginning to surface as early as 1917-18: "[Those] directly in charge of educational affairs have been so appreciative of educational measurement and so sincere in their desire to have tests and scales devised [that quality is sacrificed]. . . . Opposition, neglect, and misunderstanding will be much less disastrous to the work of quantitative science in education than a vast output of mediocre tests for measuring this, that, and the other school product, of which a large percent are fundamentally unsound. . . ."

Thorndike was a serious student of measurement. Eugenics was not his interest. He continued to raise concerns throughout the

1920s, a veritable boom period for the development and marketing of tests, about the poor quality of tests, the uncritical acceptance of test scores, and what he conceived to be the unjustified judgments being made about individuals. His goal was the production of *better* tests and more knowledgeable test use. While it is possible to disagree with Thorndike's basic assumptions about education and learning, one must, as I do, respect the way he carried out his commitments.

The technical quality of tests improved in the decades following the Depression. And these improvements are noted often in the educational literature. The eugenics advocates who were prominent in the formative years were gone, to be replaced by a growing corps of psychometric technicians. Norming and validation procedures became increasingly more sophisticated. And testing became a part of the conventional wisdom of schools. Debates were few; criticism almost nonexistent. (In fact, the literature includes very little criticism until the 1960s.) It should be noted that the standardized tests, while accepted and used as a basis for many positive and negative decisions about students, were not viewed as overbearing. They did not dominate curriculum or teaching. The amount of testing was not great in most districts, and had what appeared to be a benign quality in many others.

Standardized testing received a boost with the ascent of Sputnik I, when questions about the quality of schools began to accelerate. In 1965, with the passage of the Elementary and Secondary Education Act, testing, and the industry supporting it, began to expand rapidly. With the heavy influx of federal dollars came increasing demands for evaluation. And evaluation in most instances became synonymous unfortunately, with outcome data produced by standardized tests. In part this occurred because standardized tests and the technology supporting them were available and evaluation paradigms which might have been more appropriate were not well developed, or lacked the narrow "scientific construct" that was increasingly demanded by a "single-score" mentality.

We are now in a period where standardized tests are a major issue in schools. (It should be noted again that intelligence tests are not as much an issue now as they once were. Their use is diminishing rapidly.) The level of criticism is related closely to the volume of testing that occurs. Producers of tests have been surprised at the

harshness of the criticisms, feeling that the technical quality of their products is higher now than at any time in the past. They argue that the problems—and they now acknowledge that there are problems—are related to use, or more specifically, misuse.*

While these are certainly considerations, the issues are, from my point of view, deeper than use and misuse. We are seeing teachers, and in many settings parents, going back to fundamental questions about the purpose of the schools, the ways in which children's development is being supported. This is, in large measure, what part of this fastback is about, and it underlies much of what I am trying to communicate.

*It needs to be acknowledged that the test producers are attempting to reduce misuse by providing test users with very carefully prepared manuals relating to their tests. These manuals tend to point out painstakingly the particular test's limitations as well as constructive uses. They typically caution, in the case of achievement tests, that the test results ought not to be used to evaluate teachers, that grade-level equivalency scores are misleading, that growth is not unidimensional, that many external factors might affect a child's score, that the test measures a sampling of curriculum only, etc. But, having said all of this, it needs also to be stated that few teachers have ever seen the manuals, and my experience is that few schools act on the cautions that the manuals provide. This point will be discussed more fully later in this text.

The Tests Themselves

Having provided a basic position statement and historical review, I will address several issues which relate generally to the technical aspects of testing and which tend to create misunderstandings and often misuse. I make no attempt to be all-encompassing; the problem areas are just too large.

In discussions of standardized testing one often confronts terms/concepts such as *objectivity*, [illegible] *validity*. The [illegible] give them an aura of science, however, what do they mean in basic, nontechnical language?

A test is considered *objective* if everyone takes it under the same basic conditions. The multiple-choice format, buttressed by a single "right"-answer pattern, supports objectivity. But objectivity has nothing to do with whether a test is fair, contains items of importance, or has ambiguous questions and answers. Objectivity has, in other words, *no relationship to quality.*

A test is *standardized* if norms have been established. Whether the norm populations are representative in more than a statistical sense is not the defining characteristic. This term, as in the case for objectivity, *has no relationship to quality.*

Reliability relates to the consistency of the test: How close are the results for an individual or group at two different testings? Or how close are the scores of individuals on two different forms of a particular test? Reliability is rather simple to establish. But a test can have very high reliability (most popular standardized achievement tests carry reliability coefficients of .87 to .93) and yet be a very poor test, measuring little considered important to large numbers of people who use or take the test. The latter point relates to *validity*, which has an interdependent relationship with reliability. Validity doesn't, however, receive as much attention.

Validity refers, at the most simple level, to the degree to which a test measures what it is supposed to measure and/or the degree to which the scores derived from a particular test can be related to what the test is supposed to be measuring. In other words, what are the inferences that can be drawn from the test scores? Validity, unlike reliability, is difficult to establish with any authority. It is typically determined by having an expert (or experts) examine a particular test and provide the equivalent of an imprimatur. This is content validation by opinion. Content validation is the focus of most standardized tests used in elementary and secondary schools. (Those who suggest that the tests often contain bias are, in essence, questioning content validity by claiming that the test content is not representative of the socioeducation[illegible] of minority people.) Validity [illegible]ity) is also [illegible] compar[illegible] the results with other measur[illegible] i.e., other tests, grades, or teacher judgments.

So much for the terms. What can be said about test content? For many individuals with reservations about standardized tests, the content is the principal issue. This relates to the concept of validity but is, at the same time, broader. Most currently used standardized achievement tests, as was noted in my prefatory comment, have been constructed to conform to instructional programs with predetermined objectives and materials which everyone is expected to work through. They have less relationship to programs which stress high levels of individualization and flexibility of objectives.

How are the tests constructed? In preparing items for an achievement test, authors typically survey curricular patterns and basal materials; they attempt to learn about the sequence, if any, that tends to exist in various content/subject areas; and they make decisions about how to establish a balance between information items and concept items. Questions are prepared and generally tried out in a variety of school settings. (The particular items selected to try out represent, in effect, a statement about what the test authors consider important. This is not intended as a negative observation; it is, however, a condition that needs to be understood.) Those items which most individuals get right or wrong are discarded. *Distribution* is desired, inasmuch as the entire battery of items is designed to produce, among a sample of students who take the test, a *normal curve*, a construct in which half the students score below the average

and half-above. In such a process, items which many would suggest are important might well be discarded and items of limited importance retained. Teachers, school administrators, and parents would do well to examine closely the questions which appear in the standardized tests used in their schools, or being considered for use, in order to make a judgment about the importance of the questions and their relationship to the local curriculum.

Once the items are developed, standardization procedures begin. This involves testing a norm population and constructing norm scores. How long does all of this take? From start-up to publication, two years might pass. But what if curriculum changes are rapid and/or new goals emerge in schools? Might there then be a gap between the curricular assumptions of the tests and the curricula that actually exist? This certainly is the case in mathematics where math tests have been concerned for the past decade, as they were in the previous decade, with computation in a base 10 system only—hardly evidence of the "new math." What if teachers really believed that a shift in educational direction was necessary? Is it possible that such a shift might not occur because of the risk of lowering results on an achievement test designed for different purposes? And what if the populations taking the tests change significantly from one standardization to another? Do the scores derived then really have the same meaning?

It is important, I believe, to comment briefly on test statistics—the derivations which bring scores relating to a norm population and provide a basis for giving meaning to the *raw* scores (the number of "correct" responses on a test or subsection of a test). While it could be argued that everyone using tests should know a good deal about *derived scores* and their meaning, my experience is that far too many do not.*

Test results are most often reported as percentile scores, stanine scores, or grade-level equivalency scores. Forty-three correct answers out of 80 items on the language section of a very popular

*I must note that the test manuals accompanying most popularly used standardized tests are replete with information about derived scores, i.e., how to interpret them and what limitations need to be taken into account when using them. But as noted earlier, the manuals are not often available in schools for teachers to read, and little information about derived scores goes out to the public to increase their understanding.

achievement test (1974 version) which students take at the end of the seventh grade is converted to a percentile score of 52. This indicates that 52% of those who took the test as part of the norm population scored 43 or less and 48% scored higher than 43. Stanine scores, unlike percentile scores and grade-equivalency scores, are suggestive of a range. (For this reason test publishers increasingly are encouraging their use.) All raw and percentile scores are grouped to make up a nine point, or *stanine* scale. A stanine score of 5 is average; 40% of the scores will then fall above this average and 40% below. On the test cited above, the percentile score of 52 falls within the fifth stanine, along with all percentile scores between 40 and 58. The diagram below presents range raw scores converted to percentiles and then to a stanine scale.

Stanine	1	2	3	4	5	6	7	8	9
Percentile	1-2	4-10	11-22	23-39	40-58	59-76	77-88	89-94	95-99
Raw Score	14-17	18-22	23-28	29-36	37-46	47-55	56-63	64-68	69-79

The *grade-level equivalency* score is derived essentially by assigning to the median score of a seventh-grade norm population a *grade-level equivalency of 7.0.* Scores above and below the median are assigned grade-level equivalencies above and below 7.0. It is an estimation, nothing more. The score of 43 on the language section of the test under discussion (taken at the end of the seventh grade) converts to a percentile score of 52 and a grade-level equivalency score of 8.3 (eight years, three months). A score of 41 converts, on the other hand, to a grade-level equivalency of 7.8 and 45 converts to 8.7. *Two questions right or wrong cover a range of 11 months.*

The publisher of the foregoing test lists a *standard error* for the language section as 3.9. This standard error indicates that two-thirds of the time one could expect a fluctuation of 3.9. As the test manual notes, "We could expect with about 68% certainty that the true score [for a student with a raw score of 43] would fall between [39 and 47]." This is between the 44th and the 60th percentiles; the grade-level equivalency range is 7.2 to 8.9. And one-third of the time there may be even more error. The point of all of this is that *the scores are very imprecise; one has to be very careful in attaching too much importance to them.*

Of all of the derived scores, grade-level equivalency is the most

commonly used, even though it is the most misleading. Test publishers now regularly, in their manuals, point out that grade-level equivalency scores "are being questioned as an appropriate means of interpreting the test performance of individuals and groups." They suggest further that "grade equivalents are not an equal-unit score scale . . . statistical computations based on grade-level equivalency values, are not, strictly speaking, legitimate." In some cases they admonish users not to report grade-level equivalencies at all. Henry Dyer, possibly the most respected authority in testing, has called grade-level equivalency scores "absurd, wrong, and misleading." And in even stronger language he commented (*The United Teacher*, April 14, 1971, p. 15) that they are "statistical monstrosities . . . [that they] lure educational practitioners to succumb to what Alfred North Whitehead called the 'fallacy of misplaced concreteness.' " But grade-level equivalencies continue, in part because school people have been lured, with many accountability models, to measure *growth* (or "misplaced concreteness"). If children are in school for eight months, "then they should make eight months' gain." Only grade-level equivalency scores report in year-month terms.

In one school district in which we recently conducted a review of the accountability system, the following appears in the *Statement of School District Objectives:* "Students [are] expected to gain, on the average, eight months in academic achievement between September and May." And in a booklet related to the particular standardized testing program used in the school district, the following is presented: "A student is expected to grow academically one month for each month that he/she is in school. Since there are eight months between the administration of the pretest and the posttest, it is desirable for the student to gain eight months during the time."

I won't describe what teachers engage in in order to achieve eight months' gain, but to enter into this kind of gain score mentality one has to again misrepresent the tests and abuse the scoring system. Most test publishers make reasonably clear—though they are not as direct as they could be—that gain scores are "fraught with problems," inasmuch as the tests were originally constructed only to identify present status. (And even the problems of interpreting present status are immense, according to the test manuals.) In order to address the question of eight months' growth, one has to assume

that growth is unidimensional and linear and that eight months of schooling corresponds empirically with eight months' gain on a standardized test. Neither is the case!

What does it mean to learn that a third-grade child (or class) is reading at a 7.9 grade level? Does it suggest, as I have often heard or read in the newspapers, that this particular child (or class) is reading as well as average-achieving youngsters completing the seventh grade? To begin with, what the tests measure as reading ability in grade three is not necessarily the same thing as reading ability measured in the seventh grade. A score of 7.9 is nothing more than an extrapolation above the mean. It has in no real sense anything to do with how well a youngster completing the seventh grade reads. Conversely, what interpretation is to be made for a seventh-grade child with a grade-level equivalency score of 4.0 on a reading test designed for seventh-graders? It implies clearly that the youngster's reading score is much lower than average-achieving seventh-graders who were part of the norm population. But does it establish that the youngster reads only as well as a fourth-grader? The chances are that if this seventh-grader took the test designed for fourth-graders his grade-level equivalency score might be 7.0. Remember, the tests were normed at particular grade levels. Third-graders didn't take a test designed for seventh-graders and seventh-graders didn't take a test designed for fourth-graders. We are contending at best with a statistical construct. Yet the use of grade-level equivalency scores goes on unabated.

Much of this section has dealt with test scores in a broad sense. The problems, however, grow as the context narrows and the tests are used to say something about an *individual* student's achievement of particular skills or to determine specific instructional needs. (Even the test producers are promoting some of this direction.) But given the content sampling that is involved, the manner in which the tests have been constructed, the paper and pencil multiple-choice format, and the sources which exist for error, this isn't an eminently valid use. In order to make decisions about individuals, individual student scores on specific test items or subparts of a test must be used. Even from a technical perspective this is a significant problem. Let us examine some of the sources of error. The health of a child on the day the test is given can affect the score. Noise in the classroom, teacher attitudes toward the test, whether a child has taken similar

tests, a broken pencil, and any number of similar disturbances can influence a particular score. The mental state of a child—depression, boredom, elation, anxiety about the test—can also make a difference in how the student performs.

Simple mechanical errors such as marking the wrong box on the test sheet by accident, overlooking a question, or missing a word while reading are relatively common test-taking problems. Children experiencing difficulty with reading will perform poorly on tests concerned with reading; but they will also tend to perform poorly on social studies, science, and math sections of achievement tests, inasmuch as these require reading skills. Thus a child's real knowledge may be considerably underestimated.

Many of the sources of difficulty outlined above—and the surface has scarcely been scratched—can affect an individual's score. And such sources of difficulty have little to do with how "good" the test is, how carefully it was prepared, or the validity of its content. Difficulties are, in general, intrinsic to the nature of standardized testing.

How serious is the kind of error described briefly above? It depends to a large degree on how the test results are used. In reports of test scores for large groups of children, it is possible to expect that many of the mechanical errors and related difficulties suggested above will balance out; some children will score above their "true" score and others below. The larger the group tested, the more likely it is that such a balancing will occur. But for a single individual, no other score exists. Nothing can compensate for error when a single score for an individual is used for such purposes as curriculum placement, advancement, etc.

Inasmuch as reading tests are common and are used for placement or for establishing skill assignments, I will offer some general comment on them. The criticisms that I offer, however, are to some degree also relevant to the reading section of a general standardized achievement test as well as to other subject areas. At the lower grade levels, reading tests are heavily dependent upon a particular vocabulary. If a youngster's vocabulary does not include many of the words in the test, are we really to assume the problem is reading? In addition, many of the questions which appear depend on information that is not provided. A child can read the items and all of the responses and then select the "wrong" answer. The problem is

a lack of information and not a lack of reading skill. It is also possible, of course, given the ambiguities in many items, for a child to select the "wrong" answer but read very well. Another issue is the obvious cultural bias that appears in reading tests, especially at the primary level. At this level, the tests make heavy use of pictures which often reflect particular experiences that are not necessarily part of the experience of large numbers of children in the United States.

The foregoing, and similar critiques that could be offered, are possibly small issues. A more serious question, at least for those interested in reading as an area of inquiry, is related to the assumptions which underlie most standardized reading tests. In general, reading tests assume a hierarchy of specific skills. Exercises relating to words in isolation, decoding, syllabication, and the like are common. But there is no agreement among reading experts that any hierarchy of skills exists.* Many of the skill sheets that children are seen struggling with are related directly to a hierarchy of skills. In fact, several of the tests have correlative materials which can be assigned to students who score poorly on a particular skill area. There is some evidence that such activities will increase scores on reading tests, but there is little evidence that such activities enlarge a child's capacity to gain understanding from the printed page (the way in which many individuals define reading). The time taken doing skill sheets on syllabication, for example, might have been better spent reading, enlarging one's experiences with words in new contexts, etc.

Having raised in an indirect manner the issue of time, I wish to pursue it further. Testing takes time. Does it add significantly to a child's learning? Or does it take time away from other, more significant, learning experiences?

In many schools actual standardized testing time for most children takes four days in the fall and four days in the spring. But how much time goes into preparing children directly for the tests themselves? And if a child is "targeted" under Title I, he is likely to be in for another dose of pretests and posttests in reading and math. If

*A group of reading experts came together under the auspices of the International Reading Association in 1973 to discuss reading and reading tests. They agreed almost unanimously that the existing norm-referenced reading tests were without a theoretical base. They agreed further that there is "no definitive knowledge regarding either the sequential learnings or component skills that children must acquire in order to read successfully."

the "targeted" child is also in a Follow Through program, he may well receive another battery of tests related to the National Evaluation Project as well as other batteries mandated by the Follow Through sponsor. The possibilities proliferate the more one thinks about testing in schools.

What is learned through all of the testing? The question that must always be asked in addition to all that has been said is: *Do the tests provide more information about a child's achievement in most subject areas than the child's teachers typically possess?* In general, no! Teachers can, in most cases, provide more precise information to a parent about the quality of a child's reading or math skills than any standardized test score can. Do the test scores inform teachers about what they should do? There is nothing inherent in the tests or the scoring mechanisms that provides a capacity for informing teachers of what they should do.

A Moratorium?

Thus far I have provided a position statement, some historical background, and a brief introduction to test characteristics about which it is important to know something. I now wish to address a question that looms large on the horizon of the standardized testing controversy: namely, *ought there to be a moratorium?*

To raise the question of a moratorium among teachers, school administrators, parents, school board members, and legislators appears to elicit fear, even when there is a negative orientation toward standardized testing. A moratorium seems for many to be an ultimate step that might throw education into a chaotic state. Such is the authority that standardized testing has come, over the years, to wield. Nonetheless, the gauntlet has been thrown. The National Education Association passed a resolution calling for a moratorium on the use of standardized intelligence and achievement tests in 1972. In the past year, after several years of relative indifference to the resolution, the NEA has become aggressive in its support of a moratorium. The National Association for the Advancement of Colored People issued a moratorium statement in May, 1975. The Association for Childhood Education International gave support to a moratorium in 1976, and the Association for Supervision and Curriculum Development, American Association of School Administrators, National Association of Elementary School Principals, and the National Council of Teachers of English, while not calling directly for a moratorium, from 1974 to 1976 used particularly

vigorous language in agitating for a reconsideration of all uses of standardized intelligence and achievement tests.*

There are many individuals and groups who share the perspectives of the organizations listed above but feel that a moratorium, if there is to be one, should be aimed exclusively at *group-administered* intelligence and achievement testing. They hold to a belief that tests yielding normative scores can be used "if the tests are administered on an individual basis by a skilled examiner who makes sure that the child understands what he is supposed to do and wants to do it," says Millie Almy in *The Early Childhood Educator at Work*. Such a position seeks too much! A moratorium, I believe, has the potential for encouraging the development of—and legitimization for—alternatives to existing standardized testing practices. This is a crucial direction, inasmuch as evaluation, as noted earlier, is clearly essential to the qualitative improvement of educational practice in schools and the learning of children and young people. A moratorium also holds promise for intensifying critical reexamination of the politics of testing, the problems of misuse,** and the negative effects of standardized testing practice on children, teachers, and programs. The more deeply the foregoing are understood, the higher will be the potential for future reform efforts.

Would a moratorium on standardized testing disrupt school practice and bring an end to all evaluation? There is no reason to believe that either would occur. Many school districts do not use any standardized testing program; yet their evaluation practices are intense. Will standards decline? There is no evidence that standards

*It should be noted that group-administered intelligence tests have been banned in California and New York and that legislation prohibiting all intelligence testing is pending in Massachusetts. Legislation of this sort may well proliferate, inasmuch as intelligence testing is even being questioned by testing proponents. Henry Dyer has suggested that the scores derived from intelligence tests are "dubious" and "based upon an impossible assumption about the equivalence of human experience and the opportunity to learn." William Turnbull, president of the Educational Testing Service, commented at a symposium on testing (Arlington, Virginia, May 7, 1976) that "the sooner we end the use of all so-called intelligence tests, the better."

**Test publishers, as has been noted, acknowledge much of the misuse that occurs; a moratorium might provide time for test publishers to reestablish the authority of their efforts, enabling them to develop procedures that assure proper use of their tests and also to enter into collaboration with those who are developing alternatives to norm-referenced evaluation procedures.

have any relationship to the use or lack of use of standardized measures. One could even argue, suspect, that school standards in the United States have declined as standardized tests have increased in use.

Would a moratorium on standardized tests cause schools to fall back on "unsystematic evaluation processes," fostering an increase in "discrimination and ignorance?" Such an argument is popular but has little, if any, empirical data to support it. A moratorium would provide an excellent opportunity to assess such a belief.

To call for a moratorium is, for the most part, an appeal to moral authority. It can't really be more than that. Few wish to see federal or state legislation or court orders as the base for the reexamination that cries out for attention.

Alternatives to Standardized Tests

This brings me to the concluding section of this fastback and a discussion of alternatives to standardized tests.

How might teachers and schools proceed with an evaluation program that does not include standardized tests? Some alternative directions follow. They are clearly not all-inclusive; and many are, in fact, merely reaffirmations of practices that many teachers engaged in before they experienced the disruptive pressures of increasing numbers of standardized tests.

Supporters of standardized tests often argue that the tests are "objective" measures that serve as a check on the "subjective," inadequate assessments made by teachers. (Yet, interestingly enough, one source for validity checks of many standardized measures has been teacher judgment.) I do not accept the assumption that teachers have inadequate record-keeping and assessment skills. When standardized test supporters acknowledge that teachers do possess some of these skills, they argue that the tests are necessary because teachers and schools are not often organized sufficiently to describe children's learning or school programs. It is true that to engage in a systematic process of documentation is to expend considerable effort. Fortunately, increasing numbers of teachers at all levels wish to make such an effort.

Systematizing Documentation

What might a group of teachers in a given school want to look at? What might they view as especially important to document? *Answers need to come from the local school.* (Individuals external to the school have typically determined what it is important to document. And such a process has contributed significantly to the negative

character which evaluation has tended to assume.) I do not mean to imply that for a school as a whole (or particular clusters within a school) to make such decisions it is necessary to have standard record-keeping procedures in all areas in every classroom; I do mean that there must be some consensus about what areas will be looked at closely by a group of teachers. Where such a consensus exists, individual teachers not only receive the support of their colleagues, but they also have others with whom to share their documentation and reflection. Moreover, such a condition provides a climate in which teachers can feel comfortable while observing each other's classrooms, interviewing each other's students, and seeking and providing assistance.

Process, Content, Context

In documenting the *process of learning*, teachers in a school might wish to include information about the children's originality, responsibility, initiative, and independence of effort. In relation to the *content of learning*, they might wish to consider materials the children produce (such as writings and drawings); evidence that instruction deals with important concepts as well as necessary skills; and evidence that children find meaning in their learning, that it is not merely rote. And in relation to the *context of learning*, they might consider the basic human relationships that exist—child to child, child to teacher, and teacher to teacher—and see how much respect there is for the efforts and feelings of others.

The Prospect School in North Bennington, Vermont, uses some of the following records for its basic documentation: children's work (for example, drawings and photos); children's journals and notebooks of written work; teachers' periodic assessments of children's work in math, reading, and other activities; curriculum trees; and sociograms. The documentation is so complete that few individuals ask about a standardized test score. A synthesis of these records with *precise* statements regarding work in math, literature, reading, etc., is prepared at the conclusion of each year. This provides the subsequent year's teacher with rather full information about where to begin with a child. It should be noted, however, that within the school communication among teachers is sufficiently high to enable teachers to go beyond the year-end statements to the fuller documentation that is available in relation to each child. The

year-end statements serve as *the* record for a youngster who transfers to another school. These are far more precise operationally than test scores and are typically viewed as more helpful.

Interviews

At the University of North Dakota a process for documentation has been developed that includes interviews conducted with teachers, children, and parents.* These interviews have been used extensively as a base for program evaluation and staff development. The *teacher interview* provides a context for individual teachers to reflect on their intentions, use of materials, relationships with children, organization of time and space, difficulties, successes, and so on—in other words, the teacher's own perspective of the classroom. The *child interview* provides another important perspective, focusing on such issues as how the child uses materials, pursues learning, understands what is occurring in the classroom, uses the teacher, and relates to other children. The *parent interview*, bringing in a third perspective, is aimed at a description and understanding of parents' perceptions and attitudes about what is occurring in the classroom, the degree and kinds of their involvement in the classroom, what they believe is important, how they view their children's progress, and their overall level of support (or lack of support). The three interviews provide an enormous amount of qualitative evaluation information about classrooms and schools. No standardized test can provide as much data or make as much difference in what teachers do and how children learn. This is especially true when the information gathered in the interviews is seriously considered and discussed.

Broadening the Base of Operations

Teachers can keep up on children's progress in such areas as reading, language development, and math through systematic observation and frequent conferences (recorded). Can a standardized achievement test [illegible] reveal as much as carefully kept records maintained over a period of time?

*This effort has been supported, to a large degree, by a National Institute of Education research grant (No. [illegible]-0979).

Many teachers make use of informal reading inventories as a means of monitoring reading, especially when they wish some rough comparative information. Brenda Engel recently devised a number of reading tasks (similar to those used in informal reading inventories) to sample the reading level in the Cambridge (Massachusetts) Alternative School. Defining reading as "the ability to get meaning from the printed page," she categorized children as "those who can read," "those who are still in the process of acquiring reading," and "those who are nonreaders." Children are asked individually to read a story (approximately 100 words in length) silently. The interviewer says: "Tell me what the story was about." After recounting the story, the child is then asked to read the story to the interviewer and, after reading, to add to what he had related previously. In *very general terms,* a "reader" is one who "can read the text silently and relate the story adequately and/or can read the text aloud with fluency." A "nonreader" is one who "indicates he cannot read the text silently or is unable, after reading the text silently, to convey the principal meaning of the story and is unable to read aloud more than a few sight words." A range of reading needs can be identified in the process of this reading exercise.

Math checklists which teachers find useful are often provided, along with the various math programs used in schools.* And individual teachers or groups of teachers can prepare their own checklists; they can also devise informal inventories of math understandings.**

Project Mathematics (Minneapolis: Winston & Sons, 1974) is a program that provides particularly effective checklists for teachers and children. The Nuffield Mathematics Project provides "check-up" guides to determine children's growth in a variety of concepts. See also Nancy Langstaff, *Teaching in an Open Classroom: Informal Checks, Diagnoses, and Learning Strategies for Beginning Reading and Math* (Boston: National Association of Independent Schools, 1975), for some excellent ways of using informal checks productively. Langstaff's case studies provide a realistic context and should be useful to teachers and principals. We need many more such descriptions, written by classroom teachers or careful classroom observers, in order to enlarge teachers' understandings of such record-keeping and evaluation processes.

**Readers may wonder why there has been no mention yet of criterion-referenced testing. Many feel that criterion-referenced testing has enormous potential and may be a useful replacement for existing norm-referenced achievement testing. In many respects, criterion-referenced testing programs are an improvement. They have potential, unlike the norm-referenced testing that has dominated the schools, for providing some useful information about children's performance in relation to the direct instructional purposes of teachers or of the particular math, reading, or social

Teachers' Roles

All of the foregoing kinds of records can be particularly helpful to a teacher in planning learning activities that relate to a specific child or group of children. Teachers with whom we work feel this is a critical aspect of their work.

For teachers to make a conscious effort to document in some of these ways, they must step back and observe from time to time. To make such observations meaningful it is necessary to have a wide range of learning activities available for children to engage in during the observation. Otherwise the activities are so undifferentiated that the observations will provide limited insight into children and their learning patterns, interests, and needs. Being free of standardized tests might encourage such classroom environments.

Children's Roles

How can children themselves contribute to alternatives to the testing? When children participate in record keeping—maintaining daily or weekly journals, filing samples of their writing, recording the books they have read or the math concepts they understand—they not only provide information to the teacher but they have an increased sense of where they are and what they need to do to extend their learning. Learning takes on a personal character, encouraging students to assume greater responsibility for their own learning. (Can any of the standardized tests do as much?)

---

studies programs that teachers are actually using. They also have significant limitations. Criterion-referenced tests have been typically constructed around items that lend themselves easily to [illegible] measurements "directly interpretable in terms of specified performance [illegible]" (Robert Glaser and Anthony Nitko, "Measurement in Learning and Instruction," [illegible] *Educational Measurement* (Robert Thorndike, ed., Washington, D.C.: American Council on Education, 1971, p. 626). They tend to measure *simple* tasks at the expense of higher-level thought processes (Robert Stake and Dennis Gooler, "Measuring Goal Priorities," *School Evaluation*, Ernest House, ed., Berkeley: McCutchan Publishing Co., 1973) and to reinforce teaching of skills in isolation. They provide no more guarantee than the norm-referenced tests that the behaviors expected are really important or that the curriculum will not be developed principally to meet objectives that have little significant challenge to children or assist them in their general deve[illegible] They tend to stress end products; processes of learning and thinking may [illegible] importance (Vito Perrone and Warren Strandberg, "A [illegible]spective of [illegible]," *Teacher College Record*, February, 1972). While I do [illegible] on the bas[illegible] [illegible] have very much confidence in criterion-referenced testing program[illegible] [illegible] that efforts in this direction are still in their formative stages.

Parents' Roles

In addition to children, parents can be actively engaged in the documentation process. For example, parents can conduct observations on the use of space and materials in a classroom, the task persistence of individual children, and various social relationships. They can also take photographs at various times during the year to record classroom changes, three-dimensional projects, and so on, and they can summarize reading biographies, questionnaires, and other materials. In the process there is potential for parents to gain increased knowledge about schooling and to enlarge their overall contribution to their children's education. And, of course, the information has enormous potential for the classroom teacher.

In Sum

The foregoing suggestions, as mentioned earlier, are hardly meant to be all-inclusive; they ought to indicate that the means for evaluation are accessible if teachers organize their resources for such a purpose. Teachers need only decide what kinds of records they want to maintain—recognizing, of course, that they can't do everything in any one year.

The outcome of engaging in alternative processes such as those suggested is the establishment of a basis on which individual teachers and schools can improve the quality of their efforts. This, after all, is what evaluation must do to have any meaning, and it is what many of us wish to foster.

A Closing Note

What is very clear as I bring this fastback to a close is that so much needn't have been left out. I can anticipate questions but won't be close enough to the readers to respond to them. For many of the readers, the content will appear radical; to others it will appear conservative. I have, however, attempted to produce a moderate statement, one that will encourage discussion and promote an examination of tests, testing practices, and test uses. If the fastback serves such purposes, my objectives will have been met.

Bibliography

Allen, Virginia. *What Does a Reading Test Test?* Philadelphia: Temple University College of Education Monograph, 1974.

American Psychological Association. *Standards for Educational and Psychological Tests and Manuals.* Washington, D.C.: American Psychological Association, 1972.

Anastasi, Anne, ed. *Testing Problems in Perspective.* Washington, D.C.: American Council on Education, 1966.

Association for Childhood Education and the National Association of Elementary School Principals. Position paper in *Childhood Education.* November 1976.

Bane, Mary Jo. *Tests and Testing.* McClean, Va.: National Council for the Advancement of Education Writing, 1974.

Bussis, Anne; Chittenden, Edward; and Amarel, Marianne. "Alternative Ways in Evaluation." *Testing and Evaluation: New Views.* Washington, D.C.: Association for Childhood Education International, 1975.

Carini, Patricia. *A Methodology for Evaluating Innovative Programs.* N. Bennington, Vt.: Prospect School, 1971.

_____. *Documentation of an Alternative Approach to Program Accountability.* N. Bennington, Vt.: Prospect School, 1972.

_____. *Observation and Description: An Alternative Methodology for the Investigation of Human Phenomena.* Grand Forks: North Dakota Study Group on Evaluation, 1975.

_____. "The Prospect School: Taking Account of Process." *Testing and Evaluation: New Views.* Washington, D.C.: Association for Childhood Education International, 1975.

Cazden, Courtney. "Hypercorrection in Test Responses." *Theory Into Practice.* December 1975.

Chittenden, E. A. and Bussis, A. M. "Open Education: Research and Assessment Strategies." *Open Education: A Sourcebook for Parents and Teachers.* Edited by E. B. Nyquist and G. R. Hawes. New York: Bantam Books, 1972.

Cohen, Dorothy and Stein, Virginia. *Observing and Recording the Behavior of Young Children.* New York: Teachers College Press, 1972.

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

Cottle, Thomas. "What Tracking Did to Ollie Taylor." *Social Policy,* July/August 1974.

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

Duckworth, Eleanor. *"Evaluation of the African Primary Science Program."* Newton, Mass.: Educational Development Center, 1970.

_____. "The Having of Wonderful Ideas." *Harvard Educational Review,* May 1972.

Dyer, Henry. "Is Testing a Menace?" *Bulletin, Ontario Council for Leadership in Educational Administration*, January 1976 (also *New York State Education*, October 1961).

_____. "Testing Little Children: Some Old Problems in New Settings." *Childhood Education*, April 1973.

_____. *The United Teacher*, April 14, 1971.

Educational Testing Service. *Assessment in a Pluralistic Society*. Princeton, N.J.: Educational Testing Service, 1973.

Eisner, Elliot. "Emerging Models for Educational Evaluation." *School Review*, August 1972.

_____. "Instructional and Expressive Educational Objectives: Their Formulation and Use in Curriculum." *AERA Monograph Series on Curriculum Evaluation*, No. 3 (1969).

Engel, Brenda. *An Evaluation of the Cambridge Alternative School*. Cambridge, Mass.: Cambridge Alternative School, 1975.

_____. *Handbook on Documentation*. Grand Forks: North Dakota Study Group on Evaluation, 1975.

Farrell, Edmund. "The Vise/Vice of Standardized Testing: National Depreciation by Quantification." National Council of Teachers of English, Spring 1976.

Fine, Benjamin. *The Stranglehold of the I.Q.* New York: Doubleday, 1974.

Garrett, Henry. *General Psychology*. New York: American Book Co., 1955.

Goslin, David. *Teachers and Testing*. New York: Russell Sage Foundation, 1967.

_____. *The Search for Academic Ability: Standardized Testing in Social Perspective*. New York: Russell Sage Foundation, 1963.

Green, Robert. "The Awesome Danger of Intelligence Tests." *Ebony*, August 1974.

_____. "Tips on Educational Testing: What Teachers and Parents Should Know." *Phi Delta Kappan*, October 1975.

Hawes, Gene. *Educational Testing for the Millions: What Tests Really Mean for Your Child*. New York: McGraw-Hill, 1974.

_____. "Testing, Evaluation and Accountability: Managing Open Education." *Nation's Schools*, June 1974.

House, Ernest. *School Evaluation: The Politics and Process*. Berkeley: McCutchan, 1973.

Kagan, Jerome. "The I.Q. Puzzle: What Are We Measuring?" *Inequality in Education* 14 (1973).

Kamin, Leon. *The Science and Politics of I.Q.* New York: John Wiley, 1974.

Karier, Clarence. *Shaping the American Educational State*. New York: Free Press, 1975.

Kendrick, S. A. "The Coming Segregation of Our Selective Colleges." *College Board Review*, Winter 1967.

Kohl, David. "The I.Q. Game: Bait and Switch. A Review Essay." *School Review*, August 1976.

Levine, Murray. "The Academic Achievement Test: Its Historical Context and Social Functions." *American Psychologist*, March 1976.

Lindeman, Richard. *Educational Measurement.* New York: Scott, Foresman, 1967.

McClelland, David. "Testing for Competence Rather Than Intelligence." *American Psychiatrist*, January 1973.

McDonald, James. "An Evaluation of Evaluation." *The Urban Review*, September 1973.

McKenzie, Moira and Kernig, Wendla. "Evaluating Learning." *The Urban Review*, Spring 1976.

Maddan, Richard; Gardner, Eric F.; Rudman, Herbert; Karlsen, Bjorn; and Merwin, Jack C. *Stanford Achievement Test Manuals*: Part I, "Teacher's Directions for Administering"; Part II, "Norms Booklet"; Part III; "Teacher's Guide for Interpreting"; Part IV, "Administrator's Guide"; Part V, "Technical Data Report." New York: Harcourt Brace Jovanovich, 1973.

Meier, Deborah; Cook, Ann; and Mack, Herb. *Reading Tests, Do They Help or Hurt Your Child?* New York: Community Resources Institute, 1973.

Mercer, Jane. *Labelling the Mentally Retarded.* Berkeley: University of California Press, 1972.

Messick, Samuel. "The Standard Problem: Meaning and Values in Measurement and Evaluation." *American Psychologist*, October 1975.

National Association for the Advancement of Colored People. *Report on Minority Testing.* New York: National Association for the Advancement of Colored People, May 1976.

National Council of Teachers of English. *Uses, Abuses, Misuses of Standardized Tests in English.* Urbana, Ill.: National Council of Teachers of English, 1974.

_____. *Common Sense and Testing in English.* Urbana, Ill.: National Council of Teachers of English, 1975.

National Education Association. *Report of the NEA Task Force on Testing.* Washington, D.C.: National Education Association, July 1975.

Nunnally, J. C. *Psychometric Theory.* New York: McGraw-Hill, 1967.

Olson, Ruth Ann. *Internal Evaluation Techniques for Teachers.* Minneapolis: Minneapolis Public Schools, 1976.

Parlett, M. and Hamilton, D. "Evaluation as Illumination: A New Approach to the Study of Innovatory Programs." *Occasional Paper No. 9*, Center for Research in the Educational Sciences, University of Edinburgh, 1972.

Patton, Michael. *Alternative Evaluation Research Paradigm.* Grand Forks: North Dakota Study Group on Evaluation, 1975.

_____. Pattun, Michael. "Understanding the Gobble-dy-gook: A People's Guide to Standardized Test Results and Statistics." *Testing and Evaluation: New Views.* Washington, D.C.: Association for Childhood Education International, 1975.

Pidgeon, Douglas. *Evaluation of Achievement.* New York: Citation Press, 1972.

Shapiro, Edna. "Educational Evaluation: Rethinking the Criteria of Competence." *School Review*, August 1973.

Silvaroli, Nicholas. *Classroom Reading Inventory.* 2d ed. Dubuque, Ia.: W.C. Brown, 1973.

Stake, Robert. "Testing Hazards in Performance Contracting." *Phi Delta Kappan*, June 1971.

———. "The Countenance of Educational Evaluation." *Teachers College Record*, April 1967.

Termin, Lewis. *The Measurement of Intelligence*. Boston: Houghton Mifflin, 1916.

Thomas, Lawrence, ed. *Philosophical Redirection of Educational Research*. The Seventy-First Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1972.

Thorndike, R. L., ed. *Educational Measurement*. Washington, D.C.: American Council on Education, 1971.

Thorndike, Robert and Hagen, Elizabeth. *Measurement and Evaluation in Psychology and Education*. New York: John Wiley, 1962.

Tobier, Arthur, ed. *Evaluation Reconsidered*. New York: Workshop Center on Open Education, 1973.

Tyler, R. W., ed. *Educational Evaluation: New Roles, New Means*. The Sixty-eighth Yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1969.

Tyler, Ralph and Wolf, Richard, eds. *Crucial Issues in Testing*. Berkeley: McCutchan, 1974.

United States Office of Education. *A Procedural Guide for Validating Achievement Gains in Educational Projects*. No. 2 in a series of monographs on evaluation in education, 1976.

Venezky, Richard L. *Testing in Reading: Assessment and Decision Making*. Urbana, Ill.: National Council of Teachers of English, 1973.

Wallach, Michael. "Tests Tell Us Little About Talent." *American Scientist*, January/February 1976.

Weber, George. *Uses and Abuses of Standardized Testing in the Schools*. Washington, D.C.: Council for Basic Education, 1974.

White, Sheldon. "Social Implications of I.Q." *National Elementary Principal*, March/April 1975.

Williams, Robert. *Position Paper on Standardized Testing and Evaluation of Potential Among Minority Group Members*. American Personnel and Guidance Association, March 1975.

Zasloff, Barbara. *A Bibliography on Bias in Intelligence and Achievement Testing of Children and Youth*. Grand Forks: North Dakota Study Group on Evaluation, 1975.

This book and others in the series are made available at low cost through the contributions of the Phi Delta Kappa Educational Foundation, established in 1966 with a bequest by George H. Reavis. The foundation exists to promote a better understanding of the nature of the educative process and the relation of education to human welfare. It operates by subsidizing authors to write booklets and monographs in nontechnical language so that beginning teachers and the public generally may gain a better understanding of educational problems.

The foundation exists through the generosity of George Reavis and others who have contributed. To accomplish the goals envisaged by the founder, the foundation needs to enlarge its endowment by several million dollars. Contributions to the endowment should be addressed to the Educational Foundation, Phi Delta Kappa, Eighth and Union, Box 789, Bloomington, Indiana 47401. The Ohio State University serves as trustee for the Educational Foundation.

You, the reader, can help us improve the PDK foundation publications program. We invite you to comment on the strengths and weaknesses of this fastback. Let us know what topics you would like us to deal with in future fastbacks. Address Director of Publications, Phi Delta Kappa, Eighth and Union, Box 789, Bloomington, Indiana 47401.

All 94 titles can be purchased for $33 ($27.50 for Phi Delta Kappa members).

Any six titles $4 ($3 for members); any eight titles $5 ($4 for members). Single copies of fastbacks are 75¢ (60¢ for members).

Other quantity discounts for any titles or combination of titles are:

Number of copies	Nonmember price	Member price
10- 24	48¢/copy	45¢/copy
25- 99	45¢/copy	42¢/copy
100-499	42¢/copy	39¢/copy
500-999	39¢/copy	36¢/copy
1,000 or more	36¢/copy	33¢/copy

These prices apply during 1977. After that, they are subject to change.

Payment must accompany all orders for less than $5. If it does not, $1 will be charged for handling. Indiana residents add 4% sales tax.

Order from PHI DELTA KAPPA, Eighth and Union, Box 789, Bloomington, Indiana 47401.

The fastback titles now available are:

1. Schools Without Property Taxes: Hope or Illusion?
2. The Best Kept Secret of the Past 5,000 Years: Women Are Ready for Leadership in Education
3. Open Education: Promise and Problems
4. Performance Contracting: Who Profits Most?
5. Too Many Teachers: Fact or Fiction?
6. How Schools Can Apply Systems Analysis
7. Busing: A Moral Issue
8. Discipline or Disaster?
9. Learning Systems for the Future
10. Who Should Go to College?
11. Alternative Schools in Action
12. What Do Students Really Want?
13. What Should the Schools Teach?
14. How to Achieve Accountability in the Public Schools
15. Needed: A New Kind of Teacher
16. Information Sources and Services in Education
17. Systematic Thinking About Education
18. Selecting Children's Reading
19. Sex Differences in Learning to Read
20. Is Creativity Teachable?
21. Teachers and Politics
22. The Middle School: Whence? What? Whither?
23. Publish; Don't Perish
24. Education for a New Society
25. The Crisis in Education is Outside the Classroom
26. The Teacher and the Drug Scene
27. The Liveliest Seminar in Town
28. Education for a Global Society
29. Can Intelligence Be Taught?
30. How to Recognize a Good School
31. In Between: The Adolescent's Struggle for Independence
32. Effective Teaching in the Desegregated School
33. The Art of Followership (What Happened to the Indians?)
34. Leaders Live With Crises
35. Marshalling Community Leadership to Support the Public Schools
36. Preparing Educational Leaders: New Challenges and New Perspectives
37. General Education: The Search for a Rationale
38. The Humane Leader
39. Parliamentary Procedure: Tool of Leadership
40. Aphorisms on Education
41. Motivation, American Style
42. Optional Alternative Public Schools
43. Motivation and Learning in School
44. Informal Learning
45. Learning Without a Teacher
46. Violence in the Schools: Causes and Remedies
47. The School's Responsibility for Sex Education
48. Three Views of Competency-Based Teacher Education: I Theory
49. Three Views of Competency-Based Teacher Education: II University of Houston
50. Three Views of Competency-Based Teacher Education: III University of Nebraska
51. A University for the World: The United Nations Plan
52. Oikos, the Environment and Education
53. Transpersonal Psychology in Education
54. Simulation Games for the Classroom
55. School Volunteers: Who Needs Them?
56. Equity in School Financing: Full State Funding
57. Equity in School Financing: District Power Equalizing
58. The Computer in the School
59. The Legal Rights of Students
60. The Word Game: Improving Communications
61. Planning the Rest of Your Life
62. The People and Their Schools: Community Participation
63. The Battle of the Books: Kanawha County
64. The Community as Textbook
65. Students Teach Students
66. The Pros and Cons of Ability Grouping
67. A Conservative Alternative School: The A+ School in Cupertino
68. How Much Are Our Young People Learning? The Story of the National Assessment
69. Diversity in Higher Education: Reform in the Colleges
70. Dramatics in the Classroom: Making Lessons Come Alive
71. Teacher Centers and Inservice Education
72. Alternatives to Growth: Education for a Stable Society
73. Thomas Jefferson and the Education of a New Nation
74. Three Early Champions of Education: Benjamin Franklin, Benjamin Rush, and Noah Webster
75. A History of Compulsory Education Laws
76. The American Teacher: 1776-1976
77. The Urban School Superintendency: A Century and a Half of Change
78. Private Schools: From the Puritans to the Present
79. The People and Their Schools
80. Schools of the Past: A Treasury of Photographs
81. Sexism: New Issue in American Education
82. Computers in the Curriculum
83. The Legal Rights of Teachers
84. Learning in Two Languages
84S. Learning in Two Languages (Spanish edition)
85. Getting It All Together: Confluent Education
86. Silent Language in the Classroom
87. Multiethnic Education: Practices and Promises
88. How a School Board Operates
89. What Can We Learn from the Schools of China?
90. Education in South Africa
91. What I've Learned About Values Education
92. The Abuses of Standardized Testing
93. The Uses of Standardized Testing
94. What the People Think About Their Schools: Gallup's Findings

See inside back cover for prices.