Deborah Meier

Why Reading Tests Don't Test Reading

Hard-headed realists tell us, much as we may wish otherwise, standardized tests prove that millions of America's schoolchildren (approximately half) are failing to learn even the basic educational skills. Despite all the efforts of the Great Society (Head Start, Titles I-IV) to educate disadvantaged children, they are still as far behind as ever. The hard data—standardized test scores—suggest that the goals themselves were impossible and the efforts to achieve them even harmful.

Standardized tests, then, are central to the current attack on the idea of democratic schooling. For all their seeming scientific sophistication and objectivity, such tests are based on a set of unexamined value judgments and social biases. As long as we rely on such tools to tell us what our children can and cannot do, we will indeed have to concede the argument to the pessimists: ordinary people cannot be taught the cognitive skills necessary for making complex decisions. In short, democracy becomes impossible.

To make matters worse, acceptance of this pessimism leads school systems to adopt ever narrower and more trivial educational objectives and these, whether or not they are met, leave most children unprepared to make important decisions. In this vicious cycle, even the optimists get caught. That is why the nature of testing, and the quality of this "hard data," cannot be left to the psychometricians. It is a matter of large social concern.

An Autobiographical Account

I GREW INTERESTED IN TESTS rather slowly and reluctantly. Having been a good "testee," I tended to accept tests smugly. And having a general aversion to conspiracy theories, I was offended at the notion that such expertly put-together instruments could be seriously flawed. In fact, during the 1960s in Chicago, I supported an effort to sneak test results out of that city's Board of Education in order to demonstrate publicly the inferiority of the education offered to the black community—by exposing their poor test scores. We were right about discrimination, but we were wrong in thinking the tests were a useful tool for documenting the failure of our schools. It's painful today to see the very same testing patterns used to demonstrate the pointlessness of providing additional funds for these same neighborhood schools. The trouble was that we didn't look carefully at the tests, the way children handled them, the assumptions that lay beneath their development, or the manner in which scores were derived.

The first time I bothered to look at a test with any critical scrutiny was only when my own son did badly on one. I was puzzled by his poor reading score at the beginning of the third grade. I refused to be alarmed when his teacher tried to steer him into a tutor's arms, since this son was an early and avid bookworm. The following spring, before the next test was administered, I decided to go over a sample test with him. I noticed that he placed the palm of his hand over the reading passages as he set about selecting the right choices. Surprised, I asked him why. "There wouldn't be any point in the test if you could look back at the passage," he replied. The measure of "good reading," he assumed, was to see how much you retained after you fin-

ished reading. He thought this was eminently the "fairest" way to proceed and I had a hard time persuading him to shift his technique for the mere value of getting a higher score.

That same year I began teaching kindergarten in Harlem. I was full of confidence that I could turn my students into good readers. All I needed was time. I persuaded the principal to create a kindergarten-through-third-grade wing of four classes that I could oversee for four years.

When the first group reached the age at which New York City starts testing (second grade), I was confident that we would have demonstrated the advantage of our methods. I was startled to discover that children who I thought to be reading well did not, in fact, generally score well. For the first time I was forced to look the tests over to see what was happening.

A Look at Children Taking Tests

I PROCEEDED to look over the tests in what was, apparently, an unusual way. I tape-recorded several hours of interviews with young children in which we went over parts of the test together. (The results were published in *Reading Failure and the Tests*, in 1973. See Note [2] p. 466.) I discovered that there were students who could read the stories with apparent oral fluency and discuss the contents with me intelligently, but still got the answers wrong. Reading skill usually helped—but it was far from sufficient. Children got the wrong answers for a variety of reasons, often reasons that demonstrated both their intelligence and their determination to invest reading with meaning.

I realized that I had a kind of intuitive sixth sense that had made me a good testee. Pressed for the first time to justify my right selections against the logic of an intelligent and self-confident seven-year-old, I felt less sure of myself.

Sometimes their answers were based on experiences, interpretations, and associations that would not have occurred to me, yet that seemed eminently reasonable. Here are some examples * :

> Some days I should stay in bed. Today was one of those days.
>
> "Good morning," Mom said. "Don't you have a clean shirt to wear? That one looks dirty."
>
> "Sam," said Dad, "your shoes are on the wrong feet."
>
> I got dressed all over again. By the time I ate breakfast, my cereal was soggy. Then I stopped, as usual, for Bill. He was not home. He had already gone to school. I walked there alone.
>
> When I got to school, Bill yelled, "Here comes Sam, the snail."
>
> Why was Sam so slow in getting to school?
> ☐ He overslept.
> ☐ He had to get dressed twice.
> ☐ He fooled around.
> ☐ He did not like school.

Most of the children with whom this question was discussed said "Sam was fooling around." A few children said "He probably didn't like school; that's why he was fooling around and his friends called him a snail." The correct answer is: **He had to get dressed twice.**

> To *keep* means to_____ ☐ carry ☐ hold

After the test, Derrick said, "When I want to **keep** something, I **carry** it."
"No," said Yvette, "when I want to **keep** something, I **hold** it."

• There were others who arrived at different answers because they were clearly still thinking in what Jean Piaget calls "the concrete operational stage"—using a kind of logic that the test is not designed to tap or accept:

---

* Text of the tests, and authors' comments (in contrasting type), are excerpted from *Standardized Reading Tests*, a pamphlet by Ann Cook and Deborah Meier, Published by the North Dakota Study Group on Evaluation (Grand Forks, N.D.: University of North Dakota, Center for Teaching and Learning).

An architect's most important tools are his—
(E) pencil and paper
(F) buildings
(G) ideas
(H) bricks

Most children selected (E) pencil and paper. Was it "reading ability" or something else that led others to the "right" answer, (G) ideas?

A *giant* is always (E) huge (F) fierce (G) mean (H) scary

The correct answer is **huge**. But if children pick **scary**, does that mean they are "wrong" or that they can't read **huge, fierce**, or **mean?** or **giant?**

• Still others arrived at the wrong conclusion because they were engaged in a somewhat different task (like my son), due to a misunderstanding of the directions.

Choose the word that *best* completes each sentence.
A sage individual is
5. touchy 7. testy
6. old 8. wise

Mark selected "old" to complete the sentence. The testmaker's answer was "wise." What was actually wanted was a *synonym* though this was never stated in the instructions.

• Some select the wrong answer because they simply do not possess a specific piece of information that is not generally assumed to be part of the curriculum:

The frequency of a sound determines its—
(1) treble (3) pitch
(4) volume (4) harmony

• I tried another strategy: I read the passages and the questions aloud. I was surprised to realize that many children did not do any better now that they were entirely relieved of the task of "reading" and only required to "comprehend." Their difficulty lay not in the "reading qua reading"! Rather they still were dealing with difficulties that related either to unfamiliarity with content, the abstract reasoning required, or difficulty in deciphering the meaning of the task itself.

It struck me that it was much trickier than I had supposed to test the skills of "reading." Reading involves content—always a *particular* subject matter. If I took a test—such as the State of New York's Regents' examination in physics—and got a low score, the assumption would not be that I needed remedial reading. Most people might suggest I take a course in physics. Yet the New York Physics Regents' exam is actually a kind of reading test. One reads passages and then selects from among multiple-choice answers—though we recognize that knowledge of physics is essential to dealing with this reading material. We know also that the New York Regents' test presupposes a particular kind of background in physics (it helps if one has been schooled in the areas on which this test focuses)—particularly the terminology of the testing instrument.

The apparently simple paragraphs on even the second- or third-grade reading tests also have their subject matter and a certain technical terminology to describe that subject matter. They also have a particular syntactical style that may or may not be familiar to the testee. Finally, they demand skill in taking tests, a technical subject in itself.

The test my seven-year-olds were measured by that spring included passages that required knowing that the terms "English" and "British" referred (inaccurately!) to the same peoples; that a tree trunk had "rings" inside (and not the kind you wear on your finger); that oceans have tides (and that they go in and out every 12 hours); that penicillin is a drug; that 1865 is 100 years before 1965—and other facts that are unlikely material for a sound second-grade curriculum (unless the teacher has peeked at the test ahead of time)!

Even when children were not baffled by the subject matter, or not led astray by false but

clever guesses at unfamiliar terms, they still often missed the testers' logic for selecting one particular answer over all others. For reasons I did not yet understand, there was always at least one "almost right," "misleading," or "tricky" selection to distract the unwary testee from the "right" answer.

This pattern was so clear that I began to suspect that it couldn't be just poor test construction. Rereading Banesh Hoffmann's brilliant book *The Tyranny of Testing* (New York: Collier, 1964), I noticed that at a moment of surprising honesty, one publisher admitted that the correctness of an answer was not actually the fundamental criterion. This seemed outrageous, but correctible.

Impact on School Reform

SO I BECAME A CRUSADER and mini-expert. I read about tests, worked to improve their content and eliminate ambiguity and bias, to explain their limitations to parents and teachers—and to demand alternative methods of reporting data and gathering information. I found myself growing angrier about the subject and probably shriller.

A new promising movement in early childhood education, which respected intelligent activity and aimed at an intellectually purposeful curriculum, was derailed. Instead, we were being pushed into programs aimed at ever narrower and more trivial subskills that, it was hoped, would show up quickly on paper-and-pencil tests. The heart of a good education—respect for subject matter and intellectual inquiry—was sidetracked in favor of prescribed kits filled with disguised reading tests, hundreds of one-paragraph reading "tasks" followed by multiple choices. Schools became coaching institutions preparing their subjects for the tricky world of tests (the "real world"). If test-coaching had actually also been a good way of making children into readers, it might have been less disastrous. But since "good reading" requires knowledge about the world, a narrow focus on testing "skills" deprives children of the substance of literacy in a quest for good scores. Furthermore, many of the strategies that seem to support good test-taking are really counterproductive to those we need when we read competently.

A good reader, for example, must take a lot of risks in the process of reading most material. These risks lead to errors. But the importance of these errors, and the extent to which they are even "noticed" and thus corrected, depends upon their impact on meaning. Redundancy and involvement in the meaning of the material generally corrects the errors that good readers make as they proceed fairly rapidly through the written page. But reading tests, which involve fairly short and generally pointless passages followed by trick questions and answers, require being constantly alert to precisely the kinds of errors that are unimportant in most real reading. Also, the heavy emphasis on phonics, syllabification, and pronunciation in most tests for young children require schools to overemphasize slow oral reading, the conscious mastery of phonic rules and word-perfect, pronunciation-perfect reading. Thus regardless of whether a child is in need of such help or getting it, and despite evidence that an emphasis on such skills may reduce many children's speed, fluency, and capacity to read silently—such instruction is doggedly pursued *because it is needed for the tests.*

Test Scores

THE BIGGEST IRONY of all is that none of this can lead to its promised end: an overall improvement in the test scores of our nation's children, or even an upward shift by any substantial subgroup of children within the school system.

Newspaper editors, school boards, and groups of aroused citizens clamor for higher standards and an end to the high rate of illiteracy. Despite millions in federal subsidies, the number of ill-educated children remains constant, they say, and is still increasing in low-income urban areas.

In vain, it seems, do a few educators occasionally point out that by definition scores on these kinds of tests can never change nationwide. Half the students must always be

labeled "above" and half "below" grade level. It's as pointless to demand that most students score in the top half as it would be to insist that all the teams in the American League make it into the playoff, or have won/lost records of better than 0.500. A sportswriter making such a silly demand would be at the end of his career. The same does not apply to those who write about education. Even the more "modest" demand that all high-school graduates be brought up to a "tenth-grade level" is statistically nonsense. For what does it mean to say a twelfth-grade student is reading at a "tenth-grade level"? It means that a student's score is at a point along the normal curve labeled by the test-makers "tenth grade." A "tenth-grade level" is merely a percentile point. Although different test publishers put grade levels at somewhat different places on the curve, "two years below grade level" is always still a fixed percentile (generally around the 25th percentile). It does not, in fact, mean the student reads like a tenth grader. The percentage of students on the "tenth-grade level" is as fixed in stone on any particular test as is the midpoint. For 25 percent of those tested must always be in the top 25 percent, and 25 percent in the bottom 25 percent.

Let us not forget that the scores on such tests are only stand-ins for rank-order place. If we greatly improved our techniques for teaching either reading or testing, the test-makers would in turn either have to change the scoring system so that more right answers were needed to obtain the same "score" or develop a new and more difficult test. A more difficult test might consist of one with more difficult reading material, more difficult questions, or simply trickier questions with subtler nuances.

As long as we have both tests and a scoring system designed to produce a rank order, there will not be and cannot be a rise in the reading scores of all the students of the United States. Everyone can and should read better, be better educated, and have a larger fund of knowledge. But our measuring instrument guarantees that, through it, we can never know if we are achieving that goal.

How Is a Test Designed?

IF A TEST SCORE IS simply a statement about a particular child's rank order, it is only as accurate as the test-makers' ability to predict how all the rest of the children in the country would respond to the same test. The test-makers must figure out accurately how to design items that will conform to such a prediction. They must figure out which items will be answered wrong by which particular kind (group, category) of children. Only in this manner could the test-makers live up to the guarantee that their sample (the population they used to "norm" the test in the process of development) will reflect the population as a whole.[1] For this purpose they need items that are not too easily susceptible to teaching or coaching, since that would quickly cause a test to be out-dated and its predictions to be incorrect. This requires a complex and arduous task that takes publishers five to seven years of conscientious labor to produce. The amount of technical skill, money, and effort spent is considerable.

As a result of this effort, the test-makers know, before the test is actually put into use in our classrooms, what kinds of children will get what particular kinds of items wrong, and what kinds of children will get them right. The test-makers know that their test is relatively "teacher-proof," and that certain children from certain population subsamples will have trouble of a predictable sort with the items the test-makers have selected, while others will not. This requires a test that is as impervious as possible to the impact of schooling. The test-producers are staking their reputations and investments on their predictions.

To accomplish this, the test-makers need items that almost certainly will discriminate between children in the top 1 percent and the other 99 percent, the top 5 percent and the other 95 percent, and so forth. As I tell children before I begin a test these days, "Remember, they went around the country asking kids your age questions. They had to find some questions that they don't expect you to know the answer to. So don't be upset when you

come upon them." What matters is not the absolute "fairness" or even "accuracy" of the right choices, but only the test-makers' ability to distinguish certain types of readers from other types and in so doing to produce a predetermined scoring pattern known as the normal curve.[2]

For this to work, one needs to start with fairly clear and preconceived ideas about who the "good" readers are. Most tests use other tests—both other reading tests *and* IQ tests—as their external standard or validation. Given this background, it begins to be clear why the passages and the right answers often seem "unfair" to my East Harlem pupils. (It's the same process used by the developers of early IQ tests who used their subjects' occupations as the external standard or criterion of intelligence.) This explains why the test-makers do not select those terms or associations that would create a built-in bias favoring the language or experience of my Harlem students over that of children of highly educated and powerful parents.

Where Are the Built-in Prejudices?

A CLOSE EXAMINATION of the tests shows many examples in which choice of vocabulary, subject matter, style of language, metaphors selected, word and idea associations, and values presuppose a certain kind of social and personal history. I'm good at guessing the right answers because I have been steeped in the assumptions upon which the test-makers build their system. I do not even have to stop and imagine what the test-makers want me to think, feel, or say. The mark of an easy test is precisely that it does not require us to call upon consciously learned data; the right answer seems obvious, like "common sense."

When I was preparing for the National Teachers' Exam, I recognized that since my particular educational experiences might not coincide with the test-makers', I would have to spend time looking over old tests so that I could get a better handle on the biases of the NTE. For a young child this kind of conscious distancing and abstracting is virtually impossible. At any age it imposes a handicap on those who are "outside" the culture that the test is rewarding.

This explains why we are right to advise certain children to give their first and unthought-out response to multiple-choice items—"don't stop and think deeply," we tell them. Quick intuition, not deep thought, is the best guide—but only if you're part of the subculture that the test was designed to reward. It's bad advice for equally logical and intelligent children whose intuition will lead them astray. Such children need to exercise the kind of intellectual caution I had to rely on with the NTE—a process that at the very least slows them down.

When the issue of test bias was first raised, the test-makers responded by obligingly changing rural stories to urban ones, coloring their stick figures black, and eliminating some archaic terms (for instance, spectacles). But our victory was irrelevant. For the tool *requires* a bias, and it does so for reasons that are not deliberately mean or biased.

Let us consider a typical example of such subtle bias. In a recent test the following story was presented:

> The children lived on a pleasant tree-lined street. One morning trucks came and chopped down the trees in order to widen the road for a new 4-lane highway. When spring came, the birds and squirrels, who used to live in the tree, did not come back.

The test-makers asked the seven-year-old testees this question:

> When the truck came the children were . . .
> (a) excited (b) angry (c) sad (d) away

How was it that I immediately knew the right answer was (c) "sad"? And that the teachers and graduate students at a prestigious college to whom I posed this question also knew the right answer?

In real life, of course, children might be excited, or they might be angry when they see those trucks. But *no* child any of us has ever met—rich or poor, black or white, urban, suburban, or rural—is actually likely to feel

"sad." Yet somehow "we" all knew that we were supposed to feel sad when the environment was damaged for modernization, and thus we intuitively picked (c). If we had read more carefully, we would perhaps have hesitated. Fortunately, experience has taught us to trust our intuition.

Indeed, in trying this item out on children in my school it "works" much as the test-makers intended. It may differentiate the good readers from the poorer readers, but it mainly differentiates children with one background from those with another.

We can get angry at the test-makers for this kind of question. We can get them to eliminate the grossest examples. But this only forces them to find ever subtler ways of doing the same thing. Instead, we can recognize that items such as this do indeed "work" for the purposes of this kind of test. They discriminate "appropriately." It's the test that is inappropriate for our purposes.

No wonder these tests correlate so highly with standardized IQ tests, as all publishers' manuals boast. For IQ tests are strewn with precisely such questions, which test children's knowledge about what the "society" expects of them,[3] rather than their knowledge about the reality they know, or the way in which they think through problems.

While all children, for example—rich or poor, black or white—act pretty much the same way when they lose a friend's ball, what they think they are supposed to do and feel differs. In most reputable IQ tests for young children this kind of sociological information is put to use. Children are seen as more intelligent if they offer to pay for the ball, less so if they say they're sorry, and dumb if they say they would tell their mother.

The way children answer this question tells us little about what they would really do, or how well they can reason. Low-income children in fact may well be less likely to tell their mother than middle-class children. What it does tell us is what the adults in the child's world tell children they "should" do and what children think the adults who are administering the test want them to say. Indeed, one of the characteristics of lower-class parents and of teachers in lower-class schools is the constant reiteration of this "tell the grown-up" refrain; do not trust your own judgment! It's hardly surprising that so many of these children give us the answer they think, based upon their intelligent internalizing of experience, we wish to hear.

Items that my Harlem students would get right, because here their experiences or their social values (or what they think are the school's values) give them an edge, would be thrown out in the years of test development as being poor "discriminators."

The tests reflect the bias of reality, we are told, and this is a reality all children have to live with.

In a sense, of course, this is true. For the knowledge base, syntax, terminology, and implicit value system of the test are those we euphemistically label "mainstream." As such, they are the yardstick. And if, given this yardstick, those who are not in the mainstream (or better still, in the elite stream) are at a disadvantage in such tests, this is an unfairness built into life, not the fault of the tests. Tests are not, we are reminded, in business to change the world but merely to reflect it. They are charged with predicting who will do well and who will not, given the world as it is and as it probably will remain. The educational tests are designed, like their famous cousins the IQ tests, to be predictive, to select items that roughly correlate with future success.

Reading tests require a certain ability to decode visual symbols. But a child who can decode visual symbols won't necessarily do very well on this kind of test, for the test differentiates the child "neutrally" from other equally competent "decoders" on the basis of their relative ability to handle a certain esteemed culture quickly, intuitively, and consistently. *Remember! The better the test, the less susceptible it is to the impact of teaching and schooling.* That's what the creators of the IQ tests promised us to start with—an examination that would test what could not be influenced by do-good environmentalists. And that, too, is the technological model for the reading test, as well as its validation.

How odd indeed to create a measure for differentiating good from bad teaching based on a model that, by definition, is intended at its best to be impervious to all teaching!

My rather sweeping suggestion that our method of assessment is a fraud may seem presumptuous. I'm inclined by nature to try to state it more cautiously and with more hedges. Yet the case seems to be clear, documentable, and thoroughly damaging. If I'm correct, we cannot even determine the utility of any particular school reforms or set of reforms if we use such tests as our yardstick. Furthermore, we have distorted these very reforms in an effort to make them look good by meeting the criteria we unwittingly accepted for "success."

Efforts to produce models of education that can demonstrate the educability of the vast majority will founder similarly unless we have more legitimate ways to assess success and failure. The issue of educational evaluation and assessment indeed is critical and cannot be left to the industry that now both produces the technology and the "experts," and then sells the "curriculum" materials that promise to produce better results!

What's the Alternative?

AS LONG AS WE REMAIN wishy-washy in our criticism of these tests and seek mere improvement and modification, alternatives will not be produced. Our alternatives will be rejected because they cannot meet the myriad purposes for which present tests are used, and they cannot be as cost-efficient nor as easily administered.

If the present tests were benign—somewhat misused, or overused, in short, merely in need of improvement—such criticism of alternative means of assessment would be just. But if we really want to find out about how Americans read, including comparative data on different subgroups in the population, we shall have to design a different kind of testing with that purpose alone in mind. It should not be so very difficult. It would require some in-depth, "standardized" interviewing of a relatively small sample of the population. It would be interesting; but as of now we have incredibly sparse data on this most important topic.

If we want to find out what teachers and parents can do to assist children in improving their reading, we will have to seek data on the ways a child is learning. For these purposes, we must frankly admit that we have only the most imprecise skills for making assessments. The tools for measuring what is going on in a child's head are neither accurate nor scientifically objective. The best tool around is still the old-fashioned one of our grandparents: sitting down and listening to a child read and then talking together afterward. The only thing that has improved is our knowledge of what to listen for and how to recognize the array of meanings behind a child's common errors.

Parents and teachers need assistance in seeing and hearing the diagnostic richness that can come from structured reading encounters or interviews. The work of Paul and Yetta Goodman, the investigations of psycholinguist Frank Smith, the interviewing methods and insights of Jean Piaget—all these can help us make sense of a child's expressed views in a more systematic way than we once did. This can only be done on a one-to-one basis. It's time-consuming but it's also part of good teaching. Good listening can be informed by science, although in the end it remains an art. The art of good teaching begins when we can answer the questions our children are really trying to ask, if only they knew how to do so.

There are those who say you can't sell this approach to a public determined to have a single test that can measure success/failure, that can be replicated nationally, be used diagnostically, administered relatively cheaply, and that can produce a single rank order. It's hard to resist the pressure to pretend we have such a magic elixir. But it is immoral to pretend to comply. It would be as if the medical profession were to support a fraudulent common-cold pill because it was afraid to admit that at the present stage it did not possess another sure-fire cure.

If we are absolutely unable as a society to wean ourselves completely from norm-referenced tests, we might at least use them only for their claimed purpose: to compare groups, and not to make decisions about individuals. If we used the tests for that purpose, we could administer them on a random-sample basis. This would eliminate some of their worst aspects, including the impact of measurement error, coaching, curriculum distortion, and the widespread cheating that now distorts school life.[4]

We might also try reporting scores only in the manner that all test publishers and experts suggest—using either percentiles or stanines or both, rather than the alleged "grade-level" equivalence. We would thus relieve the *New York Times* of its annual lament that 50 percent of the students are either below the other 50 percent, as are 10 percent below 90 percent, and so forth.

IN THE NAME OF OBJECTIVITY AND SCIENCE, the testing enterprise has led teachers and parents to distrust their ability to know their own children; this is the greatest disservice it has done us. It's like people who need to call the weather bureau to see if it's really raining, or take marriage tests to see if they are really happy. Nothing is more harmful to a good education than this withering away of the expectation that it takes human beings to make judgments.[5]

The task of returning testing to its proper place will be difficult. And the job of eliminating those tests that serve antieducational purposes will take time. But the more we cater to the testing craze, even for short-term political and/or career ends, the more difficult it will be for us to remember what we are all about, and what is, at bottom, the problem.

Those of us who believe in educational equity are facing enormously difficult problems. We do not have all the forces of society on our side, and many are happy to find evidence for our "naivete" in regard to the educability of the majority of human beings. Our belief in democracy—that normal everyday people can make sense of their world and learn to make decisions about it—is at stake. Literacy is a major aspect of this claim. Literacy defined as more than knowing how to write one's name or spend a certain number of years within a school house remains a distant dream for most of humanity. It is a new ideal even in our own country. But doubts regarding such aspirations are rooted not in the objective data but in our present lack of will to deliver such education. This condition is aided and abetted by our acceptance of data that we either do not understand or largely misinterpret.

Nor is it surprising in a period when conservatism is on the rise to see ideologies surface that aim to prove democracy a sentimental dream. There now is arising a renewed interest in educational tracking, new trust, on the part of some academics, in the "innate" differences between classes and castes and in the expansion of programs for the gifted, new legislative proposals to support private education, and a new set of innovations that involve so-called minimum competencies and "time-on-task."[6] All these antiegalitarian trends that are now flourishing are nourished by the renewed focus on testing.

To the ideologues of the New Right, the focus on testing appears correct and proper, since the free play of market forces "naturally" produces inequality in the shape of the "normal curve." Such nonsense will someday fade. But the "normal curve," so deeply and perniciously embedded in American education through the use of standardized norm-referenced testing, will not fade as quickly. A testing mechanism that is based upon a priori assumptions about the "naturalness" of that "normal curve" is bound to produce results reinforcing such assumptions. The much-needed development of a theory and practice consistent with democratic schooling is crippled by the existing means of evaluating and documenting educational success. It will not flourish, even in better times, until we break with the dominant ideology of 20th-century psychometrics.

Notes

[1] Such tests are thus properly called "norm-referenced standardized tests." There are other forms of standardized tests—the driver's test is one, and it's not norm-referenced. Presumably everyone could pass it with flying colors. The tests we are describing are, in addition to being norm-referenced, all group-administered paper-and-pencil tests, and meet several other technical requirements with regard to their ability to predict future school success, stability, and reliability within agreed-upon limits of measurement error. The level of acceptable "error" is, for example, far too high to make decisions about individuals, although it is "acceptable" when comparing groups or making statistical predictions.

[2] The particular requirement of reputable norm-referenced tests for a scoring pattern that resembles a so-called normal curve is interesting in itself, and has its own particuar history. See two articles that take a critical look at this and other particular hypotheses with regard to "natural" traits: Stephen Jay Gould, "Jensen's Last Stand," *New York Review of Books*, May 1, 1980, and Philip Morrison, The Bell-Shaped Pitfall," in *The Myth of Measurability*, ed. Paul Houts (New York: Hart, 1977), pp. 82-89.

For some other critical background material on testing, see James Fallows, "The Tests and the 'Brightest' — How Fair Are the College Boards," *Atlantic*, March 1980. Gene Hawes, *Educational Testing for the Millions: What Tests Really Mean for Your Child* (New York: McGraw-Hill, 1964). Banesh Hoffmann, *The Tyranny of Testing* (New York: Collier, 1964), mentioned on p. 460. Deborah Meier, "Reading Failure and the Tests," occasional paper (New York: City College of New York, 1973).

[3] The "society" does its best, of course, to train different children in different ways—through the messages delivered by parents and teachers. "Low SES" children are actually taught certain "expectations" of what they ought to say and do to please adults that are different from the ways in which most "high SES" children are taught. Thus when they answer questions on IQ and other "norm-referenced tests" differently, they reflect their past experiences. There is no single standard for what "society" expects, but there is a single standard for what produces a high score on these tests. The standard used is the one reflecting "high SES" backgrounds—more or less.

[4] New York City has had cheating scandals almost every year for the past five or six years, including one in which the lower courts ruled that the evidence was sufficient to prohibit the use of any of the test data. Despite the fact that virtually all those concerned have a stake in burying such claims (teachers, supervisors, students, parents, school boards, etc.) they continue to surface. It is my belief—based upon both personal knowledge and an examination of annual score fluctuations—that cheating is so rampant today that most major city test scores are not even good "norm-referenced" data. In short, a child's score doesn't even reflect rank-order ability to handle the test material. A school or city that reports a major score rise is either testing a different population or cheating. Such cheating includes everything from coaching for a specific test to actual doctoring of tests and scores.

[5] The Fall 1980 *American Educator* reports "striking differences" in effective teaching of low vs. high socio-economic-status students. "Effective teachers of low SES pupils . . . kept instruction at a low level of complexity. . . . They tended to ask simple-answer or multiple-choice questions instead of encouraging pupils to analyze, synthesize, or evaluate." The "effective teacher" of high SES students "is more likely to discuss the student's answer than is the effective low SES teacher. The latter seldom amplifies, discusses, or incorporates the student's answer. "Low SES students," in short, "made fewer gains when teacher-student interaction was highly cognitive in nature." This conclusion is the shocking but logical outcome of basing educational decisions on test scores. It is an example of how testing not only fails to be helpful but sabotages good education.

[6] Unfortunately, some of the critics of standardized norm-referenced testing are treading a new, equally pernicious path, led by a new school of experts. This latest fad is variously called "competency" testing, mastery learning, testing through objectives, and minimum teaching objectives. All these various terms refer to schemes for breaking down all subject matter into small "teachable" bits that can easily be tested and scored, and when placed in an inevitable sequence from the "easiest" to the "hardest" will guarantee success. There may be some skills for which this approach is useful but, on the whole, this anti-intellectual and anticognitive approach, defying most of what we have learned about how human beings absorb information and organize ideas, is intended for children at the bottom end of the academic and social-economic scale. While today such methods are largely unmanageable, the advent of new technologies may indeed soon turn many classrooms into impersonal factories in which children process small bits and pieces of so-called knowledge or skill.

There is also a third group of people working on a very different paradigm for testing and evaluation. They are focusing on observation, interviews, description, and documentation as the primary tools of assessment. This approach seems the most appropriate and promising. □