Assessment and Accountability: Myths and Realities

IRIS C. ROTBERG

Remarks presented at the Philadelphia Federation of Teachers Health and Welfare Fund Conference, February, 1997

It is a pleasure to be here today and to discuss with you a set of public policy issues that are receiving a great deal of attention nationally--as well as in Philadelphia.

It seems fitting--in a talk about the perils of testing--to begin with a few examples of students' responses to exam questions:

The moon is a planet just like the earth, only it is even deader.

Three kinds of blood vessels are arteries, vanes and caterpillars.

The process of turning steam back into water again is called conversation.

Humans are more intelligent than beasts because the human branes have more convulsions.

Algebraical symbols are used when you do not know what you are talking about.

Parallel lines never meet, unless you bend one or both of them.

Let us turn now to other, but less entertaining, misconceptions. My remarks today are straightforward: Test score comparisons are highly misleading indicators of the quality of education and they are irrelevant to decisions about the wisdom of any particular school reform. I believe they are used for inappropriate purposes and they are blamed for perceived failures in our economy and in education reforms such as national standards, charter schools, and vouchers. In some school systems, testing has become synonymous with teacher accountability. And as we continue this process, we ignore real educational problems.

The justification for test comparisons is based on a set of myths. First, though, let me acknowledge at the outset that tests can be valuable when used for appropriate purposes. For example, tests have been useful in diagnosing students' learning problems, in providing teachers and students with ongoing feedback about students' progress, or in encouraging changes in curriculum and teaching methods. However, our current public policies have gone well beyond the rather narrow reality of what tests can and cannot accomplish.

I will begin with a myth which is widely accepted both by public officials and by the general public:

1. Test score comparisons between nations, between states, or between schools provide valid measures of the quality of education.

The fact is these comparisons are flawed methodologically and do not reflect educational quality in any country, state or school.

The international science and mathematical comparisons illustrate the problem. The international test scores have little to do with the quality of education. They reflect instead differences in student selectivity, in poverty levels, and in curriculum emphasis.

The basic problem is student selectivity: The more highly selected the students who take the test, the higher will be the average score. That score does not reflect the overall quality of the educational system. It simply reflects the fact that the students represented in the test comparisons have been much more highly selected in some countries than in others.

The first set of international comparisons, conducted in the 1960s and early 1970s, did not take into account the percentage of the age group actually enrolled in upper-secondary school. These attendance rates are higher in the United States than in most other countries. At the time the tests were administered, only about 20 percent of the age group in Europe attended upper-secondary school--the *highest achieving* 20 percent--compared with 80 percent of the age group in the United States. While the European attendance rates have increased substantially, they still remain lower than those in the United States.

More recent studies have tried to deal with the sampling problem by testing only those twelfth grade students who are in an academic track and taking mathematics or advanced science. These changes, however, do not address the problem. Consider, for example, the results of a recent assessment of mathematics students in Hungary and England. Hungary ranks near the top in the eighth-grade comparison. By the twelfth grade, when Hungary retains more students in mathematics than any other country, Hungary ranks among the bottom countries. Have Hungary's schools gone downhill between the eighth and twelfth grades, or is it simply a matter of more students, lower scores?

England, by contrast, scores in the bottom half in most of the eighth-grade comparisons, but ranks among the top countries by the twelfth grade, when only a highly select group of students there takes the test--indeed, students who have studied science and mathematics almost exclusively since the age of 16 in preparation for their secondary school completion examination. Similarly, in eighth-grade comparisons, Japan ranks first, with Hong Kong in the middle of the rankings. By the twelfth grade, when only three percent of Hong Kong's young people are taking mathematics (compared with 12 percent in Japan), Hong Kong comes in first and Japan second.

When a country's rank can change so dramatically between the eighth and twelfth grades, it simply shows that the test comparisons are meaningless as a measure of school quality.

I will turn now to Harold Stevenson's highly-publicized comparison of 11th-grade students in several countries: the United States, as represented by Minneapolis and Fairfax County, Virginia; Canada, as represented by the province of Alberta; China, as represented by Beijing; Taiwan, as represented by Taipei; and Japan, as represented by Sendai. Clearly, these sites are not representative of the nation as a whole, nor were the schools selected within each site necessarily representative of that site.

I will use the comparison between China and Japan to show how these flaws lead to misleading findings. China ranked first in Stevenson's study even though Japan educates a much higher proportion of its young people than does China, and many Japanese students also spend up to 20 hours a week in cram courses--in addition to their regular schooling.

The reality is that the test score rankings reflected student selectivity, not the overall performance of the education system. Like many other developing countries with scarce resources, China has an elitist education system that provides upper-secondary education to only a small proportion of its young people. While most Japanese students complete high school, a majority of Chinese students already have left school by the 11th grade. As a result, only a small proportion of the age group in China is represented in Stevenson's test results--the highest-achieving students, in the capital city, in a country with particularly wide disparities between urban and rural education.* We can understand, therefore, why the Chinese sample outscored not only American but also Japanese students. The point is the *fewer and more highly selected the students who take the test*, the higher will be the average score. The result has little to do with the quality of education.

I am sometimes asked whether I believe we can learn something from other nations' education systems or teaching practices. Of course we can. For example, Japanese teachers spend much less time in the classroom than do teachers in the United States. As a result, the teachers have considerably more time to prepare for their classes, work with each other, and work with their students. I am certain that no one here would complain about having that opportunity, but it is not likely to happen, particularly in a time of budget cuts.

Two other factors, in addition to student selectivity, also affect international test score comparisons:

First, the proportion of low-income children *in the test-taking population*. The United States, for example, has a relatively high proportion of low-income students compared to many other industrialized countries. And, unfortunately, the number of low-income

---

* Indeed, a recent study by Jianjun Wang, a researcher at California State University, showed no significant difference between U.S. and Chinese ninth-grade scores when students were selected from both urban and rural areas. While the samples clearly are more representative than those in Stevenson's study, selectivity still remains a problem because a large number of Chinese students have already left school by the ninth grade--and, therefore, are not tested.

students has grown substantially in the past decade. We tend to hold the education system responsible for our broader societal problems--whether these are major problems like poverty, drugs, crime, family breakup, or teenage pregnancy, or less dramatic problems like TV-watching or wearing earphones while studying.

And to compound these problems, we allocate the fewest resources to the children with the greatest needs. Connie Clayton put it this way in a paper she submitted for a Title I study we conducted: "We must face every day the realities of the unequal hand dealt to out children and to our schools." When we rank states on the international test score comparisons, the powerful influence of poverty and inadequate resources are all too evident. For example, the test scores of Iowa, North Dakota, and Minnesota were similar to the top-scoring countries--Taiwan and Korea. At the other extreme, Alabama, Louisiana, and Mississippi scored about the same as Jordan, the lowest-scoring country in that comparison.

Differences in curriculum emphases among nations also affect the international rankings. The test may not reflect what students have learned in school. For example, countries make different choices about the proportion of twelfth-grade students who study calculus, the degree of subject-matter specialization after age 16, and the amount of time devoted to cram courses in addition to regular schooling. The point is, the decision about whether or not to adopt a particular practice should be based on a careful consideration of the implications of the proposed change, not on rankings on standardized tests which are comparing quite different curricula, but which then take the place of public dialogue and debate.

The reasons for the ranking of a particular country differ from country to country. For example, Japan's high scores reflect the curriculum, the emphasis on mathematics, and the time spent learning the material--both in regular school and in cram courses. Japan's high scores also reflect its relatively low levels of poverty.

Selectivity continues to play a part in the relative ranking of many countries. In the most recent study (the Third International Mathematics and Science Study), for example, Israel included only Hebrew-speaking schools. Participation rates were low in the Netherlands. In Germany, the pair of grades tested was one grade higher than the international target.

The researchers conducting international studies do their best to avoid selectivity, but the real world and a statistician's sampling design may be two different matters. Twenty-two of the 41 participating countries did not meet the study's standards. And even for those that did, there were many other opportunities for selectivity to influence the rankings.

I will turn now to state level comparisons. The sampling problems found in international comparisons apply as well to the ranking of states on the SAT in the United States. The states with the *highest* proportions of students taking the SAT tend to have the *lowest* average SAT scores. The states with the lowest percentage taking the SAT have the highest test scores simply because the students are more highly selected.

The sampling problems are not limited to international and state comparisons. They occur as well when we compare schools within a school system, when we attempt to evaluate teachers based on the test scores of their students, or when we use test score gains to show that one school reform or another has "worked." Yet, increasingly, these are the purposes for which test scores are used.

Let me give some examples of what happens in comparisons between schools. Schools can raise their scores simply by excluding low-performing students. A *New York Times* article describes an elementary school that was put on probation by the state for low test scores. Within only a single year, the third-graders had made major gains. According to the article, "officials simply stopped testing most of the third-graders. Between 1988 and 1992, even though enrollment doubled, the number of third-graders tested dropped by nearly half, from 76 in 1988 to 44 in 1991. By 1992, only 28 percent of the class took the standardized test, according to documents obtained from the state and the district." Again, the point is that the more highly selected the students who take the test, the higher will be the average score. That score has little to do with the quality of the school.

Test score inflation at the college level has received widespread publicity. Many colleges, anxious to convince parents and students of their quality and selectivity, inflate the average SAT scores by excluding economically disadvantaged, remedial, learning disabled, or foreign students with limited English proficiency. The college, therefore, "looks" better.

Schools also inflate their scores by encouraging students to drop out of school before the examination, or by retaining them in grade. An educator put it this way: "I'm concerned because we have fewer students after grade 9 and it looks like it's to a school's advantage to get a kid to drop out [rather] than to keep him on the rolls and have poor test scores at grade 12."

This technique is not limited to the United States. According to a World Bank study, similar exclusions have been reported, for example, in Kenyan primary schools with high proportions of passing students. And the same study found that as many as 20 percent of Chinese students may be retained in grade in upper-middle school in order to increase that school's scores (and, therefore, its reputation) on university entrance examinations.

Studies also show significant fluctuations in test scores from year to year in schools throughout the United States. While we may not always know the reasons, we do know they are not related to changes in the quality of education. A recent study of Title I, the federal education program for disadvantaged children, found that about one-half of the schools identified as needing "program improvement," based on test scores, appeared to be doing just fine only one year later--without making *any* changes in their Title I programs. A similar statistic comes from a state that holds a test score competition: 28 percent of the schools have won the test score competition once, 11 percent have won twice, and only four percent have won three times. Is it likely that the quality of the schools changed so much from year to year that only a small portion of the schools could

retain their first place status over time, or are the test score fluctuations related instead to such factors as demographic changes, methods of testing, or measurement artifacts?

Perhaps the problem in using test scores to measure school quality is best illustrated by a BBC interview with the headmaster of a school that had ranked first--as measured by the largest test score gain from the previous year--under a national assessment in Britain. The headmaster made the point that his first place ranking meant little because it reflected only the school's very low baseline performance. He also predicted, based on his knowledge of the incoming students, that his ranking would fall the next year and then rise again the following year when he expected a particularly high-achieving group of students to enter the school. The headmaster knew well that these test score increases, and decreases, did not reflect changes in the quality of his school. Perhaps educators are more astute about these matters than are policymakers.

The emphasis on test scores I discussed has led to a second myth, that is :

2. <u>The quality of our schools has declined: That is why we are no longer "competitive</u>."

We incorrectly conclude from the flawed test comparisons that our schools, or our parents, or our students, or our scientists, or our research institutions have failed.

We add to the test scores our nostalgia for the past which leads us to overestimate the quality and rigor of education in our parents' and grandparents' generations--and even in the schools we attended. We ignore the enormous strides that have been made in educating a large proportion of the population. In 1940, 38.1 percent of 25-29 year olds in the United States had graduated from high school. By 1993, that percentage had risen to 88.2 percent. In the same time period, graduation rates from four-year colleges rose from 5.9 percent to 23.7 percent.

Our students' educational accomplishments equal and in many cases surpass those of previous years--*even measured by tests*. According to a recent study by RAND, student mathematics and reading performance improved *for all racial and ethnic groups between 1970 and 1990*.

None of these statistics suggest that our schools do not need improvement. Clearly, the United States faces serious educational problems, which I will turn to in a moment--but they are not the problems identified by the current preoccupation with test score comparisons. A tendency to overstate--or to misstate--the problem is not new. In the 1950s, we responded to Sputnik by blaming the schools for our perceived inferiority to the Soviet Union in science and technology. When (much later) we realized that perhaps we could hold our own in that competition, we turned to another concern: an imminent shortage of scientists and engineers, predicted to occur in the 1990s--again due to the failures of our education system. We have not heard much about these predicted shortages now that we are well into the 1990s, perhaps because many new Ph.D.'s are having a difficult time finding jobs.

Let me turn now from the inaccurate assumption that our schools have declined to a third myth:

3. We can "fix" our schools by administering more tests.

Or, put another way, if we hold teachers accountable for students' standardized test scores, our schools will improve. This myth is reminiscent of the theme of the movie, *Field of Dreams* (translated into educational terms): Build a test and they will learn.

I am afraid that just the opposite has occurred. An emphasis on multiple choice standardized tests encourages the teaching of a narrow set of measurable skills that often have little to do with what educators and parents value most. In the United States, the mandated tests--and the rote learning associated with them--are particularly common in classrooms with high proportions of low-income and minority children.

Don Reeves, President of the D.C. Board of Education, put it this way in a recent *op ed* piece entitled, "Why D.C. Schools are Abysmal:"

"The obsessive focus on standardized test scores keeps expectations low because the tests themselves expect so little and fail to measure the important things. Nevertheless, test scores have become the coin of the realm because the media and the city's elites naively (or cunningly) accept these tests as legitimate measures of student performance. The school system strives to placate the press and city leaders by engineering minuscule rises in the scores; the teachers are prodded to focus on these minimal skills and therefore have less time to teach students what they really will need in this world."

In addition to their negative effects on instruction, test comparisons do not provide a valid basis for an accountability system. The RAND study I referred to earlier put it this way:

"Comparisons of simple, unadjusted test scores from one year to the next or across different schools or districts do not provide a valid indicator of the performance of the teachers, schools, or school districts unless the differences in scores are very large compared to what might be accounted for by changing demographic or family characteristics. This is rarely the case; so, any use of unadjusted test scores to judge or reward teachers or schools will inevitably misjudge which teachers and schools are performing better."

The critique of current standardized tests has led to a fourth myth:

4. The problems in current standardized testing programs can be solved simply by developing new and "improved" tests.

It is argued that new, innovative tests--the terms used are performance tests, portfolio assessments, and essay examinations--will take care of any flaws in current testing programs. Little attention, however, is paid to how long such tests would take to

develop, how much they would cost, whether they could be administered on a large scale, how much they would substitute for instruction in schools, and whether valid comparisons could be made.

Research and practical experience in a diverse set of states show that the new tests are unlikely to be appropriate for large-scale use to compare school districts, schools, or students because they are not designed psychometrically for making such comparisons.

The scoring is highly unreliable: In some cases, the problem is rater reliability; in others, task reliability (low correlations between performance on different subtasks). Further, measures of test validity (for example, the extent to which the tests predict students' future academic performance) are lacking--not surprising, given the low reliability of the test scores. The point is that the new tests don't increase the validity of test comparisons--they further decrease them. And they clearly don't address the basic problem: Test-score differences from year to year or from school to school tell us little about the quality of the educational program.

Some state testing programs have tried to use complex statistical formulae to control for student background variables that might affect gain scores. This worthy attempt to make the comparisons "fair" has not worked--the measures simply cannot capture key variables --but it also has resulted in test-score measures that are incomprehensible even to the educators working within the system. One superintendent put it this way: "If you ask me do I understand the formula, no. No I don't . . . for me to sit here and you ask me to explain it to you, I'd be in deep trouble."

It is reasonable to ask, though, whether the tests' technical problems may be outweighed by positive instructional effects. Teachers do report that the new tests have served to draw their attention *toward* writing and problem solving skills--and *away* from rote learning (which, ironically, is a legacy of the traditional, multiple-choice, standardized tests). But these benefits could be obtained by using the tests for *instructional* purposes within schools--without attempting to make comparisons that provide misleading information.

Clearly, any instructional benefits that might arise from the new tests are not automatic. One researcher puts it this way: "It is one thing to communicate to mathematics teachers, for example, that they should put more emphasis on problem solving; it is quite another to communicate what that means, what types of problems should be used to embody those skills, and how that increased emphasis ought to be implemented in the curriculum."

In addition, the testing programs, which are extremely costly and time-consuming, use resources that might more productively be applied to other instructional purposes. In testimony before the U.S. House of Representatives Committee on Education and Labor, several researchers estimated the cost of administering tests nationally in five subject matters in only three grades at more than $3 billion per year. By comparison, the entire

Title I program, the largest federal program for elementary and secondary education, spends about $7 billion a year.

In Vermont's portfolio assessment program, teachers reported spending an average of 30 hours per month (excluding training) working on mathematics portfolios. In Kentucky, fourth-grade teachers were "overwhelmed" by the administration and grading of the writing and mathematics portfolios. In Montgomery County, Maryland, reading specialists spend an average of one month per year on language arts tests--time taken from instruction children otherwise would receive.

It is clear that testing programs contribute to bureaucracy, paperwork and costs. They raise a serious question, therefore, about the allocation of scarce resources. Yet, the use of test scores for accountability purposes often is proposed as a way to *reduce* regulatory burden--and its associated costs--by replacing traditional regulations (for example, on class size or teacher credentials) with student test scores.

California recently eliminated a performance evaluation system that was critiqued for being too costly, giving too *much* attention to *non-standard assessment*, and giving too *little* attention to *multiple-choice tests*. In Kentucky, a number of state legislators have advocated major changes in the performance testing program, including a return to standardized tests.

Perhaps the best example of what happens to testing programs comes from England. In 1988, Parliament mandated national curricula and assessments. In the first year of assessing 7-year-olds, the assessments took two to four weeks out of the school year. For the 1993 assessment of 14-year-olds, the marking and reporting form for mathematics was 112 pages long. As a result, the teachers, with strong parental support, boycotted against administering the tests and reporting test scores. They cited a range of concerns similar to those emerging from testing programs in the United States--overwork, bureaucracy, disruption of regular schooling, flawed tests, use of scores to compare schools and opposition to national curriculum and testing. This version of the British testing program has been abandoned.

A student in one of the U.S. state testing programs summed up the problem this way:

"I feel that the portfolio problems were really dome. They didn't make any sense and they didn't accomplish any thing. If they were a little harder I mean alot harder, mayby it would of help me but it didn't. There was probobly help some people. They also wasted time in class. I know why you are doing this because you want to be better than the Japanies."

I will conclude with one more myth, perhaps the most counterproductive of all:

5. We can compensate for the inadequate resources spent on poor children by increasing testing requirements.

Or, put another way, money doesn't matter.

However, research shows that per-pupil expenditure, teacher expertise, and class size do make a difference in student achievement. Increasing testing requirements will not buy better teachers or the individual attention children can receive in small schools or small classes. They will not provide low-income inner-city or rural students with science laboratories, computers, or decent facilities--amenities that affluent students take for granted. Tests will not fix the broken windows or the crumbling buildings.

Nor will tests reduce the school finance inequities, which mean that children from families with the lowest incomes attend the most poorly funded schools. You are all familiar with the large disparities among states, among districts within a state and, in some cases, even among schools within a district. For example, the 100 poorest districts in Texas spend an average of just under $3,000 per student. The 100 wealthiest districts, however, spend about $7,200 per student. In Illinois, school districts spend between roughly $2,400 and $8,300 per student. If money doesn't matter, rich districts haven't heard the message.

An educator summed up the results of his state's testing program this way:

"Basically, when you look at about three factors, and that is the number of students on free lunch, the economic base of the district and then the amount spent per child for instruction, it correlates almost 100 percent with the districts who fall at level 1, 2, 3, 4 and 5. The ones that are level 4 and 5 [the highest scoring districts] have very few students on free lunch."

The point is, we cannot improve our schools by giving more tests. A preoccupation with test score comparisons, I am afraid, encourages public officials to look for "quick fixes" and deflects attention from real problems: the large proportion of children who live in poverty and the vast differences in educational resources between rich and poor schools. My greatest concern is that a focus on test scores takes attention away from the problems in our most troubled schools, the real work that needs to be done to address them, and the resources needed to do it.

But let me conclude on a lighter note with a few additional quotes from student responses to exam questions:

A circle is a line which meets its other end without ending.

The earth makes one resolution every 24 hours.

If conditions are not favorable, bacteria go into a period of adolescence.

Dew is formed on leaves when the sun shines down on them and makes them perspire.

Blood flows down one leg and up the other.

The theory of evolution was greatly objected to because it made man think.