MYTHS IN INTERNATIONAL TEST SCORE COMPARISONS

Greater Edmonton Teachers' Convention
February 25, 1994

It is a pleasure to be here this morning and to discuss with you a set of public policy issues that are receiving a great deal of attention both in the United States and Canada. I would like to add, as well, that a visit to Canada is always a homecoming for me. My mother was Canadian--from Montreal. My early impressions of Canadian education were based on my experience in Quebec. If I make some assumptions about education in Alberta that don't sound quite right, I hope you will attribute my errors to that early imprinting--in a different province and, of course, in a different generation.

Coming to Alberta, reading the newspapers over the past two days, talking to Kathy and some of the rest of you, I am reminded again of the similarity in issues and problems between the United States and Canada. The discussion here--as in the States--is about budget cuts, unequal funding, international comparisons, and increased student testing.

In my remarks today, I will discuss major shortcomings of the international test score comparisons and draw implications for the recent comparison of Alberta with Beijing, Sendai, and Taipei. I will conclude with comments about the typical response--in Canada, in the United States, in England--to scares about international comparisons: increased requirements for standardized testing in schools.

Since the 1960s, there have been a number of international test score comparisons of science and mathematics achievement, sponsored by different organizations and involving different countries.

I have two main concerns about the international comparisons:

o First, international test score comparisons are flawed methodologically and do not reflect the quality of education in any country.

o Second, a reliance on a narrow criterion--answers on multiple choice tests--ignores far more important measures of our strengths and shortcomings in education and leads us to recommend solutions that are irrelevant at best and often are counterproductive to resolving or even addressing our most important problems.

The basic problem in any comparisons of this type is that the more students who take the test, the lower will be the average score. That score has little to do with the quality of education. 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.

o For example, in the 1960s, high school attendance rates in the United States were substantially higher than those in most other countries. United States: close to 80 percent; European average: 20 percent. While the European attendance rates have gone up substantially, they remain lower than those in the United States and Canada.

o In recent studies, major reversals of rankings between higher and lower grades: e.g., Hungary and England/Wales, Japan and Hong Kong, the former Soviet Union, Slovenia, and the United States. The reversal in rankings simply reflected the larger or smaller number of students taking the test.

o Analogous to SAT scores: ranking of states; letter from resident of affluent (and therefore high SAT) district that merged with low-income, low SAT district.

Again, the important point is that the more students who take the test, the lower will be the average score. That finding holds whether the comparisons are among nations, among provinces, among states, among school districts, or among schools within a district.

The second major factor contributing to the international test scores is 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 the numbers of low-income students have 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 more commonplace problems like TV-watching or wearing earphones while studying.

Third, differences in curriculum emphases among nations also affect the international rankings--e.g., differences in timing of calculus courses.

Finally, there are differences in how the material is presented. Bette Bao Lord (the author of Spring Moon) puts it this way:

"As a fifth grader in Brooklyn's P.S. 8 . . . even before I had mastered fifty words of Brooklynese my teacher . . . began asking me for my opinion on every matter . . . I was flabbergasted by the fact that an adult--and not just any adult; on the contrary, my most honorable teacher--would solicit the opinion of a child--not just any child; on the contrary, an eight-year-old immigrant just off the boat. . . . And before long I came to realize that the merits of one's opinions were not the crucial point of the exercise. The crucial point was to air whatever opinions one had, and today I value this aspect of what we Americans delight in praising as our way of life perhaps more than any other."

The emphasis on classroom discussion that Bette Bao Lord describes, while highly desirable, is not necessarily reflected in higher scores on multiple choice tests of isolated pieces of information.

I am often asked whether with all our expertise in statistics and sampling design we can't simply improve the validity of the international comparisons. I don't believe we can. The problems are endemic to all of the studies since they began in the 1960s. The problems are not a matter of statistical expertise, but of the societal and educational diversity among countries. This diversity cannot be controlled for by any statistical design.

There are large differences among countries in <u>which</u> students take the test. For example:

o Exclusion of 20 percent of the classes.

o Enrollment in apprenticeship programs.

o Tracking by age 11.

o Exclusion of regions, language groups; in most recent ETS study, for example, Italy included only one province (Emilia-Romagna), the former Soviet Union (when it still was the Soviet Union) included only Russian-speaking schools, Israel included only Hebrew-speaking schools.

o Highly specialized curriculum in some countries. Even Princess Diana (not then a princess, of course) did not continue in an academic program past the age of 16.

o Problems magnified by inclusion of broad range of developing countries, with highly elitist school systems (because of scarce resources) and substantial proportions of children who are out of school and therefore do not take the tests. The international comparisons are no more useful to a developing country trying to make difficult choices about how to use scarce resources than they are to the United States or to Canada.

I would like to turn now to Harold Stevenson's comparisons of Alberta's eleventh grade test scores with scores in other countries. Each of the points I have made apply as well to these comparisons. Indeed, unlike the larger studies I described, Stevenson's study did not even attempt to draw representative samples.

Holding aside Minneapolis and Fairfax County, where the differences were small, consider the other test score comparisons.

Beijing students not only scored higher than students in Alberta, but they also scored higher than students in Taipei and Sendai. Why? Because of scarce resources, a relatively small proportion of Chinese youth are still in school in the eleventh grade. According to the World Bank's <u>World Development Report</u>, only 48 percent of

Chinese children (and 41 percent of girls) attend secondary school, defined as seventh to twelfth grade. Because of scarce resources, a much smaller percentage, of course, attend upper secondary school (the eleventh grade in Stevenson's study)--certainly fewer than one-third of the age group. (Only two percent of Chinese young people attend college.) When these attendance rates are combined with the fact that the Chinese education system is elitist, with a highly selected group of students attending "key schools," we can understand why the Chinese students scored not only higher than Canadian and American students but also higher than students in Japan who are noted for their accomplishments in mathematics.

While I do not have separate statistics for Taiwan (Taiwan is considered a province of China and not a separate country for purposes of World Bank membership and World Bank statistical reports), informal evidence suggests that Taiwan too has a highly elitist school system, with many students--although perhaps not as many as China--leaving school before the eleventh grade.

Where does Japan fit into all of this? Japan, like the United States and Canada, has a high proportion of its students completing high school. As you know, it stresses mathematics in its school curriculum, and many Japanese students also are in cram courses--in addition to their regular schooling--for as much as 20 hours a week.

In short, Stevenson's study, like the larger studies I described earlier, reflects societal differences in school attendance rates, in the elitism of the schools represented in the study, and in the choices each country makes with respect to what students should learn, how much time they should spend learning it--both in and outside school, and when they should learn it (e.g., calculus). We should be careful not to hold schools accountable for results that are caused by quite extraneous factors.

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, Stevenson notes that Asian mathematics teachers spend much less time in the classroom than do teacher in Canada or 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.

The challenge in international studies is to identify those practices that can realistically be transferred from one nation to another.

However, in most cases, it would involve a basic restructuring of a nation's social, cultural, and political norms, for example, changes in the respective roles of national and local governments in education, the role of the teacher in society, teachers'

salaries, comprehensive high schools, competitive sports in schools, summer vacations, the elitism of the school system, our value system with respect to pluralism, open access to higher education across socioeconomic groups, the role of industry in vocational education and apprenticeship programs, and similar issues that each country looks at differently. These issues are tough ones. And they are likely to generate tremendous conflict. Instead, we "dump" on our schools, as if they were a magic factor which, alone, could be changed without addressing the far more systemic and underlying social and political systems on which an education system is based.

Even when there is a public discussion within a country about making basic changes in education, the nation's social and cultural norms make it very difficult to accomplish. For example, in the United States, the current debate is about giving our elementary and high school students a more demanding curriculum, and then administering national tests; Japan would like its students to express their own views more readily; Taiwan would like its students to play more! A matter of culture, not the education system. In the area of international competitiveness, there are similar problems in trying to adapt industrial policies from Japan or Germany to the United States because government/industry links differ so fundamentally between the countries.

I would like to turn now to the issue of whether it matters if we focus on test score differences, or if we exaggerate the problems in the United States and Canada, when we all agree that our education systems can be better than they are.

I believe it does matter.

First, the rhetoric is not supported by the facts. We incorrectly assume that adverse test score differences mean that our schools, or our parents, or our students, or our scientists, or our research institutions have failed. The fact is that test scores are highly misleading indicators of the merits of a nation's education system, the expertise of its students, or the quality of its scientific research.

Second, the rhetoric has detracted, at least in the United States, from real problems--the large proportion of our children who live in poverty, the vast differences in educational resources between rich and poor schools, and the combination of rising costs of higher education, reductions in the real value of student financial aid, and decreasing state expenditures for higher education--and what these trends do to student motivation. My concern is that a focus on test scores deflects attention from what we can do to solve our real problems.

Consider school finance inequities in the United States, 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. There also are wide gaps in per-pupil expenditures among states and among schools within districts. These are real problems.

Third, the focus on international test score comparisons inevitably leads to more testing requirements in our schools. We assume that our children will learn more if we give them more standardized tests, or that we can test our way to school improvement. The assumption is sort of like a "Field of Dreams" argument: Build a test and they will learn. Evidence from research and practical experience suggests quite the opposite for several reasons.

First, 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.

Second, test-score differences from year to year, or from school to school, tell little about the quality of the educational program. The quality of an education system, or of an individual school, cannot be measured simply by comparing test score fluctuations from one year to another, or by comparing schools or classrooms on test scores. The reason is that the results do not control for changes in student population, incentives for encouraging certain students to take (or not to take) the test, or the consistency (or lack of it) between the test and the instructional program. We all know that it is not difficult to raise test scores if we spend a lot of time teaching to the test, or if we exclude more students from taking the test. We all know as well that the higher scores under those circumstances do not reflect improved education.

It is sometimes argued that testing can be improved by developing innovative new tests, which would include performance assessments, essay exams, and portfolio assessments. Little attention 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, and how much they would interfere with instruction in schools. These tests might be useful for research or diagnostic purposes in individual schools, but in my view are unlikely to be appropriate for large-scale use to compare countries, or provinces, or even schools within a province.

The point is that increases in testing requirements often contribute more to bureaucracy, paperwork, and costs than they do to the quality of education.

Perhaps the best example of what happens when standardized testing is carried to an extreme 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 math was 112 pages long. I would

like to quote from a description in the press of what happened in the Summer of 1993:

"Citing a range of concerns such as overwork, bureaucracy, disruption to regular schooling, flawed tests, use of scores to compare schools, and opposition to national curriculum and testing, all but one of Britain's teacher unions joined in a boycott against administering and reporting tests for 14-year-olds and reporting test scores for 7-year-olds.

. . . "Schools made substantial efforts to inform parents of the reasons for the boycott. The government responded by publishing the tests to persuade parents that the boycott was not worth the trouble. However, independent polls and a government report all indicated strong parental support for the teachers.

"The boycott was initiated in April by the National Association of Teachers of English. They viewed the reading and writing tests for 14-year-olds as particularly narrow and flawed. Other unions quickly joined.

"As opposition to the test for 14-year-olds grew, the teachers also decided to boycott the test for 7-year-olds. Since most of that assessment had already been administered, a decision was made to refuse to report the scores to the government. . . .

"The 1993 assessments of 7-year-olds were to have been the first reported nationally and were to include comparisons among schools in a region. The boycott . . . eliminated that possibility. The government reportedly spent 35 million pounds (about $55 million) to conduct the now-useless 1993 exams."

The result is that at least for the present, the British testing program has been abandoned.

This is an example of what happens when government policy-makers decide that "silver bullet" solutions--whether they are standardized tests, or vouchers, or school "restructuring"--can solve the basic problems of underfunding of schools, school finance inequalities, or the educational problems associated with poverty. Those are real problems and deserve a lot more attention than international--or domestic--test score comparisons and rankings.

Thank you.