For more information please contact:

Canadian Teachers’ Federation
110 Argyle Avenue
Ottawa, ON K2P 1B4
Tel: (613) 232-1505
Fax: (613) 232-1886
E-mail: info@ctf-fce.ca
Internet address: http://www.ctf-fce.ca

The Canadian Teachers’ Federation was created in 1920 to ensure that teachers’ views were taken into account whenever any authority outside the provincial jurisdiction was considering action which would affect teachers, or impinge on their work with students. Active at the interprovincial, national and international levels, CTF is the national voice of teachers in promoting quality education, the status of teachers and equality of opportunity through education. CTF also coordinates and facilitates the sharing of ideas, knowledge and skills among its provincial/territorial Member organizations which, collectively, represent over 240,000 teachers.

ISBN: 0-88989-326-8

© Canadian Teachers’ Federation, 1999

The production of this publication including translation, layout and printing was made possible through the collaborative efforts of CTF staff.

Cover Design by Prosar Canada Inc.
# Table of Contents

Preface ............................................................................................................................... i

Introduction ....................................................................................................................... 1

Problems with Standardized Testing – An Overview ....................................................... 3

  Inability to assess student learning and development .................................................. 5

  Other problems ............................................................................................................... 12

Testing and Inequity in Education .................................................................................... 13

  Inappropriate use of tests ............................................................................................. 13

  High-stakes, measurement-driven accountability .......................................................... 18

  Bias in testing .................................................................................................................. 23

A Few Words About Authentic Assessment ................................................................. 31

Politics of Testing ............................................................................................................. 35

  International comparisons .............................................................................................. 35

  Testing, schools and market competition ..................................................................... 38

  ‘Pan-Canadianization’ of curriculum and assessment ................................................... 40

  The bigger picture of educational reform ...................................................................... 45

Some General Observations .............................................................................................. 49

Appendix A – Summary of Problems with Standardized Testing .................................. 53

Appendix B – Principles for Fair Student Assessment Practices for Education in Canada (Joint Advisory Committee) ................................................................. 55

Appendix C – CTF Principles on Curriculum, Assessment and Evaluation ................. 57

References ......................................................................................................................... 61
Standardized Testing: 
Undermining 
Equity in Education

Report prepared for 
the National Issues in Education Initiative

by Bernie Froese-Germain 
Canadian Teachers’ Federation

November 1999
Canadian schools are awash in student testing. Changes in assessment policy have increased standardized testing at provincial, national and international levels, introduced testing at more (and earlier) grade levels, increased the reporting of test results and attached more significance to those results. Advocates claim that more testing will result in greater school accountability – testing seems to have become synonymous with accountability.

In this context of an increased focus on testing as an instrument of educational reform, CTF’s Standardized Testing: Undermining Equity in Education argues that there is little evidence to support standardized testing from a pedagogical perspective and looks at the effects of standardized testing on educational equity. Problems with standardized testing are well-documented. They stem from the nature of the tests and the way in which tests and their results are used. This report considers some of these problems, with particular emphasis on the impact of testing bias and the misuse of test results on equity in education. It also discusses performance-based assessment, examines the politics of testing, offers some general observations on the assessment of student learning, and provides principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

Intended to increase awareness of critical issues related to the impact of standardized testing, this book will be a useful resource for teachers at all levels as well as administrators, researchers, policy-makers and anyone with an interest in student assessment and evaluation.

Bernie Froese-Germain is a researcher with the Canadian Teachers’ Federation in Ottawa.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>i</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Problems with Standardized Testing – An Overview</td>
<td>3</td>
</tr>
<tr>
<td>Inability to assess student learning and development</td>
<td>5</td>
</tr>
<tr>
<td>Other problems</td>
<td>12</td>
</tr>
<tr>
<td>Testing and Inequity in Education</td>
<td>13</td>
</tr>
<tr>
<td>Inappropriate use of tests</td>
<td>13</td>
</tr>
<tr>
<td>High-stakes, measurement-driven accountability</td>
<td>18</td>
</tr>
<tr>
<td>Bias in testing</td>
<td>23</td>
</tr>
<tr>
<td>A Few Words About Authentic Assessment</td>
<td>31</td>
</tr>
<tr>
<td>Politics of Testing</td>
<td>35</td>
</tr>
<tr>
<td>International comparisons</td>
<td>35</td>
</tr>
<tr>
<td>Testing, schools and market competition</td>
<td>38</td>
</tr>
<tr>
<td>‘Pan-Canadianization’ of curriculum and assessment</td>
<td>40</td>
</tr>
<tr>
<td>The bigger picture of educational reform</td>
<td>45</td>
</tr>
<tr>
<td>Some General Observations</td>
<td>49</td>
</tr>
<tr>
<td>Appendix A –</td>
<td></td>
</tr>
<tr>
<td>Summary of Problems with Standardized Testing</td>
<td>53</td>
</tr>
<tr>
<td>Appendix B –</td>
<td></td>
</tr>
<tr>
<td>Principles for Fair Student Assessment Practices for Education in Canada (Joint Advisory Committee)</td>
<td>55</td>
</tr>
<tr>
<td>Appendix C –</td>
<td></td>
</tr>
<tr>
<td>CTF Principles on Curriculum, Assessment and Evaluation</td>
<td>57</td>
</tr>
<tr>
<td>Principles on Curriculum</td>
<td>57</td>
</tr>
<tr>
<td>Principles on Assessment and Evaluation</td>
<td>58</td>
</tr>
<tr>
<td>References</td>
<td>61</td>
</tr>
</tbody>
</table>
Introduction

Addressing a Canadian Teachers’ Federation workshop on curriculum and evaluation in 1998, Benjamin Levin, Dean of Continuing Education at the University of Manitoba, provided an overview of the changes occurring across the country with respect to educational policy. On the curriculum side, he noted that policy changes include an increased focus on the ‘basics’, fewer options and more restrictive time allocations, the development of regional and Pan-Canadian curricula, and shorter implementation times for new curricula. Changes in assessment policy include more provincial testing, more testing at the national and international levels, increased public reporting of test results despite the generally poor quality of this reporting, and more importance given to the results. Canada is not alone in placing an increased focus on testing. According to Levin (1998),

an emphasis on standards, accountability and testing has been a feature of reforms in many countries. Almost everywhere we find more large-scale testing of students and more reporting of the results of these tests than was the case a few years ago. Increasing national assessment is complemented by more and more international assessment and in both cases the results are used more overtly for public comparisons. (p. 133)

---

1 The National Issues in Education Workshop on “Curriculum and Evaluation: Teacher Organizations and Pan-Canadian Developments” was convened by the Canadian Teachers’ Federation in January 1998 to bring the leaders of teacher organizations together to discuss recent developments in curriculum and evaluation, particularly their increasingly “Pan-Canadian” nature. Workshop participants developed statements of principle on curriculum, assessment and evaluation which are contained in Appendix C at the end of this report.
While advocates claim that standardized testing will increase the accountability of teachers and schools, there is little evidence to support such claims. There is, however, a good deal of evidence documenting numerous problems associated with standardized testing (Meaghan & Casas, 1995d, p. 46). Problems stem not just from the nature (form and content) of the tests but perhaps more importantly, from the way in which tests and their results are used. This report considers some of these problems, particularly the impact of testing bias and the misuse of test results on educational equity. It will also briefly discuss authentic or performance assessment, address standardized testing from a political perspective, and offer some general observations on the assessment of student learning and development. Appendices are included which contain principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

The numerous references in the report to testing in the U.S. reflect the extensive experience Americans have had with this practice, experience which holds valuable lessons for Canada. It is estimated that American public schools administered more than 200 million standardized tests over the 1986-87 school year (Medina & Neill, 1990). According to Barlow and Robertson (1994), “if testing created excellence, American students, the ‘most-tested’ students in the world, should be winning the education sweepstakes” (p. 117); instead, the U.S. consistently scores low on large-scale international assessments (Earl, 1995, p. 5). Research conducted over the past two decades confirms that standardized tests have had “many negative consequences for the quality of American schooling and for the equitable allocation of school opportunities.” (Darling-Hammond, 1991, p. 220)

2 While the terms assessment and evaluation are often used interchangeably (as is the case in this report), there is a subtle yet important distinction between them. Assessment is defined as the gathering of reliable information pertaining to students’ knowledge and understanding including social, economic and educational factors and resources that influence student and system performance. Evaluation is the process of making judgements, based in part on assessment data.


Problems with Standardized Testing – An Overview

Developed during the First World War to identify army recruits who could be trained as military officers, standardized tests were a means of spreading the candidates out in order to spot those best suited for leadership roles (Casas & Meaghan, 1995, p. 18). Traub (1994) defines standardized achievement tests in the following way:

Generically, an achievement test is designed to assess the knowledge and understanding a student has acquired of a school subject. A standardized achievement test is further defined by its being given and scored in the same way, whenever and wherever it is used. Standardization means that the scores of all students tested can be fairly compared, one against the other ... the essential requirements are that the conditions of administration and scoring be the same for all the students who are tested so that their scores can be compared. (p. 5)

A summary of some of the major problems with standardized testing is contained in Appendix A.

The U.S. Scholastic Aptitude Test (SAT), developed in the 1920s, was based on the Army Alpha test, a test used to select potential officers from among army recruits (Lemann, 1995). For a brief history of testing in general (including the introduction in the 18th century of ranking and the invention of the quantified mark), see Madaus & Kellaghan (1993).
They are often referred to as norm-referenced tests because an individual’s scores are compared to the scores of a larger representative group (the “norm”). The test format is frequently multiple choice, making it easier to mark with the assistance of computers. Aside from advances in electronic scoring, analysis and reporting of results, many standardized achievement tests have changed very little over the past fifty years despite claims that they have become more “scientific” (Medina & Neill, 1990, p. 13) and despite advances in our knowledge about what makes for good teaching and effective learning.

A substantial body of research has documented numerous flaws in the construction, validation, administration and use of standardized tests. According to Medina and Neill of FairTest (the National Center for Fair & Open Testing in Massachusetts),

[s]tandardized tests are constructed in ways that often guarantee biased results against minorities, females and low-income students. Test results are evaluated and scored in ways that are often at odds with modern theories of intelligence and child development. The test validation process is often inadequate and far from objective. Many tests are administered in an environment that undermines any claims they may have to being “standardized”. Even those that adhere to “standard” administration practices may be biasing the results against minorities, low-income students and females by using examiners who are unfamiliar to the test-takers and by using tightly timed tests. These flaws undermine testmakers’ claims of objectivity and produce results that are inaccurate, unreliable or biased. Ultimately, many tests fail to effectively measure test-takers’ achievement, abilities or skills. (p. 8)

References


Inability to assess student learning and development

Although standardized tests may be useful for sorting and ranking students, they are inadequate in assessing student learning and development. The relevant literature consistently identifies the following reasons:

Many types of student ability are not captured by a standardized test.

Qualities such as a student’s sense of citizenship, ethics, confidence/self-esteem, aesthetic appreciation, respect for others, self-discipline, social competence, and desire to learn, although not easily quantifiable, contribute enormously to the multiple goals of schooling. By forcing schools and teachers to concentrate on basic, quantifiable skills, standardized testing poses a serious threat to this diversity of goals and content. Medina and Neill tell us that “[t]he ability of standardized tests to accurately report students’ knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score.” (p. 12)

In his seminal work The Mismeasure of Man, Gould debunked two major myths about human intelligence (as cited in Darling-Hammond, 1994, pp. 10-11): that it is an “innate, unitary, measurable commodity” and, flowing from the notion that intelligence could be quantified in this way, that individuals can be compared and ranked on this basis. Different perspectives on the concept of intelligence are emerging. Gardner (1999) notes that areas of debate include “what intelligence is, how (and whether) it should be measured, and which values should be invoked in considerations of the human intellect.” (p. 70) He proposes that we all “possess at least eight intelligences: linguistic and logical-mathematical (the two most prized in school and the ones central to success on standard intelligence tests), musical, spatial, bodily-kinesthetic, naturalist, interpersonal, and intrapersonal” and that “no two of us exhibit the
same intelligences in precisely the same proportions.” (p. 71) This of course has important implications for educational assessment and evaluation practices.

**Tests may be standardized, but students are not.**

A student’s capacity to learn is affected by a host of factors including: the impact of poverty; parents’ educational level; mental, physical and emotional health; the effects of racial and other forms of discrimination; and language of origin. For this reason, the absence of information on the characteristics of students in the schools taking part in testing programs leads to misleading and inappropriate use of the data derived (Canadian Teachers’ Federation, 1995b). In a 1994 survey of province-wide assessment programs carried out by the Canadian Teachers’ Federation, a few jurisdictions reported collecting some background data on students in conjunction with the tests (principally demographic in nature). One province collected information on such factors as parental involvement in education and amount of homework. However, no information on the socio-economic origin of students was collected in any jurisdiction (Gilliss, 1994, p. 5).

Gilliss (1993) states that “[a]nyone who expects students who are disadvantaged, impoverished, undernourished, neglected or severely disabled to score at the same level as students who are affluent, healthy, well cared for or able [bodied] is doomed to disappointment.” She also observes that average test scores for schools provide more information about the pooled characteristics of individual students than about the schools themselves. Rather than incur the expense of developing, administering, grading, and ranking test scores to determine which schools score the highest, a socio-economic analysis of census data of the communities where the schools are located would provide much the same information. McLaughlin (1991) echoes this point. Responding to the call for standardized testing in “America 2000” (the national education strategy for the U.S. announced under former President George Bush), she notes that testing
Curriculum is ongoing and dynamic.

Curriculum development, implementation and evaluation are multi-phased processes that require extensive time and resources.

Teachers must have the primary role in, and control of, curriculum development.

Professional dialogue and collaborative activity are essential for professional growth and development.

Teachers require time and resources to engage in ongoing professional development, and in order to effectively implement curriculum change.

Teachers must have the autonomy to choose the methods, timing and curriculum adjustments that are necessary to adapt the curriculum in order to meet students’ needs.

The goals society sets for students and schools must be challenging but attainable, and progress towards these goals must be measured fairly and in context.

The conditions under which teaching and learning take place have a direct effect on what teachers and students can achieve together.

The process for curriculum development must include opportunity for input from the broadest possible community representation.

Principles on Assessment and Evaluation

Assessment is the gathering of reliable information pertaining to students’ knowledge and understanding including social, economic and educational factors and resources that influence student and system performance.

Standardized tests designed for large numbers of students are of necessity very general in nature.

This leads to a mismatch between what is taught (given curriculum diversity) and what is tested (Casas & Meaghan, 1995, pp. 18-19). In his analysis of the Grade 8 results from the Third International Mathematics and Science Study (TIMSS) released in November 1996, Ireland (1997b) points out the substantial differences in test-curriculum match which exist among the provinces. For example, the TIMSS report noted that while 98% of the test items matched the British Columbia science curriculum, the corresponding figures for Ontario, Alberta, and Newfoundland were 53%, 54%, and 54% respectively. There was however little explanatory information on test-curriculum match beyond the tables and their respective figures. Ireland suggests that the TIMSS study could have been improved by addressing the following questions (excerpted from Ireland, 1997b, p. 14): How well did students do on questions derived from the curriculum they should have been taught? How representative of a jurisdiction’s curriculum was the test and how significant is the part of the curriculum not tested? What outcomes were on the test which are not in a jurisdiction’s curriculum, and how significant are those outcomes?

Standardized tests typically measure lower-order recall of facts and skills, and penalize higher-order thinking.

Meaghan and Casas (1995d) note that standardized tests “measure a student’s ability to recall facts, define words, perform routine operations, rather than higher learning skills such as analyzing, synthesizing, forming hypotheses and exploring alternative ways of
solving problems”. In addition, they “exclude the process of feedback and debate which is at the core of learning .... [and] they rule out independent creativity and criticism by eliminating the need to construct answers” (p. 47). The following analysis by McMurtry (1992) of a question from an Alberta Ministry of Education achievement test for Grade 3 social studies underscores this problem:

Question 7, one of 9 Model Questions, asks “Which picture BEST shows people keeping a custom?” Underneath are four block pictures – a pick-up truck, three Native people in traditional costume, kids in uniform playing hockey, and a group of teenagers dancing. The last three of these pictures all qualify as representing “people keeping a custom.” Yet only one picture qualifies as “correct.” This means that a grade 3 child could be marked wrong for choosing correctly in two of the three cases, that any reason or argument over what is in fact true is ruled out even if it is fully justified, that a genuinely critical intelligence will be thrown off by a test that is demonstrably arbitrary in its selection of what is right and wrong (p. 95).

In U.S. schools, the increasing emphasis on the use of “test-oriented accountability measures” parallels a decline in teaching methods such as essay writing, research projects, lab work, and student-centred discussions designed to foster higher-order learning skills (Darling-Hammond, 1991, p. 222).

Because standardized tests are designed to sort individuals into groups, test questions are chosen on the basis of how well they contribute to spreading out the scores, not on their centrality to the curriculum or their predictive validity.

In norm-referenced tests, score dispersion is required as a basis for making comparisons. Questions that don’t contribute to spreading out the scores are therefore avoided. Robertson (1998b) explains this rather technical concept in terms understandable to the lay person:

### Appendix C — CTF Principles on Curriculum, Assessment and Evaluation

The Canadian Teachers’ Federation has developed its own set of principles on assessment and evaluation. In early 1998 the CTF held a workshop on curriculum and evaluation (referred to in the introduction) for the leadership of teacher organizations across the country. An important outcome of the workshop was the creation of draft statements of principle on assessment and evaluation as well as on curriculum. These statements of principle are intended to guide the work of provincial and territorial teacher organizations in a number of areas including the development, implementation and scope of curriculum, and high-stakes testing and its use. They have since been adopted as CTF policy and are reproduced here:

### Principles on Curriculum

- Curriculum must be broad in scope and holistic in nature in order to meet the needs of children, community and the larger society.
- Curriculum must reflect realistic expectations based on knowledge of child development.
- Curriculum must promote, support, enhance and model equity and respect for diversity.
- Curriculum must be designed to prepare students to become caring, responsible and participatory members of society.
Standardized tests actually squeeze out what is most commonly taught and learned, not because of any nefarious plot, but because of the statistical requirements of data analysis. The type of question that best discriminates among learners is the one that is answered correctly only 50 per cent of the time. Questions that are consistently answered correctly by more than 60 per cent of students are routinely dropped. While these questions may be too easy, they may also represent a core area of content that has been carefully taught to almost every child. To function as they are intended, standardized tests require the evaluation of peripheral concepts, not the common core. Forced to teach to the test, teachers must focus on less important content – the kind of factoids that suit multiple-guess exams – rather than on core ideas or complex understandings. (p. 71)

Test performance is shaped by individual characteristics not related to content knowledge.

Pressure to increase scores on standardized tests and “a lack of understanding of the complexities of achievement and its causes” have resulted in what is known as test score pollution, described as “factors affecting the truthfulness of a test score interpretation.” (Haladyna, Nolen & Haas, 1991, p. 4) Test score pollution encompasses a number of test preparation and administration practices designed to increase test scores but which are unrelated to the material covered. Haladyna, Nolen and Haas (1991) describe some of these practices and whether they are ethical. They include:

- teaching test-taking skills;
- developing a curriculum and teaching objectives to match the test;
- presenting items similar to those appearing on the test;
- presenting the actual test items before the test;
√ using commercial materials specifically designed to improve test performance;
√ dismissing low-achieving students on test days; and
√ interfering with responses by, for example, giving hints or answers to students or altering response sheets.

Other factors that can influence test scores include students’ anxiety and motivation levels (Haladyna, Nolen & Haas, 1991). It has been demonstrated that anxious test-takers don’t perform as well as test-takers who are more relaxed, and they do particularly poorly on more difficult test items. A meta-analysis of over 500 studies by Hembree (as cited in FairTest, 1995) reveals “that students with high test anxiety hold themselves in lower esteem than do those who are less test anxious .... [f]emales have higher test anxiety than males, Blacks in elementary school have higher test anxiety than Whites, and Hispanics have higher test anxiety than Whites at all ages.” (p. 9)

Meaghan and Casas (1995d) note that because standardized test scores “are very sensitive to motivational factors ... [they] are not good measures of knowledge” (p. 47). Motivational factors were the focus of a study of student performance on the mathematics section of the Iowa Test of Basic Skills. In this study over 400 elementary students were randomly placed in two groups – those in the experimental group, which received special instructions from their teachers on the importance of doing their best on the test, had significantly higher scores than a control group which did not receive the instructions (Meaghan & Casas, 1995c, pp. 42-43).

_Test preparation and administration take up valuable classroom time that could be used for teaching._

In some instances, curriculum is discarded to make way for test preparations that sometimes begin months before the actual test (Medina & Neill, p. 26). A number of Ontario teachers preparing students to take the Ontario Grade 9 Reading and Writing Test noted privately that much of the latter part of the 1993/94 school year was

---

**Appendix B — Principles for Fair Student Assessment Practices for Education in Canada**

(Joint Advisory Committee)

In 1993 a Canadian working group convened by the Joint Advisory Committee of the University of Alberta’s Centre for Research in Applied Measurement and Evaluation produced a document called _Principles for Fair Student Assessment Practices for Education in Canada_. The principles and their related guidelines address both classroom assessments and large-scale standardized assessments developed outside the classroom (by commercial test publishers, ministries of education and school boards), and represent a broad consensus of the education community (participating organizations included the Canadian Education Association, Canadian School Boards Association and Canadian Teachers’ Federation among others). Intended for both developers and users of assessments (the former being those “who construct assessment methods and people who set policies for particular assessment programs”; the latter are those “who select and administer assessment methods, commission assessment development services, or make decisions on the basis of assessment results and findings” (p. 3)), the principles include the following:

12 In the U.S., a project by the National Forum on Assessment, a coalition of education and civil rights organizations, led to the development in 1995 of Principles and Indicators for Student Assessment Systems. The first principle asserts that the primary purpose of assessment is to improve student learning.
In order to achieve a proper dispersion in the scores [to facilitate making comparisons], questions that are almost always answered correctly or incorrectly tend to be avoided. Aspects of performance that should be tested may therefore not be represented at all because they make no contribution to score dispersion. This leads to overtesting of minutiae and undertesting of important conventions.

Standardized tests stress the product of learning, not the process. They measure a student’s ability to recall facts, define words, and perform routine calculations, not higher learning processes such as analyzing, synthesizing, forming hypotheses, and exploring alternative ways of solving problems.

More crucially, standardized tests have nothing to do with performance standards but only measure relative standards.

Standardized tests do not serve as a useful diagnostic tool because they do not show students where they went wrong; corrected tests are never returned, excluding the process of feedback and debate that is at the core of learning and education.

Standardized tests systematically rule out students’ independent creativity and criticism by eliminating the need to construct answers for themselves.

Rather than increasing accountability, standardized tests merely shift it from teachers and school authorities to anonymous government officials or corporate bureaucrats who cannot be confronted or held accountable if tests are poorly constructed, administered, or marked.

Many studies have shown that standardized tests are biased against socioeconomic, racial, and ethnic minority groups.

Teachers are induced to teach to the tests rather than for learning with the result that curriculum is becoming increasingly test-driven (Meaghan & Casas, 1995d, p. 47).

This phenomenon is well-documented in the research literature. Many teachers have reported that they align their curriculum with what will be tested, are concerned that standardized testing narrows curriculum, and feel pressure to improve test scores (Meaghan & Casas, 1995c, pp. 38-39). For example, a national survey of teachers conducted by the Boston College Center for the Study of Testing, Evaluation, and Educational Policy reported that teachers with a large percentage of minority students in their classrooms (over 60%) were more likely to teach to the standardized tests, emphasizing basic academic skills and time spent in test-taking activities, than were their counterparts who taught mainly White students (Madaus, 1994, pp. 83-84). Wideen et al. (1997) examined the impact of Grade 12 school-leaving examinations on the teaching of science in several British Columbia school districts. They concluded that the high-stakes tests reduced the variety of instructional approaches used, particularly at the Grade 12 level, because they “discouraged teachers from using strategies which promoted enquiry and active student learning”, and that “this impoverishment affected the language of classroom discourse.” (p. 428)
Other problems

Additional problems associated with over-reliance on standardized tests include:

✓ neglecting essential educational services by allocating scarce financial resources to expensive (and questionable) standardized testing programs at a time of severe government cutbacks to education (Meaghan & Casas, 1995b, p. 87); the testing of all Ontario Grade 3 and Grade 6 students in 1999 is expected to cost about $12 million;

✓ disregarding variations in curriculum and curriculum organization for students in particular grades among different countries when comparing international test results (Ireland, 1995);

✓ the inability of information provided by standardized tests to assist educators in identifying and improving ineffective school programs (Casas & Meaghan, 1995, p. 19);

✓ shifting responsibility over curriculum from teachers and school authorities to the testing industry and government officials, thereby lessening rather than enhancing accountability (Meaghan & Casas, 1995d, p. 47).

In the United States, the production of standardized tests is mainly in the hands of commercial publishers and non-school agencies (Darling-Hammond, 1991, p. 220). The situation in Canada is similar – standardized tests are developed and marketed by the major textbook publishers with profits going to the U.S.-based parent companies (Barlow & Robertson, 1994, p. 120). The U.S. private testing industry is a billion dollar a year business which, despite its growing influence on the education system, is virtually unregulated and unaccountable. Under such conditions, increased involvement by teachers and parents in the educational decision-making process is ruled out (Medina & Neill, 1990).

Appendix A — Summary of Problems with Standardized Testing

A considerable body of research has identified numerous problems associated with the nature and use of standardized achievement tests. The following is a partial list (excerpted from Meaghan & Casas, “On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora’s Box?” Interchange, 26(1), 1995, pp. 34-35):

✓ To permit machine scoring, standardized tests are limited to multiple-choice items which, in turn, severely limits what can be tested. For example, you can test language skills but not writing ability or how students use language.

✓ To ensure comparability of test scores over time, few substantive changes are made over the years, which hardly encourages updating of curricula.

✓ Although advocated as a means to carry out international comparisons of student learning and performance, these tests are not sensitive to differences in curricula for particular ages and grades in different countries (nor can they be made so).

✓ Because each question is assigned the same score value, the student who answers more of the difficult (and/or important from the program standpoint) questions is judged the equal of the student who gets only the easy ones.
implemented in tandem that address broad social and educational conditions. If a primary goal of our education system is to offer all students equality of educational opportunity, then standardized testing will only move us further away from this goal.

Testing and Inequity in Education

Not only is there little evidence to support standardized testing from a pedagogical standpoint, it is clear that standardized testing can have a detrimental impact on the most vulnerable students. Far from being a neutral practice, it perpetuates and intensifies educational inequities in two ways: through the misuse of test scores; and because test bias works against the interests of students from low-income groups, racial and ethnic minorities, girls and young women, and students with disabilities. Darling-Hammond (1994) reminds us that “[t]he role of testing in reinforcing and extending social inequalities in educational opportunities has by now been extensively researched ... and widely acknowledged.” (p. 10)

Inappropriate use of tests

Tests are often used inappropriately. The history of testing reveals many instances of tests developed for a specific purpose being used for other unintended purposes. The Scholastic Aptitude Test or SAT is a prime example. Originally designed to predict a student’s potential for success in their first year of college in the U.S., the SAT has been used as a criterion for determining such things as athletic eligibility, student loan eligibility and awarding scholarships (Madaus, 1991, p. 227). Despite its limitations (including its “limited predictive value”), Schwartz (1999) observes that the SAT “has arguably become the single most important test for American high-school students – an academic and psychic rite of passage that strongly influences future educational options, prompts fierce anxiety and serves as an almost mystical barometer of self-worth.” (p. 30)
Tests of individuals are used to analyze policy, program, school and teacher success, and they are being inappropriately used as “educational gatekeepers” to make important decisions about students, teachers, schools, and the school system as a whole (Medina & Neill, p. 24). Standardized test scores are becoming the mechanism that facilitates a number of questionable education practices that contribute to education inequity.

**Tracking**

In the U.S., test scores are used routinely to make important decisions about student placement. The “tracking” that results streams children from disadvantaged socio-economic backgrounds into average and low-track classes and programs. Most progressive educators conclude that the practice of tracking fosters a negative self-image and creates reduced expectations for a student’s future. It also denies those in the lower tracks the benefits of working with higher achievers. In Canada there is a danger that standardized testing will be used as a mechanism to renew interest in tracking thereby moving us away from a more inclusive model of education (Meaghan & Casas, 1995a). Tracking accounts in large part for the disparity in achievement observed between American White students and those from minority groups, as well as between students from higher- and lower-income groups (Darling-Hammond, 1991, p. 222). Perhaps because curriculum in “low-track” classes tends to be limited and rote-oriented, students in these classes don’t do as well as their peers of comparable aptitude in academic or untracked programs. Oakes (1985) concludes that tracking hurts low-income and minority group children, and that it not only reflects but perpetuates class and racial inequalities in the larger society. Using a method of testing already biased against certain groups of students to subsequently assign them to lower track programs only adds insult to injury.

**Grade promotion and retention**

Standardized test results have also been used to determine grade promotion and retention. The practice of grade retention can contribute to an increased drop-out rate (Medina & Neill, p. 29) and a means by which individuals or organizations take responsibility for their actions so that those who depend on them can be assured that there are some safeguards in place to encourage good practices and to prevent bad practices or abuses, to have some course of redress for problems that arise, and to have some assurance of equitable and fair treatment.

With this in mind, she notes that in a genuine, comprehensive system of educational accountability, the various players, from the federal and provincial governments down to the individual school and teacher, have important roles and responsibilities they must fulfill and for which they must be held to account. For example, the state or province “should be responsible for resourcing schools so that they can provide adequate education, for ensuring that there is equity in the distribution of those resources, [and] for assuring that there is a means by which qualified, well-prepared staff will be in all schools in those settings.” Local school districts are responsible “[f]or practices that are supportive of good teaching and learning in schools, [and] for processes that support [these practices].” Teachers are responsible for “making good educational decisions on behalf of children”. Darling-Hammond (1999) describes the latter as professional accountability which means investing in teachers’ skills and knowledge. This is a central component of a genuine accountability system, one which requires providing teachers with ongoing access to the knowledge and support necessary to make sound decisions in order to improve student learning. Unless and until we hold all of the actors accountable for their diverse responsibilities, we can only speak of accountability in a vacuum.

In sum, there is little evidence to support standardized testing from a learning standpoint. Moreover, standardized testing has been shown to have a negative impact on learning, and is particularly harmful to at-risk students – it perpetuates and worsens educational inequities. Assessment, including authentic assessment, must be viewed first and foremost as part of a diverse strategy to improve the quality of education and address equity issues – other systemic reforms must be
improvement originates. There are three groups of
decision-makers who determine the quality and therefore
the impact of American schools. In order of importance,
they are students themselves, their teachers and their
parents. All three are informed in their decision-making
by the same source of information on student
achievement and that source is not national, state or local
standardized testing. They all rely on teachers’ day-to-
day classroom assessments of student achievement. For
this reason, the path to school improvement lies not
through more standardized testing, but through the
development and use of the best quality classroom
assessment we can generate!” (as quoted in Earl, 1995,
p. 7)

In addition to acknowledging the centrality of teachers to the
assessment and evaluation process, we must recognize the value of
the multiple goals of education – goals that go beyond simply preparing
students to enter the labour force as important as that may be. There
is also a need to evaluate evidence of all of these goals, not just cognitive
abilities, with a view to improving the quality of education. Although
effectively evaluating such goals as preparing students to assume the
responsibilities of good citizenship is an important and complex task,
“that complexity is precisely what makes those goals so important to
teach.” (Barlow & Robertson, p. 119)

Reliance on a single form of evaluation, particularly one as flawed as
standardized tests, for making important decisions about students,
teachers and schools is irresponsible. Equally important is the danger
of relying on one form of evaluation for several different purposes
(whether student progress, teacher performance, or program and
system effectiveness). The correct evaluation tool needs to be used
for the task at hand (Earl, 1995, p. 6).

We need to revisit the concept of accountability and what it actually
means. Contrary to current trends in assessment policy, testing by
itself does not constitute accountability. Darling-Hammond (1999)
reminds us that accountability is

according to Shepard and Smith, “‘[c]ontrary to popular beliefs,
repeating a grade does not help students gain ground academically and
has a negative impact on social adjustment and self-esteem.’” (as quoted
in Darling-Hammond, 1991, p. 222) Grade retention has also resulted
in lower achievement. Given the lower levels of self-esteem generally
observed among girls as compared to boys, the implications of this
policy for further eroding self-esteem among girls and young women
requires serious consideration.

Graduation

Test scores have been used to determine a student’s eligibility for high
school graduation. According to Darling-Hammond (1994),

[the rationale for this practice is that students should show that they have mastered the “minimum skills”
needed for employment or future education in order to
graduate. The assumption is that tests can adequately
capture whatever those skills are. While this appears
plausible in theory, it is unlikely in reality, given the
disjunction between multiple-choice tests of
decontextualized bits of information and the demands
of real jobs and adult tasks …. In fact, research indicates that
neither employability nor earnings are significantly affected by
students’ scores on basic skills tests, while chances of employment
and welfare dependency are tightly linked to graduation from high
school [italics added] …. Thus, the use of tests as a sole
determinant of graduation imposes heavy personal and
societal costs, without obvious social benefits. (p. 14)

Nineteen U.S. states currently require students to pass state tests to
graduate from high school (Education Week on the Web, 1999). In
Canada the majority of provinces now require graduating students to
take high school exit exams with the stakes ranging from 30 to 50 per
cent of the final course grade (Dunning, 1997). Once taken as evidence
of years of cumulative development, the high school diploma may
eventually come to be replaced by a one-shot, high-stakes test.
Other manifestations of test misuse

Misuse of the results of province-wide assessment programs in Canada includes “inadequate representation of test results by politicians and the media and inadequate explanation of test results to the public (for example, no background information or comparison of samples with differing characteristics).” (Gilliss, 1994, p. 6) A survey of provincial ministries and departments of education as well as a sample of local school boards across the country also reported that a major drawback of province-wide testing programs is the misinterpretation of results by the public and the media (Traub, 1994, pp. 1, 3).

When 90% of the students in an Ontario Grade 9 reading and writing test fell into the category of having “adequate” reading skills, results that received broad coverage in the media, the North York School Board vice-chair “announced that these results were not good enough for his Board” – this despite the fact that he had not been involved at any level in the preparation of the test, had no indication of how reliable the scores were, or whether the student population in North York [in Metropolitan Toronto] has special characteristics such as a higher-than-average proportion of ESL [English as a second language] students. (Meaghan & Casas, 1995b, p. 83)

Standardized test results are used in some U.S. jurisdictions to make decisions about teachers with respect to merit pay awards as well as certification and recertification. Test-based decisions impacting on schools and school systems include such things as allocation of funds, and school system certification and decertification (Medina & Neill, p. 24). To use average school test scores as the basis for allocating rewards and sanctions to teachers and schools, a practice several U.S. states and school districts have already put in place, is to endorse a market-based approach to stimulating improvement in the school system. Former U.S. President Bush proposed using a national test to

Some General Observations

Shifting from the global to the classroom context, it is important to emphasize that teachers recognize the importance of evaluating student learning and of finding the best means of classroom-based assessment to accomplish this. They regularly evaluate learner progress using various methods which they design and administer – these include term exams, quizzes, essays, projects, portfolios and classroom observation to name a few. Robertson (1998b) describes the diverse formal and informal evaluations carried out by teachers as being “remarkably good predictors of their students’ future success, both in other courses and in life” (p. 65). According to Lorna Earl (1995), Professor of Theory and Policy Studies in Education at OISE/UT, “classroom teachers and their daily classroom assessments may form the basis for the most important accountability relationships.” (p. 6) Because of their regular contact and close relationship with students, teachers are in the best position to conduct ongoing assessments of students, provide regular feedback on their performance to parents and to the students themselves, and recommend ways in which students can learn more effectively. Quoting from R. Stiggins’ book Student-Centered Classroom Assessment, Earl (1995) makes the point that the rationale underlying the importance of classroom assessment in the improvement of learning is also the reason that large-scale assessments cannot achieve this:

“...all of the centralized testing programs operate on the assumption that the decisions that exert the greatest influence on school quality are made at school management and policy levels; that is, somewhere outside the classroom....This view of the origin of school improvement is naive and obviously counterproductive. The halls of political power are not where school
At the same time as the state appears to be devolving power to individuals and autonomous institutions that are themselves increasingly competing in a market, the state remains strong in key areas…. In many ways, this is actually one of the hallmarks of neo-liberalism. Indeed, one of the key differences between classical liberalism and its faith in ‘enterprising individuals’ in a market and current forms of neo-liberalism is the latter’s commitment to a regulatory state. Neo-liberalism demands the constant production of evidence that one is in fact ‘making an enterprise of oneself’…. Thus, under these conditions, not only does education become a market commodity like bread and cars in which the values, procedures, and metaphors of business dominate, but its results must be reducible to standardized ‘performance indicators’ (as quoted in Apple, 1998, p. 25).

In a market-based system, testing and performance indicators thus provide the necessary evidence that schools and school systems are “making an enterprise of themselves” thereby satisfying demands for market accountability – at the expense of genuine public accountability.

allocate federal funds to schools in “America 2000”. Such a policy would provide an incentive to schools to close their doors to students who might “bring down” the average scores, such as children with learning and other disabilities, those whose first language is not English, “or children from educationally disadvantaged environments.” (Darling-Hammond, 1991, p. 223) It could also result in experienced teaching staff avoiding schools which serve disadvantaged kids, thereby closing off further educational opportunities for these children.

Citing a study of a large urban American school district that used average school test scores to make decisions about rewards and sanctions for individual schools, Darling-Hammond (1991) ominously describes the impact as “the widespread engineering of student populations”. The authors of the study state that

“[s]tudent selection provides the greatest leverage in the short-term accountability game …. The easiest way to improve one’s chances of winning is 1) to add some highly likely students and 2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric.” (as quoted in Darling-Hammond, 1991, p. 223)
scores without necessarily upgrading skill levels.
(Meaghan & Casas, 1995c, p. 49)

As Elliot Eisner remarked in a presentation at the B.C. Teachers’ Federation AGM in March 1998, “[i]t is perfectly possible for a school’s faculty to raise test scores and at the same time to diminish the quality of education.”

High-stakes, measurement-driven accountability

A recent *Education Week* survey of the accountability policies of 50 states (with the less than subtle title “Quality Counts ’99: Rewarding Results, Punishing Failure”) reveals the strong push for outcome-based accountability to leverage educational reform in the U.S. According to the report, among the components (described as “essential steps”) necessary for putting a comprehensive system of accountability in place are tests, standards, report cards, rating systems, and rewards and sanctions. Using these components as the basis for its analysis, the survey found that:

√ 48 states have state-wide testing programs (13 of which use only multiple choice tests).

√ 36 states publish annual “report cards” on the performance of individual schools.

√ 19 states evaluate and issue public ratings of school performance including “low-performing” schools.

√ 16 states have the power to close, take over or “reconstitute” a failing school (reconstitution is described in the report as “replacing all or most of a school’s staff and reopening it under new management.”) Not surprisingly, schools attended by poor and minority students are more likely to have sanctions applied to them.

measures such as high-stakes standardized testing and performance indicators not only embody dominant cultural values associated with competition, efficiency, performance, and outcomes (the more concrete and measurable, the better). They are also an important link between the parallel and apparently contradictory tendencies in education reform of decentralization (weak state) and centralization (strong state). As decentralization takes shape in reforms such as site-based management in which decision-making responsibility is devolved to the school level, and school choice in which power is devolved to individual “consumers” of education, governments are centralizing power over funding, curriculum, standards and assessment. As noted earlier, assessment of centrally-set outcomes and the subsequent publication of results enables parents to make comparisons and select among schools – thus facilitating a measure of accountability to the market for the school system. In his review of Whitty, Power and Halpin’s book *Devolution and Choice in Education: The School, the State and the Market*, Apple (1998) discusses the authors’ concept of the “evaluative state”:

third are those who will do reasonably well after international economic restructuring; the [other] two-thirds will not” and “everything is ordered to suit the market and the unbridled pursuit of profit.” Noting that the “assault on public education is about creating McSchools to match and serve McWorld”, she demonstrates how high-stakes testing can be used as the
education, are hardly surprising in a world in which the prevailing ideology is one of globalization.11

The influence of international bodies such as the Organisation for Economic Co-operation and Development (OECD) and the Asia-Pacific Economic Community (APEC) on education policy is being increasingly felt. In his analysis of the APEC position paper on education prepared by the South Korean government, Kuehn (1997b) notes that the “globalization agenda for education” relies on the premise that “globalization is inevitable and education [along with every other sector of society] must comply with its requirements”. A major goal of education in the global economy is the preparation of workers for business, hence “business should have a central role in determining the content of schooling”. The content of education becomes narrowed to work ethic, attitudes, and skills “appropriate for a working life”. In this context standardized testing becomes a convenient means of regulating the production of the necessary “human capital”. To illustrate, Kuehn (1997a) cites the case of Mexico where a standardized examination, the Examen Unico, has been created to channel more students into vocational/technical programs that produce the skills required for the global economy and away from university programs “primarily aimed at the specific social and cultural development of Mexico.” (p. 86) Opposition to the Examen Unico was strong to say the least. In August 1996, anti-test demonstrators blocked busy streets in Mexico City and forced a one-day closure of the Mexican stock exchange.

Robertson’s (1998a) analysis of high-stakes testing in the context of globalization illustrates well how testing is becoming an important lever for a number of major education reforms. Globalization is moving us inexorably towards what she describes as McWorld – a highly inequitable one-third/two-thirds world in which “[t]he one-

11 For a summary of some of the main features of globalization, see “On Globalization” by Larry Kuehn, British Columbia Teachers’ Federation [available online at: http://www.ctf-fce.ca/e/what/other/onglobe.htm].

√ 14 states provide monetary rewards for individual schools based on performance.

√ 2 states have attempted to link the evaluation of individual teachers to student performance.

The report contrasts this “hard approach to accountability” with what it describes as a “soft approach” – the latter involves providing schools and students with sufficient resources, support and encouragement as well as enhancing professional accountability. The concept of professional accountability recognizes the critical link between teacher quality and student achievement and hence, the importance of investing in teacher professional development and growth.

In March the Alberta government announced its own version of outcome-based accountability for K-12 education. According to an Alberta Education news release (March 11, 1999), it intends to spend $66 million for a School Performance Incentive Program (SPIP) described as an “innovative initiative designed to improve student learning” by awarding additional funding to school boards “based on how well they achieve improvement targets.” Targets will be set by the province in conjunction with local boards. Allocation of funding awards (ranging from 2-4 per cent of a board’s salary budget) is conditional on improvements in learning as measured by provincial achievement test and diploma exam results, high school graduation rates, and “locally determined performance indicators” such as student attendance and results from local achievement tests. The program which is voluntary (and apparently unprecedented in Canada) will be implemented in the 2000/2001 school year. As noted above, Americans are more familiar with such initiatives – 14 states including North Carolina, Maryland and Texas have instituted some form of financial incentive program (“Alberta Proposes Bonuses for Student Achievement Gains”, 1999).

The Alberta Teachers’ Association (ATA) is concerned that a program which essentially offers cash bonuses in return for improved test scores, particularly at a time of budget cutbacks, “will pit schools and boards against each other and unfairly favor well-to-do districts.” (McMahan,
Other questions raised by the SPIP include: What will be the impact on teacher collegiality? Will the program result in increased teacher requests to transfer to schools with higher student achievement? What is the potential fallout from identifying schools and classrooms that don’t meet the target objectives? (“Alberta Proposes Bonuses for Student Achievement Gains”, 1999) Arguing that such incentive programs are inconsistent with what is known about effective teaching and student evaluation practices, teacher motivation and school improvement generally, the ATA and its provincial education partners have proposed an alternative to the SPIP designed to improve teaching and learning based on proven research and practice. The Minister of Education has agreed to put the SPIP on hold pending discussion of the alternative program (“Improvement, Not Incentives”, 1999).

The so-called hard approach as described in the Education Week survey is clearly not the path to genuine accountability in education as advocates suggest. Darling-Hammond (1994) explains the negative impact on equity of such an approach:

> Needless to say, this kind of policy that rewards or punishes schools for aggregate test scores creates a distorted view of accountability, in which beating the numbers by playing shell games with student placements overwhelms efforts to serve students’ educational needs well .... Applying sanctions to schools with lower test score performance penalizes already disadvantaged students twice over: having given them inadequate schools to begin with, society will now punish them again for failing to perform as well as other students attending schools with greater resources and more capable teachers. This kind of reward system confuses the quality of education offered by schools with the needs of the students they enroll; it works against equity and integration, and against any possibilities for fair and open school choice, by discouraging good schools from opening their doors to educationally needy students. Such a reward structure is a powerful force for curriculum reform and for shaping the goals of education. What has been and will be tested becomes a priority, not just in the public mind or ministry policy, but in time and attention in the classroom. Other instructional priorities recede, thus shifting curriculum balance effortlessly and without debate. (pp. 118-119)

This is being borne out south of the border. Over the past 150 years, testing requirements in the U.S. have had a number of unintended yet dramatic effects in terms of education policy. These include the reshaping of curriculum as well as attitudes about the goals of education (Madaus, 1994, p. 79). The impact of institutions such as the Educational Testing Service (ETS), the world’s largest private testing firm, on education are significant and cannot be underestimated. Lemann (1995) wrote in the Atlantic Monthly that the ETS and its tests are a national obsession, deeply worked into the fabric of middle-class life. They have generated a large, independent test-prep industry and have strongly influenced elementary and secondary education, one of whose central goals is preparation for ETS tests. (p. 42)

The bigger picture of educational reform

The trends in assessment cited in the introduction to this report must be situated within the context of broader trends in educational change, among them: reduced funding, reforms increasingly being justified in economic terms (i.e. education as the means to prepare individuals for the labour market and nations for global competition), and a move toward market-based solutions which view education as a commodity and students and their parents as consumers of that commodity (Levin, 1998). Such trends, motivated more by fiscal, political and ideological considerations than by a sincere desire to improve access to high quality
As the SAIP nears the end of its second full cycle and concerns begin to surface about aspects of the program (such as the process by which expectations for student performance are set), there are indications that it is influencing, as Robertson (1999) has observed, a “remarkable degree of curriculum homogenization within a country that has considered local control of curriculum to be a virtue rather than a liability.” (p. 715) The creation by regional consortia in the Western and Atlantic provinces of outcomes-based curricula aligned with the SAIP reflects this trend.

The CTF paper goes on to emphasize the importance of having broad and diverse educational goals drive appropriate educational means:

The fundamental error in the CMEC project is the failure to recognize that testing needs to be driven by the engines of educational goals and curriculum and instruction. Otherwise, the test results are likely to become an end in themselves and lead to undesirable practices in the schools .... while testing has a role to play in the development and improvement of education, it appears that the CMEC School Achievement Indicators Program falls into the category of testing which is mainly designed for monitoring and control. (pp. 44, 46)

Meaghan and Casas (1995b) describe the SAIP as a classic example of “placing the cart before the horse” – the development of a national testing program in the absence of public consensus on just what our educational goals should be and how these goals should be reflected in curriculum. This approach will result in “a handful of bureaucrats and consultants” deciding what it is they think students should know (p. 88).

As noted earlier, curriculum is becoming more test-driven and less responsive to student and community needs. Education means are dictating narrowly defined education goals. Barlow and Robertson note that standardized testing places more emphasis on score manipulations and student assignments or exclusions than on school improvement and the development of more effective teaching practices. (pp. 15-16)

In the context of U.S. proposals to introduce national testing, Madaus and Kellaghan (1993) analyze the use of testing as an instrument to drive educational policy noting that “the link between test performance and important rewards or sanctions is the key ingredient that gives tests their policy punch” – hence the term “high-stakes testing”. In an article entitled “Testing as a Mechanism of Public Policy: A Brief History and Description”, they outline six basic principles to describe the processes “through which tests used in the policy sphere affect individuals and institutions”. Their observations are insightful. (Excerpted from Madaus & Kellaghan, 1993, pp. 8-9):

1. The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon – in other words, if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false – the effect is produced by what individuals perceive to be the case.

2. If important decisions are presumed to be related to test results, then teachers will teach to the test.

3. In every setting where a high-stakes test operates, a tradition of past exam questions develops, and this tradition eventually comes to define de facto the curriculum.

4. When test results are the sole or even partial arbiter of future educational or life choices for students, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.

5. A high-stakes test transfers control over the curriculum to the agency that sets or controls the exam.
6. There is a very general principle that is true of all important social indicators, including tests, and which relates to a process that is probably always unintended and very often not anticipated. The principle embodies the proposition that the more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.

Authentic forms of student assessment as an alternative to standardized testing are not immune to these effects. There are some indications that when authentic assessment (which will be discussed in more detail in a subsequent section of this report) is tied to high-stakes accountability, the instructional benefits of such assessment are undermined (Jones & Whitford, 1997; Torrance, 1993). For example, Kentucky’s high-stakes accountability system converts the results of performance-based assessments into a “school score” and relies heavily on the use of external rewards and sanctions. According to Jones and Whitford (1997),

[i]f the score exceeds the state’s expectation for a school, the teachers and administrators can receive substantial bonuses. If the score is not high enough or does not continue to improve over time, the teachers and administrators can be placed on probation, and the school can be taken over by the state. (p. 276)

Particularly frustrating for teachers and administrators was the state’s decision, lacking empirical data, to design the formula for determining school target scores (as a means of measuring the rate of school improvement) in an arbitrary manner.

One of the effects of KIRIS (Kentucky Instructional Results Information System) has been increased state control over curriculum at the expense of local control and the professional judgement of teachers. As Jones and Whitford explain,

The CMEC’s emphasis on assessment-driven accountability derives largely from the current fixation with, and anxiety produced by, global market competition and the resultant need to know “exactly where one stands in comparison to the competition, both nationally and internationally, [as] the first step in becoming armed for the battle.” (Christensen, 1998, p. 76) While the CMEC has been unresponsive to concerns about the SAIP from educational organizations including teachers’ organizations, it has “chosen to privilege certain perspectives” (p. 92), namely those of business interests.

A Canadian Teachers’ Federation (1992) draft position paper on accountability and testing in education raised some important issues with respect to how the various jurisdictions might respond to the initial results of the CMEC’s national testing project and the longer-term, far-reaching impacts of the SAIP. It stated that:

It is axiomatic that some provinces will rank lower than others (in fact, the results are almost predictable, on the basis of knowledge already available). What will these provinces do? Will they agree to adopt the curriculum and organizational strategies of the more successful provinces (thus moving towards a national core curriculum)? Will they attempt to modify existing curriculum so that it is more closely aligned with the tests? Will school boards seek to replicate the tests, thus making identification of individual schools and teachers more of a possibility? Will teachers be encouraged to teach to the next round of tests? Will the end result of the testing program be a pedestrian, overstandardized educational system in which there is no room for creativity or autonomy on the part of students or teachers and where the most neglected goals are those indeterminate aims of education which all agree are particularly important? (p. 44)
The Committee also called for national standards in education and, in a section on the need for employability skills testing, recommended that “the assessment of basic [employability] skills, once they have been defined, be incorporated into the SAIP as it develops and expands its mandate.” (p. 23)

Christensen’s (1998) analysis of the CMEC, from its origins in the Canadian Education Association to its emergence as a self-described “national forum on educational policy”, documents the growing influence of an organization which she notes “has largely escaped critical inquiry since its birth thirty years ago” (p. iii), is not accountable to any government and, despite its stated commitment to consult and cooperate with other educational organizations, is “exclusionary in nature”. In a chapter on the CMEC’s work in curriculum and student assessment (whose early initiatives in these areas in the 1980’s signalled the start of the CMEC’s rising influence), the author describes how federal government pressure in the early part of the decade “precipitated the Council’s hasty implementation of SAIP in 1993.” (p. 83) Federal interest in addressing “public discontent about the educational ‘crisis’ in Canada” was articulated in a 1991 throne speech in which the Prime Minister referred to the need for national education goals, marking as Christensen notes “the return of the federal government to discussions about ‘national’ education polices” (p. 83). (Interestingly, at about the same time the U.S. government released its America 2000 education strategy calling for a system of national standards and assessment.) Mindful not to encroach on provincial jurisdiction, the vehicle chosen by Ottawa to address education issues was the Canadian Forum on Learning, set up through the Steering Group on Prosperity (Prosperity Initiative), to “lead a process to develop a statement of expectations for education and training in Canada’ ” (p. 84). At a meeting of the “Learning Forum” organized through the Prosperity Secretariat and the Conference Board of Canada in 1993, federal MP Bernard Valcourt stated that the legacy of the Prosperity Initiative was to “link competitiveness and learning in the minds of Canadians’ ” (Robertson, as cited in Christensen, 1998, p. 86). Christensen suggests that criticism of the CMEC’s record by the federal Steering Group on Prosperity was one of the factors propelling the SAIP agenda forward.

Rather than generating constructivist, performance-based classrooms as was perhaps the intent, KIRIS has instead emphasized “learning that can be measured reliably and validly” – the testing of observable, measurable outcomes effectively determines what is worth knowing. While performance assessment holds promise for improving student learning, this potential is corrupted in a high-stakes environment. The authors conclude that the principles underlying high-stakes accountability and performance assessment are fundamentally incompatible: the former “encourages conformity to externally imposed standards, while the [latter] grows out of emergent interactions between teachers and students in dynamic local classrooms.” (Jones & Whitford, p. 280)

**Bias in testing**

Standardized tests are typically gender-biased. Rosser et al. (1992) observe that while young women consistently obtain higher grades in both high school and college in the U.S., they receive lower scores on
Standardized Testing

the Scholastic Aptitude Test (SAT) and similar tests which allegedly predict students’ future academic performance. High school girls aspiring to attend college represent over 52% of those writing these tests. However, their abilities are systematically underestimated by these exams resulting in reduced opportunities for attending post-secondary institutions, for obtaining scholarship funding, and for acceptance to special programs for gifted students. Students who receive low SAT scores have reduced expectations of their abilities which in turn affects their decisions about which colleges they apply to (Weiss, 1991, p. 108). The longer term economic and personal consequences for women include fewer good employment opportunities, lower income levels and reduced self-confidence. Female students from low-income backgrounds face the combined gender and class biases of the exams and are effectively penalized twice (Rosser et al., 1992). In short, SAT scores tell us more about factors such as family income, race and gender than they do about one’s ability to perform well in university (Robertson, 1998b, p. 80).

Robertson (1992) discusses the impact of testing on girls in the context of math, science and technology (MST) education. She notes that “complex indicators of ‘success’ for students and for education are frequently and inappropriately reduced to single-test numbers” (p. 25) and that there is clearly an important political agenda underlying this direction in education reform. For high school students, small differences in test score results not only decide whether one passes a

5 Even the world’s largest private educational test maker, the Educational Testing Service (ETS) which develops and administers the SAT, has admitted the SAT’s bias in terms of over-predicting male college performance while under-predicting the college performance of females. ETS also acknowledges that high school grades are a better predictor of college success than their tests (Thurmond, 1994). In a landmark U.S. federal court decision on the discriminatory nature of the SAT, the judge stated that “…[T]he probability, absent discriminatory causes, that women would consistently score 60 points less on the SAT than men is nearly zero.” Even after accounting for such factors as ethnicity and parental education, a substantial gap in points remained (Rosser et al., 1992, pp. i-ii).

(SAIP)10, a national project initiated in 1989 by the CMEC to test Canadian students aged 13-16 in math, reading and writing, and science, Barker (1991) argues that the SAIP is a politically-motivated initiative developed in response to criticisms of education and calls for increased accountability from business, industry and government. She suggests that the SAIP can be viewed as a catalyst for educational reform in Canada – from more limited changes in terms of institutionalizing national testing, to sweeping reforms such as increased control over education by the federal government and the institutionalization of national education standards with economic productivity. Some of this is echoed in the 1994 report of the National Advisory Board on Science and Technology’s Committee on National Standards in Education. Among its recommendations to the Prime Minister were that:

√ “the federal government assist in the development and administration of systemic tests for all core subjects [italics added] building on the work of the School Achievement Indicators Program”, and that

√ “the frequency of the administration of the tests in all core subjects be increased to that which maximizes the information necessary to ensure continuous improvement. If maximum advantage is gained with yearly testing, then that should be instituted.” (p. 23)

10 In the first cycle of the SAIP, assessments were administered in mathematics (1993), reading and writing (1994), and science (1996). The SAIP’s second cycle included assessments in mathematics in 1997, reading and writing in 1998, and science assessment in 1999. A third cycle has since received approval. Funding for the SAIP comes from the provinces, Human Resources Development Canada at the federal level, and private corporations (Investors’ Group Incorporated and Spar Aerospace Limited). The CMEC estimates total costs for administration and implementation of the program from the planning stage in 1989 through to its completion in 1999 to be approximately $11 million, excluding costs incurred for provincial in-kind support [from the CMEC web site at: http://www.cmecc.ca/saip/]. Although the SAIP has been in existence for the past decade, there appears to be little in the way of critical analysis of the program and its impact.
The growing Pan-Canadian (i.e. national) agenda for education being led by the Council of Ministers of Education, Canada (CMEC), as outlined in its 1993 “Victoria Declaration”, includes major initiatives in the areas of curriculum, student assessment, and education indicators. With respect to curriculum, in 1997 the CMEC released its Common Framework of Science Learning Outcomes: Pan-Canadian Protocol for Collaboration on School Curriculum, a common curriculum framework which lays out the learning outcomes for K-12 science education (this project involved all provinces except Quebec). Robertson (1998b) observes that the “CMEC sidestepped the thorny constitutional problem of curriculum as a provincial matter by declaring the document to be only a framework for curriculum developers, not an actual curriculum. But no one was fooled.” (p. 86) She notes that the curriculum consultation process was also sidestepped. The time available for feedback on drafts of the long document was short and the nature of the commentary was tightly prescribed – “[t]hose wishing to comment were told that they could question neither the desirability of a national curriculum nor its contentious devotion to outcomes.” (p. 86) Teachers’ organizations, school board associations and other groups traditionally involved in curriculum development were effectively excluded from the process.

Curriculum is also being developed regionally through the Atlantic Provinces Education Foundation in the east and the Western Canadian Protocol for Collaboration in Basic Education in the west. Such curriculum reform initiatives are in keeping with an increased emphasis on standardized outcomes and the core subject areas of math, science and language arts (Dunning, 1997).

With province-wide assessment programs in place in most provinces and territories, a full-blown national testing program cannot be far off. In her analysis of the School Achievement Indicators Program course but determine which students will be accepted to select universities (p. 25). As “scarcity” of post-secondary opportunities flows from underfunding of the post-secondary education system, Canadian universities have increased tuition fees as well as reduced the number of spaces available. These factors combined will reduce access to higher education for women, particularly in areas (such as MST) in which they are already under-represented.

The following findings (excerpted from Robertson, 1992, p. 26) are taken from the American Association of University Women (AAUW) report How Schools Shortchange Girls (1992). This important study synthesizes the available research on the educational experience of girls concluding that the education they receive is inequitable in terms of quality and quantity to that received by boys:

✓ Despite girls’ receiving higher grades in high school and college, boys were twice as likely as girls to have received a college scholarship based on the results of standardized high school tests.

✓ Girls have consistently outscored boys on the SAT-verbal when test content refers to concepts and ideas rather than things, or to aesthetics, philosophy, relationships or stereotypically female tasks; whereas boys have scored better on items relating to science or sports; since the SAT-verbal was revised in 1978 to increase the number of science-related items, reading subscores favouring boys have increased from three to twelve points.

✓ As noted earlier, SAT scores under-predict the success of young women, and over-predict the success of young men in college/university achievement, especially in math.

✓ Bias in test items can occur within content, format and the skill or knowledge area being tested; for example, in the area of math, girls tend to score higher on computational tasks, algebra and logic, whereas boys tend to score higher on problem-solving and geometry – tests which give more weight to one cluster of skills over another will produce gender differences.
Girls tend to score higher on essay and open-ended items; boys tend to score higher on multiple choice items; girls complete fewer test items than boys, are more likely to not finish tests, and more likely to choose “I don’t know” as an answer when it is offered.

FairTest notes that

[the timed, speeded nature of the [SAT] works against young women and members of minority groups. Guessing is also a barrier for females, and possibly for people of colour. The forced-choice format does not allow for shades of meaning, working against girls' more complex thinking-style. (Rosser et al., 1992, Appendix I)]

In a follow-up report to How Schools Shortchange Girls entitled Gender Gaps: Where Schools Still Fail Our Children which looks at progress toward equity since the early 1990’s, the AAUW Educational Foundation (1998) finds that “high-stakes tests, with disproportionate power to affect students' lives, are the tests that most dramatically reflect gender differences in performance.” (p. 35) Among the report’s recommendations are that post-secondary institutions formulate decisions about student admissions and scholarship eligibility using a “broad range of material”; and that research focus on the relationship between student test scores and grades in the context of gender, race and ethnicity in order to provide information “to develop appropriate equity responses based on individual differences.” (p. 41)

Robertson states that although the AAUW and other data can’t be used as evidence that gender bias is disadvantaging Canadian girls in every MST-related evaluation, she says “it is curious to note how little attention is paid to the apparent risk to young women posed by a massive shift to computer-scored, single-right-answer standardized evaluation.” (p. 27)

manipulated to political advantage. When this fact is combined with the strong pressure to increase test scores and the questionable practices that ensue, the potential for compounding distortions and providing misleading information is very real.

In a system of school choice (which would include charter schools) the so-called “good” schools are able to attract and select the best students. The “creaming off” of motivated, high-achieving students and their more educationally-involved parents by choice schools is beginning to have a negative impact on local public schools which are left with fewer students (those that are higher-risk and higher-cost) and hence, fewer tax dollars (Dehli, 1998, p. 31).

According to the Canadian Teachers’ Federation (1997),

[s]chool-by-school competition creates high demand for the students who already have the greatest chance of success in our schools, and creates incentives to avoid serving harder-to-teach kids and kids with disabilities who may consume more than their ‘share’ of resources, yet not increase the school’s profile in key marketing areas such as standardized test scores. (p. 17)

It should also be noted that when high-achieving students are attracted by and concentrated in choice programs, the illusion of increased achievement may well be created. Fuller and Elmore (as cited in Dehli, 1998, p. 31) caution that “it is unlikely that choice will do anything other than simply move high achievers around from one school to another, mistaking the effect of concentrating strong and motivated students for an effect of the school or the choice system.’”
attitudes between corporate executives and the general public with respect to confidence in schools. Specifically, Hart and Livingstone found little evidence to support the notion of “an impending crisis of public or parental confidence policy preferences of the general public” (p. 16).

Testing, schools and market competition

There is also strong support from the corporate sector for national education standards and national testing. The successful lobbying efforts of organizations representing powerful business interests such as the Business Council on National Issues, the C.D. Howe Institute and the Conference Board of Canada have put standardized testing high on the government and public agendas. Standardized testing fits well with business priorities – computer-grading of multiple choice tests is cheaper and quicker than teachers marking essays and other writing assignments (Barlow & Robertson, p. 118); standardized test scores represent the educational equivalent of the corporate bottom line (Reardon, Scott & Verre, 1994, p. 2); and as noted earlier, there are large profits to be made from educational testing.

Business also supports standardized testing because it is consistent with a free market approach to education. School choice by parents acting in their capacity as consumers of education is advocated as a means of fostering competition among schools. Competition is regarded as the necessary incentive to the achievement of quality education (as “inefficient” schools close and the “good” schools rise to the “market challenge”). In this model, a basis for comparing and making decisions about schools is required: enter the ranking and reporting of average scores on standardized tests as the basis for making these decisions (Berthelot, 1995, p. 4). Of course, rank ordering schools (or districts, provinces or states) by test scores is a seriously flawed practice. Not only are scores often presented with little or no contextual information but small increases in the test scores of individual students can translate into significant changes at the state level in terms of positioning within a rank-ordered table, something which can be

Most standardized tests in the U.S. are designed for, and written by, a middle- to upper-class White population and as a result reflect a particular perspective. Studies have shown that a number of characteristics of standardized tests can bias the results against minority students and students from low-income backgrounds. These include the following (from Medina & Neill, pp. 8-10):

- Language bias manifests itself in English-language tests taken by students whose first language is not English; it manifests itself more subtly in tests written in a more formal, stylized version of English, rather than the common vernacular, for students who speak non-standard English dialects.

- The content of test questions seeks responses which ignore the cultural experiences, perspectives and knowledge of children from racial and ethnic minorities, low-income families, and inner city and rural children.

- Individuals have “different ways of knowing and problem-solving”. These different learning styles (not different abilities), often associated with such factors as race/ethnicity, income level and gender, are not considered in test design, the assumption being “that all individuals perceive information and solve problems in the same way.”

In their research on language testing of Black South African students, Peirce and Stein (1995) found that students’ multiple interpretations of a text for a reading test were socially constructed. Their findings have broad implications beyond the South African context in terms of education inequity. They acknowledge that,

[i]n essence, if test makers are drawn from a particular class, a particular race, and a particular gender, then test takers who share these characteristics will be at an advantage relative to other test takers .... To promote equity in educational assessment, different stakeholders, such
as testers, teachers, administrators, parents, and students should be able to contribute to the test development process. (p. 62)

The injustice of being denied access to post-secondary studies because of scoring low on a biased test is underscored in the following passage by Weiss (1991):

For those minority students who have overcome innumerable social and economic barriers and achieved academic success in early schooling, the denial of access to higher education because of low college or graduate school admissions tests is particularly frustrating. (p. 114)

Coaching courses to prepare students to take tests such as the SAT have evolved into a multi-million dollar industry. The largest SAT preparation companies are the Princeton Review and Kaplan Educational Centers with the latter generating annual revenues approaching $300 million (Schwartz, 1999, p. 32). Weiss notes that a coaching course can increase one’s SAT score by enough to “mean the difference between a rejection notice and college admission with a scholarship.” (p. 118) Such courses however can be expensive – a 6-8 week course including practice tests costs about $800 while private tutors can charge a few hundred dollars per hour (Schwartz, p. 32) – putting them well beyond the means of the majority. This places minorities and low-income students at a double disadvantage – according to Weiss, “[n]ot only are they unable to afford the advantages promised by coaching, but the success of coaching increases the disparity between racial groups even further.” (p. 118)

Misrepresentation of test results (whether at the international, national or provincial level) can heighten the concerns of parents, causing them to feel the education their children are getting is inferior to that received by children in other countries, provinces or districts. Politicians and policy makers respond to the perceived crisis of public confidence by placing an increased emphasis on standardized testing as a measure of accountability. Meaghan and Casas (1995b) state that testing has become the quick-fix solution for school board administrators and ministry of education officials who need to produce “results” and need to do so quickly. Tackling curricular issues, improving the quality of teacher training, [and] developing fair and efficient assessment methods would take too long to produce results and those results would not be as simple to measure and to interpret as test scores are. (p. 84)

Is there evidence to support a crisis of public confidence in schools? Not according to OISE/UT researchers Hart and Livingstone (1998). In their analysis of trends in attitudes about schools drawn from Ontario and national opinion surveys dating back to 1980, they sought to shed light on this question. Their analysis also addressed the question of whether there was “support for an agenda of fiscal austerity and privatization that is the neocorporate prescription for education” (p. 1). Their research reveals sharply divergent

9 The authors define neoconservatism generally as a “New Right political position ... which argues for: (a) a reduction in the role of the state, particularly with regard to social services, redistribution of income, and equity initiatives; (b) a reduction in government deficits and debt; and (c) a reliance on market mechanisms and competition to make public services more efficient through privatization or by bringing greater competition into the public sphere.” (p. 2)
Testing has become an entrenched practice in our society, not only in education but in other spheres. Some have described testing as a form of social technology arguing that, although there is nothing inherently unjust about technology itself, “technologies are a product of the existing structure of opportunities and constraints in society” and as such, “they are likely to extend, shape, rework, or reproduce this structure”, and that includes any inequities built into the structure (Madaus, 1994, pp. 77, 79). In other words, the values, biases and assumptions of the elite group who create technology are reproduced in those technologies. Some of the underlying values that inform testing are utilitarianism, economic competitiveness, technological optimism, objectivity, bureaucratic control and accountability, numerical precision, efficiency, standardization, and conformity. In this way says Madaus (1994), “testing not only shapes important educational values, such as conceptions of educational quality, and student- and teacher-reward systems, but also the content and method of what should be taught and learned.” (p. 79)

Problems with our schools has been fueled in part by sensationalist and distorted media coverage of Canadian performance on international standardized tests. Using performance on international tests as an indicator of problems with our education system can be misleading. Some of the methodological problems associated with international testing of math and science achievement include the following (from Canadian Teachers’ Federation, 1995a):

- Participation in international studies is often voluntary, and not all school systems are involved. Rankings therefore reflect only those countries who choose to participate.
- Participating countries have different ways of choosing the students who will write the tests (some select only their highest achieving students).
- Participating countries are not rated against their own curriculum, and there is no effort to adjust test scores to reflect circumstances in different countries.

Past experience with international studies has clearly demonstrated that “quantitative data can be misleading without detailed qualitative information to put it in context.” (Canadian Teachers’ Federation, 1995a) Nagy (1995) notes that “[i]nternational studies have the potential to develop as a useful tool for understanding and improving education, but they face a conundrum. As they become, of necessity, more sophisticated, their very complexity will make simplistic misinterpretation difficult. This, however, may result in loss of their political appeal and funding.”

Public and political misinterpretation of complex international comparative studies can in turn drive the policy development process, resulting in misinformed decisions. Nagy (1996) makes an important distinction between policy judgements and political judgements – the former are informed decisions that flow from the original reports, accompanying cautions on data interpretation and any relevant academic and specialist analyses; the latter refer to uninformed announcements based on “the media and political view of these

7 In Testing Testing: Social Consequences of the Examined Life, F. Allan Hanson examines the social and cultural assumptions behind current U.S. testing practices. He analyzes testing as a social institution, one so closely associated with the notion of meritocracy that we rarely ask the difficult questions such as “why test at all?” Drawing on the work of Foucault, he proposes that testing (from lie detection and drug use tests to vocational, intelligence and achievement tests) constructs who we are as social beings, and that tests serve as a form of surveillance and control over the individual by organizations and institutions that use and value these tests (Kleinsasser, 1994, pp. 96-97).
Politics of Testing

Standardized testing needs to be viewed with a political as well as a pedagogical eye. Advocated as a measure of educational achievement and a means of addressing our educational problems, standardized testing is politically attractive because, while creating the illusion that something is being done to improve the quality of education (via what is perceived as an objective, neutral mechanism), it allows us to ignore the complex reality of educating children and youth, and the associated demands, time and costs. McMurtry (1992) notes that standardized tests “can be a destructive distraction from what we need most to do and to achieve as educators” and that “they can be politically misused to escape the actual demands of teaching.” (p. 95) In a chapter on the politics of testing in No More Teachers, No More Books: The Commercialization of Canada’s Schools, Robertson (1998b) states that “[i]t is so much easier to increase the quantity of testing than to improve the quality of teaching – or to deal with inequality of opportunity.” (p. 72) She also observes that testing begets more testing:

Whether the alleged deficiency [with schools] is curriculum, standards, or accountability, the strategy is to use data, no matter how corrupt, to leverage education reforms. Thus a key political mechanism is to increase testing – more tests for more kids more often, producing more data, more widely reported and more susceptible to manipulation. (pp. 63-64)

International comparisons

Despite the fact that international comparisons of educational achievement are difficult to do, the available data shows that Canada is near the average for comparable countries, and in some cases fares better (Rehnby, 1996, p. 15). Nevertheless, the perception of serious
A Few Words About Authentic Assessment

Authentic or performance assessment is often proposed as an alternative to standardized testing. FairTest describes performance assessment as a “general term for an assessment activity in which students construct responses, create products, or perform demonstrations to provide evidence of their knowledge and skills” (Neill et al., 1997, Appendix B). Such assessment covers a wide range of initiatives including: oral presentations, debates, exhibitions, collections of students’ written work, videotapes of performances and other learning experiences, constructions and models and their solutions to problems, experiments and the results of scientific and other inquiries, teacher observations and inventories of individual students’ work and behaviour, and cooperative group work (Darling-Hammond, 1994, pp. 5-6). Unresolved issues related to this relatively new form of assessment include the substantial time commitment required for development and administration (authentic assessment

---

8 In her review of F. Allan Hanson’s book Testing Testing: Social Consequences of the Examined Life, Kleinsasser (1994) notes that the results of standardized testing, which, given the limitations of such testing, provide only indirect evidence of a student’s academic achievement, are retained by school authorities as part of the student’s official school record. This evidence is then used to make important education-related decisions affecting the student’s future; whereas the actual work (or direct evidence) produced by a student, whether writing samples like essays or stories, art, science fair projects, audio- and video-tapes of music and dance performance, etc., is usually kept by the student (rather than the school) because it is more valued by its creator. These products however “are much less meaningful in the hands of powerful authorities like schools or employers .... When institutions choose to prize test results that are held by authorities with the power to recommend or not recommend, they send a strong message to test takers about society’s values and evaluation criteria.” (p. 98)
is by nature time- and labour-intensive) and the associated financial costs as well as the trade-off between validity and reliability.

Changes to the form and content of assessment (as in the case of authentic assessment) will not by themselves solve the problems associated with more traditional assessment. Darling-Hammond (1994) stresses that assessment must also be viewed in terms of how it will be used and in particular, how it will be used to support and enhance equity initiatives within the education system. She argues that

*changes in the forms of assessment are unlikely to enhance equity unless we change the ways in which assessments are used as well:* from sorting mechanisms to diagnostic supports; from external monitors of performance to locally generated tools for inquiring deeply into teaching and learning; and from purveyors of sanctions for those already underserved to levers for equalizing resources and enhancing learning opportunities. [italics in original] (p. 7)

At the same time, broader reforms must be implemented including adequate funding and supports for social, health, family and educational resources, what Madaus (1994) refers to as creating “a level playing field for students and schools” in terms of social and educational conditions. He states that

[e]quity issues arising from assessment are not confined to the specific instrumentation, technique, or psychometrics inherent in the test itself. More broadly, equity issues relate to antecedent and current social and educational conditions. (p. 87)

As with any new technology and the aura that envelops it, there is a danger that authentic assessment may come to be viewed as the solution to educational problems, to the exclusion of addressing other factors such as student health and living conditions, teacher training, etc. Madaus (1994) warns against having technological solutions “blind policymakers and the public to the reality that we Americans cannot test, examine, or assess our way out of our educational problems” (p. 78), a caution we would do well to heed in the Canadian context given our growing emphasis on test-based accountability.
is by nature time- and labour-intensive) and the associated financial costs as well as the trade-off between validity and reliability.

Changes to the form and content of assessment (as in the case of authentic assessment) will not by themselves solve the problems associated with more traditional assessment. Darling-Hammond (1994) stresses that assessment must also be viewed in terms of how it will be used and in particular, how it will be used to support and enhance equity initiatives within the education system. She argues that

\textit{changes in the forms of assessment are unlikely to enhance equity unless we change the ways in which assessments are used as well: from sorting mechanisms to diagnostic supports; from external monitors of performance to locally generated tools for inquiring deeply into teaching and learning; and from purveyors of sanctions for those already underserved to levers for equalizing resources and enhancing learning opportunities. [italics in original]} (p. 7)

At the same time, broader reforms must be implemented including adequate funding and supports for social, health, family and educational resources, what Madaus (1994) refers to as creating “a level playing field for students and schools” in terms of social and educational conditions. He states that

\textit{[e]quity issues arising from assessment are not confined to the specific instrumentation, technique, or psychometrics inherent in the test itself. More broadly, equity issues relate to antecedent and current social and educational conditions.} (p. 87)

As with any new technology and the aura that envelops it, there is a danger that authentic assessment may come to be viewed as \textit{the} solution to educational problems, to the exclusion of addressing other factors such as student health and living conditions, teacher training, etc. Madaus (1994) warns against having technological solutions “blind policymakers and the public to the reality that we Americans cannot test, examine, or assess our way out of our educational problems” (p. 78), a caution we would do well to heed in the Canadian context given our growing emphasis on test-based accountability.
A Few Words About Authentic Assessment

Authentic or performance assessment is often proposed as an alternative to standardized testing. FairTest describes performance assessment as a “general term for an assessment activity in which students construct responses, create products, or perform demonstrations to provide evidence of their knowledge and skills” (Neill et al., 1997, Appendix B). Such assessment covers a wide range of initiatives including: oral presentations, debates, exhibitions, collections of students’ written work, videotapes of performances and other learning experiences, constructions and models and their solutions to problems, experiments and the results of scientific and other inquiries, teacher observations and inventories of individual students’ work and behaviour, and cooperative group work (Darling-Hammond, 1994, pp. 5-6). Unresolved issues related to this relatively new form of assessment include the substantial time commitment required for development and administration (authentic assessment

---

8 In her review of F. Allan Hanson’s book Testing Testing: Social Consequences of the Examined Life, Kleinsasser (1994) notes that the results of standardized testing, which, given the limitations of such testing, provide only indirect evidence of a student’s academic achievement, are retained by school authorities as part of the student’s official school record. This evidence is then used to make important education-related decisions affecting the student’s future; whereas the actual work (or direct evidence) produced by a student, whether writing samples like essays or stories, art, science fair projects, audio- and video-tapes of music and dance performance, etc., is usually kept by the student (rather than the school) because it is more valued by its creator. These products however “are much less meaningful in the hands of powerful authorities like schools or employers .... When institutions choose to prize test results that are held by authorities with the power to recommend or not recommend, they send a strong message to test takers about society’s values and evaluation criteria.” (p. 98)
Politics of Testing

Standardized testing needs to be viewed with a political as well as a pedagogical eye. Advocated as a measure of educational achievement and a means of addressing our educational problems, standardized testing is politically attractive because, while creating the illusion that something is being done to improve the quality of education (via what is perceived as an objective, neutral mechanism), it allows us to ignore the complex reality of educating children and youth, and the associated demands, time and costs. McMurtry (1992) notes that standardized tests “can be a destructive distraction from what we need most to do and to achieve as educators” and that “they can be politically misused to escape the actual demands of teaching.” (p. 95) In a chapter on the politics of testing in No More Teachers, No More Books: The Commercialization of Canada’s Schools, Robertson (1998b) states that “[i]t is so much easier to increase the quantity of testing than to improve the quality of teaching – or to deal with inequality of opportunity.” (p. 72) She also observes that testing begets more testing:

Whether the alleged deficiency [with schools] is curriculum, standards, or accountability, the strategy is to use data, no matter how corrupt, to leverage education reforms. Thus a key political mechanism is to increase testing – more tests for more kids more often, producing more data, more widely reported and more susceptible to manipulation. (pp. 63-64)

International comparisons

Despite the fact that international comparisons of educational achievement are difficult to do, the available data shows that Canada is near the average for comparable countries, and in some cases fares better (Rehnby, 1996, p. 15). Nevertheless, the perception of serious
standardized testing

Testing has become an entrenched practice in our society, not only in education but in other spheres. Some have described testing as a form of social technology arguing that, although there is nothing inherently unjust about technology itself, “technologies are a product of the existing structure of opportunities and constraints in society” and as such, “they are likely to extend, shape, rework, or reproduce this structure”, and that includes any inequities built into the structure (Madaus, 1994, pp. 77, 79). In other words, the values, biases and assumptions of the elite group who create technology are reproduced in those technologies. Some of the underlying values that inform testing are utilitarianism, economic competitiveness, technological optimism, objectivity, bureaucratic control and accountability, numerical precision, efficiency, standardization, and conformity. In this way says Madaus (1994), “testing not only shapes important educational values, such as conceptions of educational quality, and student- and teacher-reward systems, but also the content and method of what should be taught and learned.” (p. 79)

Problems with our schools has been fueled in part by sensationalist and distorted media coverage of Canadian performance on international standardized tests. Using performance on international tests as an indicator of problems with our education system can be misleading. Some of the methodological problems associated with international testing of math and science achievement include the following (from Canadian Teachers’ Federation, 1995a):

- Participation in international studies is often voluntary, and not all school systems are involved. Rankings therefore reflect only those countries who choose to participate.
- Participating countries have different ways of choosing the students who will write the tests (some select only their highest achieving students).
- Participating countries are not rated against their own curriculum, and there is no effort to adjust test scores to reflect circumstances in different countries.

Past experience with international studies has clearly demonstrated that “quantitative data can be misleading without detailed qualitative information to put it in context.” (Canadian Teachers’ Federation, 1995a) Nagy (1995) notes that “[i]nternational studies have the potential to develop as a useful tool for understanding and improving education, but they face a conundrum. As they become, of necessity, more sophisticated, their very complexity will make simplistic misinterpretation difficult. This, however, may result in loss of their political appeal and funding.”

Public and political misinterpretation of complex international comparative studies can in turn drive the policy development process, resulting in misinformed decisions. Nagy (1996) makes an important distinction between policy judgements and political judgements – the former are informed decisions that flow from the original reports, accompanying cautions on data interpretation and any relevant academic and specialist analyses; the latter refer to uninformed announcements based on “the media and political view of these

7 In Testing Testing: Social Consequences of the Examined Life, F. Allan Hanson examines the social and cultural assumptions behind current U.S. testing practices. He analyzes testing as a social institution, one so closely associated with the notion of meritocracy that we rarely ask the difficult questions such as “why test at all?” Drawing on the work of Foucault, he proposes that testing (from lie detection and drug use tests to vocational, intelligence and achievement tests) constructs who we are as social beings, and that tests serve as a form of surveillance and control over the individual by organizations and institutions that use and value these tests (Kleinsasser, 1994, pp. 96-97).
as testers, teachers, administrators, parents, and students should be able to contribute to the test development process. (p. 62)

The injustice of being denied access to post-secondary studies because of scoring low on a biased test is underscored in the following passage by Weiss (1991):

For those minority students who have overcome innumerable social and economic barriers and achieved academic success in early schooling, the denial of access to higher education because of low college or graduate school admissions tests is particularly frustrating. (p. 114)

Coaching courses to prepare students to take tests such as the SAT have evolved into a multi-million dollar industry. The largest SAT preparation companies are the Princeton Review and Kaplan Educational Centers with the latter generating annual revenues approaching $300 million (Schwartz, 1999, p. 32). Weiss notes that a coaching course can increase one’s SAT score by enough to “mean the difference between a rejection notice and college admission with a scholarship.” (p. 118) Such courses however can be expensive – a 6-8 week course including practice tests costs about $800 while private tutors can charge a few hundred dollars per hour (Schwartz, p. 32) – putting them well beyond the means of the majority. This places minorities and low-income students at a double disadvantage – according to Weiss, “[n]ot only are they unable to afford the advantages promised by coaching, but the success of coaching increases the disparity between racial groups even further.” (p. 118)

Misrepresentation of test results (whether at the international, national or provincial level) can heighten the concerns of parents, causing them to feel the education their children are getting is inferior to that received by children in other countries, provinces or districts. Politicians and policy makers respond to the perceived crisis of public confidence by placing an increased emphasis on standardized testing as a measure of accountability. Meaghan and Casas (1995b) state that testing has become the quick-fix solution for school board administrators and ministry of education officials who need to produce “results” and need to do so quickly. Tackling curricular issues, improving the quality of teacher training, [and] developing fair and efficient assessment methods would take too long to produce results and those results would not be as simple to measure and to interpret as test scores are. (p. 84)

Is there evidence to support a crisis of public confidence in schools? Not according to OISE/UT researchers Hart and Livingstone (1998). In their analysis of trends in attitudes about schools drawn from Ontario and national opinion surveys dating back to 1980, they sought to shed light on this question. Their analysis also addressed the question of whether there was “support for an agenda of fiscal austerity and privatization that is the neoconservative prescription for education” (p. 1).9 Their research reveals sharply divergent

9 The authors define neoconservatism generally as a “New Right political position ... which argues for: (a) a reduction in the role of the state, particularly with regard to social services, redistribution of income, and equity initiatives; (b) a reduction in government deficits and debt; and (c) a reliance on market mechanisms and competition to make public services more efficient through privatization or by bringing greater competition into the public sphere.” (p. 2)
Most standardized tests in the U.S. are designed for, and written by, a middle- to upper-class White population and as a result reflect a particular perspective. Studies have shown that a number of characteristics of standardized tests can bias the results against minority students and students from low-income backgrounds. These include the following (from Medina & Neill, pp. 8-10):

- Language bias manifests itself in English-language tests taken by students whose first language is not English; it manifests itself more subtly in tests written in a more formal, stylized version of English, rather than the common vernacular, for students who speak non-standard English dialects.

- The content of test questions seeks responses which ignore the cultural experiences, perspectives and knowledge of children from racial and ethnic minorities, low-income families, and inner city and rural children.

- Individuals have “different ways of knowing and problem-solving”. These different learning styles (not different abilities), often associated with such factors as race/ethnicity, income level and gender, are not considered in test design, the assumption being “that all individuals perceive information and solve problems in the same way.”

In their research on language testing of Black South African students, Peirce and Stein (1995) found that students’ multiple interpretations of a text for a reading test were socially constructed. Their findings have broad implications beyond the South African context in terms of education inequity. They acknowledge that,

[i]n essence, if test makers are drawn from a particular class, a particular race, and a particular gender, then test takers who share these characteristics will be at an advantage relative to other test takers .... To promote equity in educational assessment, different stakeholders, such
Girls tend to score higher on essay and open-ended items; boys tend to score higher on multiple choice items; girls complete fewer test items than boys, are more likely to not finish tests, and more likely to choose “I don’t know” as an answer when it is offered.

FairTest notes that

[the timed, speeded nature of the [SAT] works against young women and members of minority groups. Guessing is also a barrier for females, and possibly for people of colour. The forced-choice format does not allow for shades of meaning, working against girls' more complex thinking-style. (Rosser et al., 1992, Appendix I)

In a follow-up report to How Schools Shortchange Girls entitled Gender Gaps: Where Schools Still Fail Our Children which looks at progress toward equity since the early 1990's, the AAUW Educational Foundation (1998) finds that “high-stakes tests, with disproportionate power to affect students' lives, are the tests that most dramatically reflect gender differences in performance.” (p. 35) Among the report's recommendations are that post-secondary institutions formulate decisions about student admissions and scholarship eligibility using a "broad range of material"; and that research focus on the relationship between student test scores and grades in the context of gender, race and ethnicity in order to provide information “to develop appropriate equity responses based on individual differences.” (p. 41)

Robertson states that although the AAUW and other data can’t be used as evidence that gender bias is disadvantaging Canadian girls in every MST-related evaluation, she says “it is curious to note how little attention is paid to the apparent risk to young women posed by a massive shift to computer-scored, single-right-answer standardized evaluation.” (p. 27)

manipulated to political advantage. When this fact is combined with the strong pressure to increase test scores and the questionable practices that ensue, the potential for compounding distortions and providing misleading information is very real.

In a system of school choice (which would include charter schools) the so-called “good” schools are able to attract and select the best students. The “creaming off” of motivated, high-achieving students and their more educationally-involved parents by choice schools is beginning to have a negative impact on local public schools which are left with fewer students (those that are higher-risk and higher-cost) and hence, fewer tax dollars (Dehli, 1998, p. 31).

According to the Canadian Teachers’ Federation (1997),

[s]chool-by-school competition creates high demand for the students who already have the greatest chance of success in our schools, and creates incentives to avoid serving harder-to-teach kids and kids with disabilities who may consume more than their 'share' of resources, yet not increase the school's profile in key marketing areas such as standardized test scores. (p. 17)

It should also be noted that when high-achieving students are attracted by and concentrated in choice programs, the illusion of increased achievement may well be created. Fuller and Elmore (as cited in Dehli, 1998, p. 31) caution that "it is unlikely that choice will do anything other than simply move high achievers around from one school to another, mistaking the effect of concentrating strong and motivated students for an effect of the school or the choice system."
‘Pan-Canadianization’ of curriculum and assessment

The growing Pan-Canadian (i.e. national) agenda for education being led by the Council of Ministers of Education, Canada (CMEC), as outlined in its 1993 “Victoria Declaration”, includes major initiatives in the areas of curriculum, student assessment, and education indicators. With respect to curriculum, in 1997 the CMEC released its Common Framework of Science Learning Outcomes: Pan-Canadian Protocol for Collaboration on School Curriculum, a common curriculum framework which lays out the learning outcomes for K-12 science education (this project involved all provinces except Quebec). Robertson (1998b) observes that the “CMEC sidestepped the thorny constitutional problem of curriculum as a provincial matter by declaring the document to be only a framework for curriculum developers, not an actual curriculum. But no one was fooled.” (p. 86) She notes that the curriculum consultation process was also sidestepped. The time available for feedback on drafts of the long document was short and the nature of the commentary was tightly prescribed – “[t]hose wishing to comment were told that they could question neither the desirability of a national curriculum nor its contentious devotion to outcomes.” (p. 86) Teachers’ organizations, school board associations and other groups traditionally involved in curriculum development were effectively excluded from the process.

Curriculum is also being developed regionally through the Atlantic Provinces Education Foundation in the east and the Western Canadian Protocol for Collaboration in Basic Education in the west. Such curriculum reform initiatives are in keeping with an increased emphasis on standardized outcomes and the core subject areas of math, science and language arts (Dunning, 1997).

With province-wide assessment programs in place in most provinces and territories, a full-blown national testing program cannot be far off. In her analysis of the School Achievement Indicators Program course but determine which students will be accepted to select universities (p. 25). As “scarcity” of post-secondary opportunities flows from underfunding of the post-secondary education system, Canadian universities have increased tuition fees as well as reduced the number of spaces available. These factors combined will reduce access to higher education for women, particularly in areas (such as MST) in which they are already under-represented.

The following findings (excerpted from Robertson, 1992, p. 26) are taken from the American Association of University Women (AAUW) report How Schools Shortchange Girls (1992). This important study synthesizes the available research on the educational experience of girls concluding that the education they receive is inequitable in terms of quality and quantity to that received by boys:

√ Despite girls’ receiving higher grades in high school and college, boys were twice as likely as girls to have received a college scholarship based on the results of standardized high school tests.

√ Girls have consistently outscored boys on the SAT-verbal when test content refers to concepts and ideas rather than things, or to aesthetics, philosophy, relationships or stereotypically female tasks; whereas boys have scored better on items relating to science or sports; since the SAT-verbal was revised in 1978 to increase the number of science-related items, reading subscores favouring boys have increased from three to twelve points.

√ As noted earlier, SAT scores under-predict the success of young women, and over-predict the success of young men in college/university achievement, especially in math.

√ Bias in test items can occur within content, format and the skill or knowledge area being tested; for example, in the area of math, girls tend to score higher on computational tasks, algebra and logic, whereas boys tend to score higher on problem-solving and geometry – tests which give more weight to one cluster of skills over another will produce gender differences.
the Scholastic Aptitude Test (SAT) and similar tests which allegedly predict students’ future academic performance. High school girls aspiring to attend college represent over 52% of those writing these tests. However, their abilities are systematically underestimated by these exams resulting in reduced opportunities for attending post-secondary institutions, for obtaining scholarship funding, and for acceptance to special programs for gifted students. Students who receive low SAT scores have reduced expectations of their abilities which in turn affects their decisions about which colleges they apply to (Weiss, 1991, p. 108). The longer term economic and personal consequences for women include fewer good employment opportunities, lower income levels and reduced self-confidence. Female students from low-income backgrounds face the combined gender and class biases of the exams and are effectively penalized twice (Rosser et al., 1992). In short, SAT scores tell us more about factors such as family income, race and gender than they do about one’s ability to perform well in university (Robertson, 1998b, p. 80).

Robertson (1992) discusses the impact of testing on girls in the context of math, science and technology (MST) education. She notes that “complex indicators of ‘success’ for students and for education are frequently and inappropriately reduced to single-test numbers” (p. 25) and that there is clearly an important political agenda underlying this direction in education reform. For high school students, small differences in test score results not only decide whether one passes a

| 6 | Even the world’s largest private educational test maker, the Educational Testing Service (ETS) which develops and administers the SAT, has admitted the SAT’s bias in terms of over-predicting male college performance while under-predicting the college performance of females. ETS also acknowledges that high school grades are a better predictor of college success than their tests (Thurmond, 1994). In a landmark U.S. federal court decision on the discriminatory nature of the SAT, the judge stated that “...[T]he probability, absent discriminatory causes, that women would consistently score 60 points less on the SAT than men is nearly zero.” Even after accounting for such factors as ethnicity and parental education, a substantial gap in points remained (Rosser et al., 1992, pp. i-ii). |

(SAIP)\(^10\), a national project initiated in 1989 by the CMEC to test Canadian students aged 13-16 in math, reading and writing, and science, Barker (1991) argues that the SAIP is a politically-motivated initiative developed in response to criticisms of educational and calls for increased accountability from business, industry and government. She suggests that the SAIP can be viewed as a catalyst for educational reform in Canada – from more limited changes in terms of institutionalizing national testing, to sweeping reforms such as increased control over education by the federal government and the institutionalization of national education standards with economic productivity. Some of this is echoed in the 1994 report of the National Advisory Board on Science and Technology’s Committee on National Standards in Education. Among its recommendations to the Prime Minister were that:

- “the federal government assist in the development and administration of systemic tests for all core subjects [italics added] building on the work of the School Achievement Indicators Program”, and that
- “the frequency of the administration of the tests in all core subjects be increased to that which maximizes the information necessary to ensure continuous improvement. If maximum advantage is gained with yearly testing, then that should be instituted.” (p. 23)

\(^10\) In the first cycle of the SAIP, assessments were administered in mathematics (1993), reading and writing (1994), and science (1996). The SAIP’s second cycle included assessments in mathematics in 1997, reading and writing in 1998, and science assessment in 1999. A third cycle has since received approval. Funding for the SAIP comes from the provinces, Human Resources Development Canada at the federal level, and private corporations (Investors’ Group Incorporated and Spar Aerospace Limited). The CMEC estimates total costs for administration and implementation of the program from the planning stage in 1989 through to its completion in 1999 to be approximately $11 million, excluding costs incurred for provincial in-kind support [from the CMEC web site at: http://www.cmec.ca/saip/]. Although the SAIP has been in existence for the past decade, there appears to be little in the way of critical analysis of the program and its impact.
The Committee also called for national standards in education and, in a section on the need for employability skills testing, recommended that “the assessment of basic [employability] skills, once they have been defined, be incorporated into the SAIP as it develops and expands its mandate.” (p. 23)

Christensen’s (1998) analysis of the CMEC, from its origins in the Canadian Education Association to its emergence as a self-described “national forum on educational policy”, documents the growing influence of an organization which she notes “has largely escaped critical inquiry since its birth thirty years ago” (p. iii), is not accountable to any government and, despite its stated commitment to consult and cooperate with other educational organizations, is “exclusionary in nature”. In a chapter on the CMEC’s work in curriculum and student assessment (whose early initiatives in these areas in the 1980’s signalled the start of the CMEC’s rising influence), the author describes how federal government pressure in the early part of the decade “precipitated the Council’s hasty implementation of SAIP in 1993.” (p. 83) Federal interest in addressing “public discontent about the educational ‘crisis’ in Canada” was articulated in a 1991 throne speech in which the Prime Minister referred to the need for national education goals, marking as Christensen notes “the return of the federal government to discussions about ‘national’ education policies” (p. 83). (Interestingly, at about the same time the U.S. government released its America 2000 education strategy calling for a system of national standards and assessment.) Mindful not to encroach on provincial jurisdiction, the vehicle chosen by Ottawa to address education issues was the Canadian Forum on Learning, set up through the Steering Group on Prosperity (Prosperity Initiative), to “lead a process to develop a statement of expectations for education and training in Canada” (p. 84). At a meeting of the “Learning Forum” organized through the Prosperity Secretariat and the Conference Board of Canada in 1993, federal MP Bernard Valcourt stated that the legacy of the Prosperity Initiative was to “‘link competitiveness and learning in the minds of Canadians’ ” (Robertson, as cited in Christensen, 1998, p. 86). Christensen suggests that criticism of the CMEC’s record by the federal Steering Group on Prosperity was one of the factors propelling the SAIP agenda forward.

[p]ressure generated by the state test for high-stakes accountability has led school-based educators to pressure the state to be more explicit about content that will be tested. This in turn constrains local school decision making about curriculum. This dialectical process works to increase the state control of local curriculum. (p. 278)

Rather than generating constructivist, performance-based classrooms as was perhaps the intent, KIRIS has instead emphasized “learning that can be measured reliably and validly” – the testing of observable, measurable outcomes effectively determines what is worth knowing. While performance assessment holds promise for improving student learning, this potential is corrupted in a high-stakes environment. The authors conclude that the principles underlying high-stakes accountability and performance assessment are fundamentally incompatible: the former “encourages conformity to externally imposed standards, while the [latter] grows out of emergent interactions between teachers and students in dynamic local classrooms.” (Jones & Whitford, p. 280)

Bias in testing

Standardized tests are typically gender-biased. Rosser et al. (1992) observe that while young women consistently obtain higher grades in both high school and college in the U.S., they receive lower scores on

5 Robertson (1992) observes that “differences between male and female achievement do not necessarily demonstrate bias. It should be remembered that one group may genuinely know more about a particular content area than another group. Any bias is thus the result of the kind of knowledge which is valued, rather than bias in favour of or against a particular group. However, if a test fails to meet the criterion of validity, that is, if two groups know the same amount about a subject but one group consistently scores higher on a test alleged to evaluate that knowledge, or if a test fails to identify the superior level of knowledge of one group over the other, then such a test can be said to be biased.” (p. 26)
There is a very general principle that is true of all important social indicators, including tests, and which relates to a process that is probably always unintended and very often not anticipated. The principle embodies the proposition that the more any quantitative social indicator is used for social decision making, the more likely it will be to distort and corrupt the social processes it is intended to monitor.

Authentic forms of student assessment as an alternative to standardized testing are not immune to these effects. There are some indications that when authentic assessment (which will be discussed in more detail in a subsequent section of this report) is tied to high-stakes accountability, the instructional benefits of such assessment are undermined (Jones & Whitford, 1997; Torrance, 1993). For example, Kentucky’s high-stakes accountability system converts the results of performance-based assessments into a “school score” and relies heavily on the use of external rewards and sanctions. According to Jones and Whitford (1997),

[i]f the score exceeds the state’s expectation for a school, the teachers and administrators can receive substantial bonuses. If the score is not high enough or does not continue to improve over time, the teachers and administrators can be placed on probation, and the school can be taken over by the state. (p. 276)

Particularly frustrating for teachers and administrators was the state’s decision, lacking empirical data, to design the formula for determining school target scores (as a means of measuring the rate of school improvement) in an arbitrary manner.

One of the effects of KIRIS (Kentucky Instructional Results Information System) has been increased state control over curriculum at the expense of local control and the professional judgement of teachers. As Jones and Whitford explain,

The CMEC’s emphasis on assessment-driven accountability derives largely from the current fixation with, and anxiety produced by, global market competition and the resultant need to know “exactly where one stands in comparison to the competition, both nationally and internationally, [as] the first step in becoming armed for the battle.” (Christensen, 1998, p. 76) While the CMEC has been unresponsive to concerns about the SAIP from educational organizations including teachers’ organizations, it has “chosen to privilege certain perspectives” (p. 92), namely those of business interests.

A Canadian Teachers’ Federation (1992) draft position paper on accountability and testing in education raised some important issues with respect to how the various jurisdictions might respond to the initial results of the CMEC’s national testing project and the longer-term, far-reaching impacts of the SAIP. It stated that:

It is axiomatic that some provinces will rank lower than others (in fact, the results are almost predictable, on the basis of knowledge already available). What will these provinces do? Will they agree to adopt the curriculum and organizational strategies of the more successful provinces (thus moving towards a national core curriculum)? Will they attempt to modify existing curriculum so that it is more closely aligned with the tests? Will school boards seek to replicate the tests, thus making identification of individual schools and teachers more of a possibility? Will teachers be encouraged to teach to the next round of tests? Will the end result of the testing program be a pedestrian, overstandardized educational system in which there is no room for creativity or autonomy on the part of students or teachers and where the most neglected goals are those indeterminate aims of education which all agree are particularly important? (p. 44)
As the SAIP nears the end of its second full cycle and concerns begin to surface about aspects of the program (such as the process by which expectations for student performance are set), there are indications that it is influencing, as Robertson (1999) has observed, a “remarkable degree of curriculum homogenization within a country that has considered local control of curriculum to be a virtue rather than a liability.” (p. 715) The creation by regional consortia in the Western and Atlantic provinces of outcomes-based curricula aligned with the SAIP reflects this trend.

The CTF paper goes on to emphasize the importance of having broad and diverse educational goals drive appropriate educational means:

Meaghan and Casas (1995b) describe the SAIP as a classic example of “placing the cart before the horse” – the development of a national testing program in the absence of public consensus on just what our educational goals should be and how these goals should be reflected in curriculum. This approach will result in “a handful of bureaucrats and consultants” deciding what it is they think students should know (p. 88).

As noted earlier, curriculum is becoming more test-driven and less responsive to student and community needs. Education means are dictating narrowly defined education goals. Barlow and Robertson note that standardized testing places more emphasis on score manipulations and student assignments or exclusions than on school improvement and the development of more effective teaching practices. (pp. 15-16)

In the context of U.S. proposals to introduce national testing, Madaus and Kellaghan (1993) analyze the use of testing as an instrument to drive educational policy noting that “the link between test performance and important rewards or sanctions is the key ingredient that gives tests their policy punch” – hence the term “high-stakes testing”. In an article entitled “Testing as a Mechanism of Public Policy: A Brief History and Description”, they outline six basic principles to describe the processes “through which tests used in the policy sphere affect individuals and institutions”. Their observations are insightful. (Excerpted from Madaus & Kellaghan, 1993, pp. 8-9):

1. The power of tests and examinations to affect individuals, institutions, curriculum, or instruction is a perceptual phenomenon – in other words, if students, teachers, ... very little whether this is really true or false – the effect is produced by what individuals perceive to be the case.
2. If important decisions are presumed to be related to test results, then teachers will teach to the test.
3. In every setting where a high-stakes test operates, a tradition of past exam questions develops, and this tradition eventually comes to define de facto the curriculum.
4. When test results are the sole or even partial arbiter of future educational or life choices for students, society tends to treat test results as the major goal of schooling rather than as a useful but fallible indicator of achievement.
5. A high-stakes test transfers control over the curriculum to the agency that sets or controls the exam.
Standardized Testing

20 Canadian Teachers’ Federation

Other questions raised by the SPIP include: What will be the impact on teacher collegiality? Will the program result in increased teacher requests to transfer to schools with higher student achievement? What is the potential fallout from identifying schools and classrooms that don’t meet the target objectives? (“Alberta Proposes Bonuses for Student Achievement Gains”, 1999) Arguing that such incentive programs are inconsistent with what is known about effective teaching and student evaluation practices, teacher motivation and school improvement generally, the ATA and its provincial education partners have proposed an alternative to the SPIP designed to improve teaching and learning based on proven research and practice. The Minister of Education has agreed to put the SPIP on hold pending discussion of the alternative program (“Improvement, Not Incentives”, 1999).

The so-called hard approach as described in the Education Week survey is clearly not the path to genuine accountability in education as advocates suggest. Darling-Hammond (1994) explains the negative impact on equity of such an approach:

Needless to say, this kind of policy that rewards or punishes schools for aggregate test scores creates a distorted view of accountability, in which beating the numbers by playing shell games with student placements overwhelms efforts to serve students’ educational needs well .... Applying sanctions to schools with lower test score performance penalizes already disadvantaged students twice over: having given them inadequate schools to begin with, society will now punish them again for failing to perform as well as other students attending schools with greater resources and more capable teachers. This kind of reward system confuses the quality of education offered by schools with the needs of the students they enroll; it works against equity and integration, and against any possibilities for fair and open school choice, by discouraging good schools from opening their doors to educationally needy students. Such a reward structure is a powerful force for curriculum reform and for shaping the goals of education. What has been and will be tested becomes a priority, not just in the public mind or ministry policy, but in time and attention in the classroom. Other instructional priorities recede, thus shifting curriculum balance effortlessly and without debate. (pp. 118-119)

This is being borne out south of the border. Over the past 150 years, testing requirements in the U.S. have had a number of unintended yet dramatic effects in terms of education policy. These include the reshaping of curriculum as well as attitudes about the goals of education (Madaus, 1994, p. 79). The impact of institutions such as the Educational Testing Service (ETS), the world’s largest private testing firm, on education are significant and cannot be underestimated. Lemann (1995) wrote in the Atlantic Monthly that the ETS and its tests are a national obsession, deeply worked into the fabric of middle-class life. They have generated a large, independent test-prep industry and have strongly influenced elementary and secondary education, one of whose central goals is preparation for ETS tests. (p. 42)

The bigger picture of educational reform

The trends in assessment cited in the introduction to this report must be situated within the context of broader trends in educational change, among them: reduced funding, reforms increasingly being justified in economic terms (i.e. education as the means to prepare individuals for the labour market and nations for global competition), and a move toward market-based solutions which view education as a commodity and students and their parents as consumers of that commodity (Levin, 1998). Such trends, motivated more by fiscal, political and ideological considerations than by a sincere desire to improve access to high quality
education, are hardly surprising in a world in which the prevailing ideology is one of globalization.\footnote{For a summary of some of the main features of globalization, see “On Globalization” by Larry Kuehn, British Columbia Teachers’ Federation [available online at: http://www.ctf-fce.ca/e/what/other/onglobe.htm].}

The influence of international bodies such as the Organisation for Economic Co-operation and Development (OECD) and the Asia-Pacific Economic Community (APEC) on education policy is being increasingly felt. In his analysis of the APEC position paper on education prepared by the South Korean government, Kuehn (1997b) notes that the “globalization agenda for education” relies on the premise that “globalization is inevitable and education [along with every other sector of society] must comply with its requirements”. A major goal of education in the global economy is the preparation of workers for business, hence “business should have a central role in determining the content of schooling”. The content of education becomes narrowed to work ethic, attitudes, and skills “appropriate for a working life”. In this context standardized testing becomes a convenient means of regulating the production of the necessary “human capital”. To illustrate, Kuehn (1997a) cites the case of Mexico where a standardized examination, the Examen Unico, has been created to channel more students into vocational/technical programs that produce the skills required for the global economy and away from university programs “primarily aimed at the specific social and cultural development of Mexico.” (p. 86) Opposition to the Examen Unico was strong to say the least. In August 1996, anti-test demonstrators blocked busy streets in Mexico City and forced a one-day closure of the Mexican stock exchange.

Robertson’s (1998a) analysis of high-stakes testing in the context of globalization illustrates well how testing is becoming an important lever for a number of major education reforms. Globalization is moving us inexorably towards what she describes as McWorld – a highly inequitable one-third/two-thirds world in which “[t]he one-

\checkmark 14 states provide monetary rewards for individual schools based on performance.

\checkmark 2 states have attempted to link the evaluation of individual teachers to student performance.

The report contrasts this “hard approach to accountability” with what it describes as a “soft approach” – the latter involves providing schools and students with sufficient resources, support and encouragement as well as enhancing professional accountability. The concept of professional accountability recognizes the critical link between teacher quality and student achievement and hence, the importance of investing in teacher professional development and growth.

In March the Alberta government announced its own version of outcome-based accountability for K-12 education. According to an Alberta Education news release (March 11, 1999), it intends to spend $66 million for a School Performance Incentive Program (SPIP) described as an “innovative initiative designed to improve student learning” by awarding additional funding to school boards “based on how well they achieve improvement targets.” Targets will be set by the province in conjunction with local boards. Allocation of funding awards (ranging from 2-4 per cent of a board’s salary budget) is conditional on improvements in learning as measured by provincial achievement test and diploma exam results, high school graduation rates, and “locally determined performance indicators” such as student attendance and results from local achievement tests. The program which is voluntary (and apparently unprecedented in Canada) will be implemented in the 2000/2001 school year. As noted above, Americans are more familiar with such initiatives – 14 states including North Carolina, Maryland and Texas have instituted some form of financial incentive program (“Alberta Proposes Bonuses for Student Achievement Gains”, 1999).

The Alberta Teachers’ Association (ATA) is concerned that a program which essentially offers cash bonuses in return for improved test scores, particularly at a time of budget cutbacks, “will pit schools and boards against each other and unfairly favor well-to-do districts.” (McMahan,
scores without necessarily upgrading skill levels.
(Meaghan & Casas, 1995c, p. 49)

As Elliot Eisner remarked in a presentation at the B.C. Teachers’ Federation AGM in March 1998, “[i]t is perfectly possible for a school’s faculty to raise test scores and at the same time to diminish the quality of education.”

High-stakes, measurement-driven accountability

A recent Education Week survey of the accountability policies of 50 states (with the less than subtle title “Quality Counts ’99: Rewarding Results, Punishing Failure”) reveals the strong push for outcome-based accountability to leverage educational reform in the U.S. According to the report, among the components (described as “essential steps”) necessary for putting a comprehensive system of accountability in place are tests, standards, report cards, rating systems, and rewards and sanctions. Using these components as the basis for its analysis, the survey found that:

- 48 states have state-wide testing programs (13 of which use only multiple choice tests).
- 36 states publish annual “report cards” on the performance of individual schools.
- 19 states evaluate and issue public ratings of school performance including “low-performing” schools.
- 16 states have the power to close, take over or “reconstitute” a failing school (reconstitution is described in the report as “replacing all or most of a school’s staff and reopening it under new management.”) Not surprisingly, schools attended by poor and minority students are more likely to have sanctions applied to them.

Not surprisingly, schools attended by poor and minority students are more likely to have sanctions applied to them.

means to leverage support for any number of McSchool reforms that would otherwise be rejected as unconstitutional, ill-advised, or contrary to everything else we know about fostering good schools and good students. High-stakes testing is so closely and yet opaque related to the other reforms of McSchools – the narrowing of curriculum, the depprofessionalization of teachers, privatization and compulsory enthusiasm for technology – that it has been strategically brilliant to focus on this strategy.

Mechanisms such as high-stakes standardized testing and performance indicators not only embody dominant cultural values associated with competition, efficiency, performance, and outcomes (the more concrete and measurable, the better). They are also an important link between the parallel and apparently contradictory tendencies in education reform of decentralization (weak state) and centralization (strong state). As decentralization takes shape in reforms such as site-based management in which decision-making responsibility is devolved to the school level, and school choice in which power is devolved to individual “consumers” of education, governments are centralizing power over funding, curriculum, standards and assessment. As noted earlier, assessment of centrally-set outcomes and the subsequent publication of results enables parents to make comparisons and select among schools – thus facilitating a measure of accountability to the market for the school system. In his review of Whitty, Power and Halpin’s book Devolution and Choice in Education: The School, the State and the Market, Apple (1998) discusses the authors’ concept of the “evaluative state”:
allocate federal funds to schools in “America 2000”. Such a policy would provide an incentive to schools to close their doors to students who might “bring down” the average scores, such as children with learning and other disabilities, those whose first language is not English, “or children from educationally disadvantaged environments.” (Darling-Hammond, 1991, p. 223) It could also result in experienced teaching staff avoiding schools which serve disadvantaged kids, thereby closing off further educational opportunities for these children.

Citing a study of a large urban American school district that used average school test scores to make decisions about rewards and sanctions for individual schools, Darling-Hammond (1991) ominously describes the impact as “the widespread engineering of student populations”. The authors of the study state that

“[s]tudent selection provides the greatest leverage in the short-term accountability game .... The easiest way to improve one’s chances of winning is 1) to add some highly likely students and 2) to drop some unlikely students, while simply hanging on to those in the middle. School admissions is a central thread in the accountability fabric.” (as quoted in Darling-Hammond, 1991, p. 223)

As noted above, the pressure to raise average test scores can result in the rejection by the school system of more at-risk students. This includes placing students with low scores in special education programs to avoid having their scores reflected in school reports; refusing to admit students to “open enrolment schools” on the basis of low scores; and going so far as to encourage students with low test scores to drop out of school (Darling-Hammond, 1991, p. 223). In addition, the pressure to raise and maintain average test scores is such that,

as teachers, principals, and superintendents come to see test scores as having a critical impact on their own careers and on their students’ lives, testing becomes the curriculum. Dissemination of test scores places greater pressure on teachers and administrators, who respond to the higher stakes by devising ingenious ways to raise
Other manifestations of test misuse

Misuse of the results of province-wide assessment programs in Canada includes “inadequate representation of test results by politicians and the media and inadequate explanation of test results to the public (for example, no background information or comparison of samples with differing characteristics).” (Gilliss, 1994, p. 6) A survey of provincial ministries and departments of education as well as a sample of local school boards across the country also reported that a major drawback of province-wide testing programs is the misinterpretation of results by the public and the media (Traub, 1994, pp. 1, 3).

When 90% of the students in an Ontario Grade 9 reading and writing test fell into the category of having “adequate” reading skills, results that received broad coverage in the media, the North York School Board vice-chair “announced that these results were not good enough for his Board” – this despite the fact that he had not been involved at any level in the preparation of the test, had no indication of how reliable the scores were, or whether the student population in North York [in Metropolitan Toronto] has special characteristics such as a higher-than-average proportion of ESL [English as a second language] students. (Meaghan & Casas, 1995b, p. 83)

Standardized test results are used in some U.S. jurisdictions to make decisions about teachers with respect to merit pay awards as well as certification and recertification. Test-based decisions impacting on schools and school systems include such things as allocation of funds, and school system certification and decertification (Medina & Neill, p. 24). To use average school test scores as the basis for allocating rewards and sanctions to teachers and schools, a practice several U.S. states and school districts have already put in place, is to endorse a market-based approach to stimulating improvement in the school system. Former U.S. President Bush proposed using a national test to

Some General Observations

Shifting from the global to the classroom context, it is important to emphasize that teachers recognize the importance of evaluating student learning and of finding the best means of classroom-based assessment to accomplish this. They regularly evaluate learner progress using various methods which they design and administer – these include term exams, quizzes, essays, projects, portfolios and classroom observation to name a few. Robertson (1998b) describes the diverse formal and informal evaluations carried out by teachers as being “remarkably good predictors of their students’ future success, both in other courses and in life” (p. 65). According to Lorna Earl (1995), Professor of Theory and Policy Studies in Education at OISE/UT, “classroom teachers and their daily classroom assessments may form the basis for the most important accountability relationships.” (p. 6) Because of their regular contact and close relationship with students, teachers are in the best position to conduct ongoing assessments of students, provide regular feedback on their performance to parents and to the students themselves, and recommend ways in which students can learn more effectively. Quoting from R. Stiggins’ book Student-Centered Classroom Assessment, Earl (1995) makes the point that the rationale underlying the importance of classroom assessment in the improvement of learning is also the reason that large-scale assessments cannot achieve this:

“...all of the centralized testing programs operate on the assumption that the decisions that exert the greatest influence on school quality are made at school management and policy levels; that is, somewhere outside the classroom ....This view of the origin of school improvement is naive and obviously counterproductive. The halls of political power are not where school
improvement originates. There are three groups of decision-makers who determine the quality and therefore the impact of American schools. In order of importance, they are students themselves, their teachers and their parents. All three are informed in their decision-making by the same source of information on student achievement and that source is not national, state or local standardized testing. They all rely on teachers’ day-to-day classroom assessments of student achievement. For this reason, the path to school improvement lies not through more standardized testing, but through the development and use of the best quality classroom assessment we can generate!“ (as quoted in Earl, 1995, p. 7)

In addition to acknowledging the centrality of teachers to the assessment and evaluation process, we must recognize the value of the multiple goals of education—goals that go beyond simply preparing students to enter the labour force as important as that may be. There is also a need to evaluate evidence of all of these goals, not just cognitive abilities, with a view to improving the quality of education. Although effectively evaluating such goals as preparing students to assume the responsibilities of good citizenship is an important and complex task, “that complexity is precisely what makes those goals so important to teach.” (Barlow & Robertson, p. 119)

Reliance on a single form of evaluation, particularly one as flawed as standardized tests, for making important decisions about students, teachers and schools is irresponsible. Equally important is the danger of relying on one form of evaluation for several different purposes (whether student progress, teacher performance, or program and system effectiveness). The correct evaluation tool needs to be used for the task at hand (Earl, 1995, p. 6).

We need to revisit the concept of accountability and what it actually means. Contrary to current trends in assessment policy, testing by itself does not constitute accountability. Darling-Hammond (1999) reminds us that accountability is according to Shepard and Smith, “‘[c]ontrary to popular beliefs, repeating a grade does not help students gain ground academically and has a negative impact on social adjustment and self-esteem.’” (as quoted in Darling-Hammond, 1991, p. 222) Grade retention has also resulted in lower achievement. Given the lower levels of self-esteem generally observed among girls as compared to boys, the implications of this policy for further eroding self-esteem among girls and young women requires serious consideration.

**Graduation**

Test scores have been used to determine a student’s eligibility for high school graduation. According to Darling-Hammond (1994),

[t]he rationale for this practice is that students should show that they have mastered the “minimum skills” needed for employment or future education in order to graduate. The assumption is that tests can adequately capture whatever those skills are. While this appears plausible in theory, it is unlikely in reality, given the disjunction between multiple-choice tests of decontextualized bits of information and the demands of real jobs and adult tasks …. In fact, research indicates that neither employability nor earnings are significantly affected by students’ scores on basic skills tests, while chances of employment and welfare dependency are tightly linked to graduation from high school [italics added] …. Thus, the use of tests as a sole determinant of graduation imposes heavy personal and societal costs, without obvious social benefits. (p. 14)

Nineteen U.S. states currently require students to pass state tests to graduate from high school (Education Week on the Web, 1999). In Canada the majority of provinces now require graduating students to take high school exit exams with the stakes ranging from 30 to 50 per cent of the final course grade (Dunning, 1997). Once taken as evidence of years of cumulative development, the high school diploma may eventually come to be replaced by a one-shot, high-stakes test.
Tests of individuals are used to analyze policy, program, school and teacher success, and they are being inappropriately used as “educational gatekeepers” to make important decisions about students, teachers, schools, and the school system as a whole (Medina & Neill, p. 24). Standardized test scores are becoming the mechanism that facilitates a number of questionable education practices that contribute to education inequity.

**Tracking**

In the U.S., test scores are used routinely to make important decisions about student placement. The “tracking” that results streams children from disadvantaged socio-economic backgrounds into average and low-track classes and programs. Most progressive educators conclude that the practice of tracking fosters a negative self-image and creates reduced expectations for a student’s future. It also denies those in the lower tracks the benefits of working with higher achievers. In Canada there is a danger that standardized testing will be used as a mechanism to renew interest in tracking thereby moving us away from a more inclusive model of education (Meaghan & Casas, 1995a). Tracking accounts in large part for the disparity in achievement observed between American White students and those from minority groups, as well as between students from higher- and lower-income groups (Darling-Hammond, 1991, p. 222). Perhaps because curriculum in “low-track” classes tends to be limited and rote-oriented, students in these classes don’t do as well as their peers of comparable aptitude in academic or untracked programs. Oakes (1985) concludes that tracking hurts low-income and minority group children, and that it not only reflects but perpetuates class and racial inequalities in the larger society. Using a method of testing already biased against certain groups of students to subsequently assign them to lower track programs only adds insult to injury.

**Grade promotion and retention**

Standardized test results have also been used to determine grade promotion and retention. The practice of grade retention can contribute to an increased drop-out rate (Medina & Neill, p. 29) and a means by which individuals or organizations take responsibility for their actions so that those who depend on them can be assured that there are some safeguards in place to encourage good practices and to prevent bad practices or abuses, to have some course of redress for problems that arise, and to have some assurance of equitable and fair treatment.

With this in mind, she notes that in a genuine, comprehensive system of educational accountability, the various players, from the federal and provincial governments down to the individual school and teacher, have important roles and responsibilities they must fulfill and for which they must be held to account. For example, the state or province “should be responsible for resourcing schools so that they can provide adequate education, for ensuring that there is equity in the distribution of those resources, [and] for assuring that there is a means by which qualified, well-prepared staff will be in all schools in those settings.” Local school districts are responsible “[f]or practices that are supportive of good teaching and learning in schools, [and] for processes that support [these practices].” Teachers are responsible for “making good educational decisions on behalf of children”. Darling-Hammond (1999) describes the latter as professional accountability which means investing in teachers’ skills and knowledge. This is a central component of a genuine accountability system, one which requires providing teachers with ongoing access to the knowledge and support necessary to make sound decisions in order to improve student learning. Unless and until we hold all of the actors accountable for their diverse responsibilities, we can only speak of accountability in a vacuum.

In sum, there is little evidence to support standardized testing from a learning standpoint. Moreover, standardized testing has been shown to have a negative impact on learning, and is particularly harmful to at-risk students – it perpetuates and worsens educational inequities. Assessment, including authentic assessment, must be viewed first and foremost as part of a diverse strategy to improve the quality of education and address equity issues – other systemic reforms must be
implemented in tandem that address broad social and educational conditions. If a primary goal of our education system is to offer all students equality of educational opportunity, then standardized testing will only move us further away from this goal.

Testing and Inequity in Education

Not only is there little evidence to support standardized testing from a pedagogical standpoint, it is clear that standardized testing can have a detrimental impact on the most vulnerable students. Far from being a neutral practice, it perpetuates and intensifies educational inequities in two ways: through the misuse of test scores; and because test bias works against the interests of students from low-income groups, racial and ethnic minorities, girls and young women, and students with disabilities. Darling-Hammond (1994) reminds us that “[t]he role of testing in reinforcing and extending social inequalities in educational opportunities has by now been extensively researched ... and widely acknowledged.” (p. 10)

Inappropriate use of tests

Tests are often used inappropriately. The history of testing reveals many instances of tests developed for a specific purpose being used for other unintended purposes. The Scholastic Aptitude Test or SAT is a prime example. Originally designed to predict a student’s potential for success in their first year of college in the U.S., the SAT has been used as a criterion for determining such things as athletic eligibility, student loan eligibility and awarding scholarships (Madaus, 1991, p. 227). Despite its limitations (including its “limited predictive value”), Schwartz (1999) observes that the SAT “has arguably become the single most important test for American high-school students – an academic and psychic rite of passage that strongly influences future educational options, prompts fierce anxiety and serves as an almost mystical barometer of self-worth.” (p. 30)
Other problems

Additional problems associated with over-reliance on standardized tests include:

- Neglecting essential educational services by allocating scarce financial resources to expensive (and questionable) standardized testing programs at a time of severe government cutbacks to education (Meaghan & Casas, 1995b, p. 87); the testing of all Ontario Grade 3 and Grade 6 students in 1999 is expected to cost about $12 million;

- Disregarding variations in curriculum and curriculum organization for students in particular grades among different countries when comparing international test results (Ireland, 1995);

- The inability of information provided by standardized tests to assist educators in identifying and improving ineffective school programs (Casas & Meaghan, 1995, p. 19);

- Shifting responsibility over curriculum from teachers and school authorities to the testing industry and government officials, thereby lessening rather than enhancing accountability (Meaghan & Casas, 1995d, p. 47).

In the United States, the production of standardized tests is mainly in the hands of commercial publishers and non-school agencies (Darling-Hammond, 1991, p. 220). The situation in Canada is similar – standardized tests are developed and marketed by the major textbook publishers with profits going to the U.S.-based parent companies (Barlow & Robertson, 1994, p. 120). The U.S. private testing industry is a billion dollar a year business which, despite its growing influence on the education system, is virtually unregulated and unaccountable. Under such conditions, increased involvement by teachers and parents in the educational decision-making process is ruled out (Medina & Neill, 1990).

Appendix A — Summary of Problems with Standardized Testing

A considerable body of research has identified numerous problems associated with the nature and use of standardized achievement tests. The following is a partial list (excerpted from Meaghan & Casas, “On the Testing of Standards and Standardized Achievement Testing: Panacea, Placebo, or Pandora’s Box?” Interchange, 26(1), 1995, pp. 34-35):

- To permit machine scoring, standardized tests are limited to multiple-choice items which, in turn, severely limits what can be tested. For example, you can test language skills but not writing ability or how students use language.

- To ensure comparability of test scores over time, few substantive changes are made over the years, which hardly encourages updating of curricula.

- Although advocated as a means to carry out international comparisons of student learning and performance, these tests are not sensitive to differences in curricula for particular ages and grades in different countries (nor can they be made so).

- Because each question is assigned the same score value, the student who answers more of the difficult (and/or important from the program standpoint) questions is judged the equal of the student who gets only the easy ones.
√ In order to achieve a proper dispersion in the scores [to facilitate making comparisons], questions that are almost always answered correctly or incorrectly tend to be avoided. Aspects of performance that should be tested may therefore not be represented at all because they make no contribution to score dispersion. This leads to overtesting of minutiae and undertesting of important conventions.

√ Standardized tests stress the product of learning, not the process. They measure a student’s ability to recall facts, define words, and perform routine calculations, not higher learning processes such as analyzing, synthesizing, forming hypotheses, and exploring alternative ways of solving problems.

√ More crucially, standardized tests have nothing to do with performance standards but only measure relative standards.

√ Standardized tests do not serve as a useful diagnostic tool because they do not show students where they went wrong; corrected tests are never returned, excluding the process of feedback and debate that is at the core of learning and education.

√ Standardized tests systematically rule out students’ independent creativity and criticism by eliminating the need to construct answers for themselves.

√ Rather than increasing accountability, standardized tests merely shift it from teachers and school authorities to anonymous government officials or corporate bureaucrats who cannot be confronted or held accountable if tests are poorly constructed, administered, or marked.

√ Many studies have shown that standardized tests are biased against socioeconomic, racial, and ethnic minority groups.

In 1997 Ontario’s testing agency, the Education Quality and Accountability Office (EQAO), conducted a provincial assessment of all Ontario Grade 3 students in the areas of language arts and mathematics (this annual province-wide test has since been expanded to include all Grade 6 students). Some of the limitations of this testing exercise which took ten full days to administer at a cost of over $7 million are documented in Robertson (1998b) and Ireland (1997a). While the administration time for the Grade 3 test was reduced the following year, it still required three hours per day over five days. Such exercises raise ethical questions regarding the use of valuable teaching and learning time being taken up with testing activities of limited utility.

Teachers are induced to teach to the tests rather than for learning with the result that curriculum is becoming increasingly test-driven (Meaghan & Casas, 1995d, p. 47).

This phenomenon is well-documented in the research literature. Many teachers have reported that they align their curriculum with what will be tested, are concerned that standardized testing narrows curriculum, and feel pressure to improve test scores (Meaghan & Casas, 1995c, pp. 38-39). For example, a national survey of teachers conducted by the Boston College Center for the Study of Testing, Evaluation, and Educational Policy reported that teachers with a large percentage of minority students in their classrooms (over 60%) were more likely to teach to the standardized tests, emphasizing basic academic skills and time spent in test-taking activities, than were their counterparts who taught mainly White students (Madaus, 1994, pp. 83-84). Wideen et al. (1997) examined the impact of Grade 12 school-leaving examinations on the teaching of science in several British Columbia school districts. They concluded that the high-stakes tests reduced the variety of instructional approaches used, particularly at the Grade 12 level, because they “discouraged teachers from using strategies which promoted enquiry and active student learning”, and that “this impoverishment affected the language of classroom discourse.” (p. 428)
√ using commercial materials specifically designed to improve test performance;

√ dismissing low-achieving students on test days; and

√ interfering with responses by, for example, giving hints or answers to students or altering response sheets.

Other factors that can influence test scores include students’ anxiety and motivation levels (Haladyna, Nolen & Haas, 1991). It has been demonstrated that anxious test-takers don’t perform as well as test-takers who are more relaxed, and they do particularly poorly on more difficult test items. A meta-analysis of over 500 studies by Hembree (as cited in FairTest, 1995) reveals “that students with high test anxiety hold themselves in lower esteem than do those who are less test anxious ... [f]emales have higher test anxiety than males, Blacks in elementary school have higher test anxiety than Whites, and Hispanics have higher test anxiety than Whites at all ages.” (p. 9)

Meaghan and Casas (1995d) note that because standardized test scores “are very sensitive to motivational factors ... [they] are not good measures of knowledge” (p. 47). Motivational factors were the focus of a study of student performance on the mathematics section of the Iowa Test of Basic Skills. In this study over 400 elementary students were randomly placed in two groups – those in the experimental group, which received special instructions from their teachers on the importance of doing their best on the test, had significantly higher scores than a control group which did not receive the instructions (Meaghan & Casas, 1995c, pp. 42-43).

Test preparation and administration take up valuable classroom time that could be used for teaching.

In some instances, curriculum is discarded to make way for test preparations that sometimes begin months before the actual test (Medina & Neill, p. 26). A number of Ontario teachers preparing students to take the Ontario Grade 9 Reading and Writing Test noted privately that much of the latter part of the 1993/94 school year was

Appendix B —

Principles for Fair Student Assessment Practices for Education in Canada

(Joint Advisory Committee)

In 1993 a Canadian working group convened by the Joint Advisory Committee of the University of Alberta’s Centre for Research in Applied Measurement and Evaluation produced a document called Principles for Fair Student Assessment Practices for Education in Canada. The principles and their related guidelines address both classroom assessments and large-scale standardized assessments developed outside the classroom (by commercial test publishers, ministries of education and school boards), and represent a broad consensus of the education community (participating organizations included the Canadian Education Association, Canadian School Boards Association and Canadian Teachers’ Federation among others). Intended for both developers and users of assessments (the former being those “who construct assessment methods and people who set policies for particular assessment programs”; the latter are those “who select and administer assessment methods, commission assessment development services, or make decisions on the basis of assessment results and findings” (p. 3)), the principles include the following:

12 In the U.S., a project by the National Forum on Assessment, a coalition of education and civil rights organizations, led to the development in 1995 of Principles and Indicators for Student Assessment Systems. The first principle asserts that the primary purpose of assessment is to improve student learning.
Standardized tests actually squeeze out what is most commonly taught and learned, not because of any nefarious plot, but because of the statistical requirements of data analysis. The type of question that best discriminates among learners is the one that is answered correctly only 50 per cent of the time. Questions that are consistently answered correctly by more than 60 per cent of students are routinely dropped. While these questions may be too easy, they may also represent a core area of content that has been carefully taught to almost every child. To function as they are intended, standardized tests require the evaluation of peripheral concepts, not the common core. Forced to teach to the test, teachers must focus on less important content – the kind of factoids that suit multiple-guess exams – rather than on core ideas or complex understandings. (p. 71)

Test performance is shaped by individual characteristics not related to content knowledge.

Pressure to increase scores on standardized tests and “a lack of understanding of the complexities of achievement and its causes” have resulted in what is known as test score pollution, described as “factors affecting the truthfulness of a test score interpretation.” (Halady, Nolen & Haas, 1991, p. 4) Test score pollution encompasses a number of test preparation and administration practices designed to increase test scores but which are unrelated to the material covered. Halady, Nolen and Haas (1991) describe some of these practices and whether they are ethical. They include:

- teaching test-taking skills;
- developing a curriculum and teaching objectives to match the test;
- presenting items similar to those appearing on the test;
- presenting the actual test items before the test;

Assessment methods should reflect the purpose and context of the assessment.

Students should be given sufficient opportunity to demonstrate the knowledge, skills, attitudes, or behaviours being assessed.

Procedures for judging or scoring student performance should be appropriate for the assessment method used, and should be consistently applied and monitored.

Procedures for summarizing and interpreting assessment results should yield accurate and informative representations of a student's performance in relation to the goals and objectives of instruction for the reporting period.

Assessment reports should be clear, accurate, and of practical value to the intended audience.
solving problems”. In addition, they “exclude the process of feedback and debate which is at the core of learning .... [and] they rule out independent creativity and criticism by eliminating the need to construct answers” (p. 47). The following analysis by McMurtry (1992) of a question from an Alberta Ministry of Education achievement test for Grade 3 social studies underscores this problem:

Question 7, one of 9 Model Questions, asks “Which picture BEST shows people keeping a custom?” Underneath are four block pictures – a pick-up truck, three Native people in traditional costume, kids in uniform playing hockey, and a group of teenagers dancing. The last three of these pictures all qualify as representing “people keeping a custom.” Yet only one picture qualifies as “correct.” This means that a grade 3 child could be marked wrong for choosing correctly in two of the three cases, that any reason or argument over what is in fact true is ruled out even if it is fully justified, that a genuinely critical intelligence will be thrown off by a test that is demonstrably arbitrary in its selection of what is right and wrong (p. 95).

In U.S. schools, the increasing emphasis on the use of “test-oriented accountability measures” parallels a decline in teaching methods such as essay writing, research projects, lab work, and student-centred discussions designed to foster higher-order learning skills (Darling-Hammond, 1991, p. 222).

Because standardized tests are designed to sort individuals into groups, test questions are chosen on the basis of how well they contribute to spreading out the scores, not on their centrality to the curriculum or their predictive validity.

In norm-referenced tests, score dispersion is required as a basis for making comparisons. Questions that don’t contribute to spreading out the scores are therefore avoided. Robertson (1998b) explains this rather technical concept in terms understandable to the lay person:

---

**Appendix C — CTF Principles on Curriculum, Assessment and Evaluation**

The Canadian Teachers’ Federation has developed its own set of principles on assessment and evaluation. In early 1998 the CTF held a workshop on curriculum and evaluation (referred to in the introduction) for the leadership of teacher organizations across the country. An important outcome of the workshop was the creation of draft statements of principle on assessment and evaluation as well as on curriculum. These statements of principle are intended to guide the work of provincial and territorial teacher organizations in a number of areas including the development, implementation and scope of curriculum, and high-stakes testing and its use. They have since been adopted as CTF policy and are reproduced here:

**Principles on Curriculum**

- Curriculum must be broad in scope and holistic in nature in order to meet the needs of children, community and the larger society.
- Curriculum must reflect realistic expectations based on knowledge of child development.
- Curriculum must promote, support, enhance and model equity and respect for diversity.
- Curriculum must be designed to prepare students to become caring, responsible and participatory members of society.
Standardized Testing

√ Curriculum is ongoing and dynamic.

√ Curriculum development, implementation and evaluation are multi-phased processes that require extensive time and resources.

√ Teachers must have the primary role in, and control of, curriculum development.

√ Professional dialogue and collaborative activity are essential for professional growth and development.

√ Teachers require time and resources to engage in ongoing professional development, and in order to effectively implement curriculum change.

√ Teachers must have the autonomy to choose the methods, timing and curriculum adjustments that are necessary to adapt the curriculum in order to meet students’ needs.

√ The goals society sets for students and schools must be challenging but attainable, and progress towards these goals must be measured fairly and in context.

√ The conditions under which teaching and learning take place have a direct effect on what teachers and students can achieve together.

√ The process for curriculum development must include opportunity for input from the broadest possible community representation.

Principles on Assessment and Evaluation

√ Assessment is the gathering of reliable information pertaining to students’ knowledge and understanding including social, economic and educational factors and resources that influence student and system performance.

√ Standardized tests designed for large numbers of students are of necessity very general in nature.

This leads to a mismatch between what is taught (given curriculum diversity) and what is tested (Casas & Meaghan, 1995, pp. 18-19). In his analysis of the Grade 8 results from the Third International Mathematics and Science Study (TIMSS) released in November 1996, Ireland (1997b) points out the substantial differences in test-curriculum match which exist among the provinces. For example, the TIMSS report noted that while 98% of the test items matched the British Columbia science curriculum, the corresponding figures for Ontario, Alberta, and Newfoundland were 53%, 54%, and 54% respectively. There was however little explanatory information on test-curriculum match beyond the tables and their respective figures. Ireland suggests that the TIMSS study could have been improved by addressing the following questions (excerpted from Ireland, 1997b, p. 14): How well did students do on questions derived from the curriculum they should have been taught? How representative of a jurisdiction’s curriculum was the test and how significant is the part of the curriculum not tested? What outcomes were on the test which are not in a jurisdiction’s curriculum, and how significant are those outcomes?

√ Standardized tests typically measure lower-order recall of facts and skills, and penalize higher-order thinking.

Meaghan and Casas (1995d) note that standardized tests “measure a student’s ability to recall facts, define words, perform routine operations, rather than higher learning skills such as analyzing, synthesizing, forming hypotheses and exploring alternative ways of
Tests may be standardized, but students are not.

A student’s capacity to learn is affected by a host of factors including: the impact of poverty; parents’ educational level; mental, physical and emotional health; the effects of racial and other forms of discrimination; and language of origin. For this reason, the absence of information on the characteristics of students in the schools taking part in testing programs leads to misleading and inappropriate use of the data derived (Canadian Teachers’ Federation, 1995b). In a 1994 survey of province-wide assessment programs carried out by the Canadian Teachers’ Federation, a few jurisdictions reported collecting some background data on students in conjunction with the tests (principally demographic in nature). One province collected information on such factors as parental involvement in education and amount of homework. However, no information on the socio-economic origin of students was collected in any jurisdiction (Gilliss, 1994, p. 5).

Gilliss (1993) states that “[a]nyone who expects students who are disadvantaged, impoverished, undernourished, neglected or severely disabled to score at the same level as students who are affluent, healthy, well cared for or able [bodied] is doomed to disappointment.” She also observes that average test scores for schools provide more information about the pooled characteristics of individual students than about the schools themselves. Rather than incur the expense of developing, administering, grading, and ranking test scores to determine which schools score the highest, a socio-economic analysis of census data of the communities where the schools are located would provide much the same information. McLaughlin (1991) echoes this point. Responding to the call for standardized testing in “America 2000” (the national education strategy for the U.S. announced under former President George Bush), she notes that testing same intelligences in precisely the same proportions.” (p. 71) This of course has important implications for educational assessment and evaluation practices.

√ Evaluation is the process of making judgements, based in part on assessment data.

√ Assessment and evaluation are continuous processes that inform decision-making at classroom and system levels.

√ Classroom teachers have the primary responsibility for assessing and evaluating student achievement.

√ Governments and school systems have a collective responsibility to provide in-service opportunities for teachers in the areas of assessment and evaluation.

√ Teachers’ organizations, as the professional voice of teachers, must be directly involved in all policy decisions about assessment and evaluation at provincial/territorial, national and international levels.

√ The purpose of student assessment is to support student learning.

√ The design, interpretation and application of information derived from assessment and evaluation are complex tasks linked to professional beliefs, skills and knowledge.

√ Decisions regarding evaluation and assessment should focus on what is valuable and not simply on what is measurable.

√ The purposes and methods of assessing and evaluating student learning, curriculum, programs and systems are distinct and not interchangeable.

√ The tools and processes used to assess and evaluate student learning must be congruent with the learning theory that is reflected in the curriculum.
Inability to assess student learning and development

Although standardized tests may be useful for sorting and ranking students, they are inadequate in assessing student learning and development. The relevant literature consistently identifies the following reasons:

Many types of student ability are not captured by a standardized test.

Qualities such as a student’s sense of citizenship, ethics, confidence/self-esteem, aesthetic appreciation, respect for others, self-discipline, social competence, and desire to learn, although not easily quantifiable, contribute enormously to the multiple goals of schooling. By forcing schools and teachers to concentrate on basic, quantifiable skills, standardized testing poses a serious threat to this diversity of goals and content. Medina and Neill tell us that “[t]he ability of standardized tests to accurately report students’ knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score.” (p. 12)

In his seminal work *The Mismeasure of Man*, Gould debunked two major myths about human intelligence (as cited in Darling-Hammond, 1994, pp. 10-11): that it is an “innate, unitary, measurable commodity” and, flowing from the notion that intelligence could be quantified in this way, that individuals can be compared and ranked on this basis. Different perspectives on the concept of intelligence are emerging. Gardner (1999) notes that areas of debate include “what intelligence is, how (and whether) it should be measured, and which values should be invoked in considerations of the human intellect.” (p. 70) He proposes that we all “possess at least eight intelligences: linguistic and logical-mathematical (the two most prized in school and the ones central to success on standard intelligence tests), musical, spatial, bodily-kinesthetic, naturalist, interpersonal, and intrapersonal” and that “no two of us exhibit the
They are often referred to as norm-referenced tests because an individual’s scores are compared to the scores of a larger representative group (the “norm”). The test format is frequently multiple choice, making it easier to mark with the assistance of computers. Aside from advances in electronic scoring, analysis and reporting of results, many standardized achievement tests have changed very little over the past fifty years despite claims that they have become more “scientific” (Medina & Neill, 1990, p. 13) and despite advances in our knowledge about what makes for good teaching and effective learning.

A substantial body of research has documented numerous flaws in the construction, validation, administration and use of standardized tests. According to Medina and Neill of FairTest (the National Center for Fair & Open Testing in Massachusetts),

[s]tandardized tests are constructed in ways that often guarantee biased results against minorities, females and low-income students. Test results are evaluated and scored in ways that are often at odds with modern theories of intelligence and child development. The test validation process is often inadequate and far from objective. Many tests are administered in an environment that undermines any claims they may have to being “standardized”. Even those that adhere to “standard” administration practices may be biasing the results against minorities, low-income students and females by using examiners who are unfamiliar to the test-takers and by using tightly timed tests. These flaws undermine testmakers’ claims of objectivity and produce results that are inaccurate, unreliable or biased. Ultimately, many tests fail to effectively measure test-takers’ achievement, abilities or skills. (p. 8)

References


Problems with Standardized Testing – An Overview

Developed during the First World War to identify army recruits who could be trained as military officers, standardized tests were a means of spreading the candidates out in order to spot those best suited for leadership roles (Casas & Meaghan, 1995, p. 18). Traub (1994) defines standardized achievement tests in the following way:

Generically, an achievement test is designed to assess the knowledge and understanding a student has acquired of a school subject. A standardized achievement test is further defined by its being given and scored in the same way, whenever and wherever it is used. Standardization means that the scores of all students tested can be fairly compared, one against the other ... the essential requirements are that the conditions of administration and scoring be the same for all the students who are tested so that their scores can be compared. (p. 5)

A summary of some of the major problems with standardized testing is contained in Appendix A.

The U.S. Scholastic Aptitude Test (SAT), developed in the 1920s, was based on the Army Alpha test, a test used to select potential officers from among army recruits (Lemann, 1995). For a brief history of testing in general (including the introduction in the 18th century of ranking and the invention of the quantified mark), see Madaus & Kellaghan (1993).
While advocates claim that standardized testing will increase the accountability of teachers and schools, there is little evidence to support such claims. There is, however, a good deal of evidence documenting numerous problems associated with standardized testing (Meaghan & Casas, 1995d, p. 46). Problems stem not just from the nature (form and content) of the tests but perhaps more importantly, from the way in which tests and their results are used. This report considers some of these problems, particularly the impact of testing bias and the misuse of test results on educational equity. It will also briefly discuss authentic or performance assessment, address standardized testing from a political perspective, and offer some general observations on the assessment of student learning and development. Appendices are included which contain principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

The numerous references in the report to testing in the U.S. reflect the extensive experience Americans have had with this practice, experience which holds valuable lessons for Canada. It is estimated that American public schools administered more than 200 million standardized tests over the 1986-87 school year (Medina & Neill, 1990). According to Barlow and Robertson (1994), “if testing created excellence, American students, the ‘most-tested’ students in the world, should be winning the education sweepstakes” (p. 117); instead, the U.S. consistently scores low on large-scale international assessments (Earl, 1995, p. 5). Research conducted over the past two decades confirms that standardized tests have had “many negative consequences for the quality of American schooling and for the equitable allocation of school opportunities.” (Darling-Hammond, 1991, p. 220)

---

2 While the terms *assessment* and *evaluation* are often used interchangeably (as is the case in this report), there is a subtle yet important distinction between them. Assessment is defined as the gathering of reliable information pertaining to students’ knowledge and understanding including social, economic and educational factors and resources that influence student and system performance. Evaluation is the process of making judgements, based in part on assessment data.
Introduction

Addressing a Canadian Teachers’ Federation workshop on curriculum and evaluation in 1998, Benjamin Levin, Dean of Continuing Education at the University of Manitoba, provided an overview of the changes occurring across the country with respect to educational policy. On the curriculum side, he noted that policy changes include an increased focus on the ‘basics’, fewer options and more restrictive time allocations, the development of regional and Pan-Canadian curricula, and shorter implementation times for new curricula. Changes in assessment policy include more provincial testing, more testing at the national and international levels, increased public reporting of test results despite the generally poor quality of this reporting, and more importance given to the results. Canada is not alone in placing an increased focus on testing. According to Levin (1998),

an emphasis on standards, accountability and testing has been a feature of reforms in many countries. Almost everywhere we find more large-scale testing of students and more reporting of the results of these tests than was the case a few years ago. Increasing national assessment is complemented by more and more international assessment and in both cases the results are used more overtly for public comparisons. (p. 133)

---

1 The National Issues in Education Workshop on “Curriculum and Evaluation: Teacher Organizations and Pan-Canadian Developments” was convened by the Canadian Teachers’ Federation in January 1998 to bring the leaders of teacher organizations together to discuss recent developments in curriculum and evaluation, particularly their increasingly “Pan-Canadian” nature. Workshop participants developed statements of principle on curriculum, assessment and evaluation which are contained in Appendix C at the end of this report.


Preface

Canadian schools are awash in student testing. Changes in assessment policy have increased standardized testing at provincial, national and international levels, introduced testing at more (and earlier) grade levels, increased the reporting of test results and attached more significance to those results. Advocates claim that more testing will result in greater school accountability – testing seems to have become synonymous with accountability.

In this context of an increased focus on testing as an instrument of educational reform, CTF’s Standardized Testing: Undermining Equity in Education argues that there is little evidence to support standardized testing from a pedagogical perspective and looks at the effects of standardized testing on educational equity. Problems with standardized testing are well-documented. They stem from the nature of the tests and the way in which tests and their results are used. This report considers some of these problems, with particular emphasis on the impact of testing bias and the misuse of test results on equity in education. It also discusses performance-based assessment, examines the politics of testing, offers some general observations on the assessment of student learning, and provides principles developed by educators to guide decision-making in the areas of assessment, evaluation and curriculum.

Intended to increase awareness of critical issues related to the impact of standardized testing, this book will be a useful resource for teachers at all levels as well as administrators, researchers, policy-makers and anyone with an interest in student assessment and evaluation.

Bernie Froese-Germain is a researcher with the Canadian Teachers’ Federation in Ottawa.


Standardized Testing: Undermining Equity in Education

Report prepared for the National Issues in Education Initiative

by Bernie Froese-Germain
Canadian Teachers’ Federation

November 1999
Standardized Testing: Undermining Equity in Education

Report prepared for the National Issues in Education Initiative

by Bernie Froese-Germain
Canadian Teachers’ Federation

November 1999